A Unification of Lower Bounds on Prediction Dimension of Consistent Convex Surrogates

author names withheld

Editor: Under Review for COLT 2021

Abstract

Given a prediction task, understanding when one can and cannot design a consistent convex surrogate loss, particularly a low-dimensional one, is an important and active area of machine learning research. Surrogate loss construction often presents itself in one of two scenarios, typically studied by different frameworks: first, when one is given a target loss for a finite prediction task, and second, when one simply has a statistic they would like to estimate. Motivated by settings such as structured prediction where the prediction dimension of the surrogate is of central importance, we give a novel lower bound on the prediction dimension that applies to both frameworks. Our lower bound tightens existing results in the case of discrete predictions, namely the feasible subspace dimension of Ramaswamy and Agarwal (2016), showing that previous calibration-based bounds can largely be recovered via property elicitation. For statistic estimation, our lower bound gives new results for estimating variance as well as the entropy and norms of the conditional distribution.

1. Introduction

[JF: Without moving anything around, take a shot at storyboarding this. Want enough detail to understand how much time to spend on each chunk. (1 bullet per paragraph)] Surrogate risk minimization is one of the most widespread optimization heuristics in supervised machine learning. Often, we desire surrogate losses to be *consistent*, as this precedes deriving excess risk bounds and rates. Roughly speaking, consistency means that minimizing surrogate risk corresponds to solving the target problem of interest: the target risk should also be minimized, or the average deviation from the true conditional statistic function converges to 0.[JF: Not sure if I like these sentences here, or in general]

A variety of prediction tasks call for the construction of consistent, and ideally convex, surrogates. Applications ranging from extreme classification to top-k predictions to ranking have inspired an entire line of research into the construction of surrogate loss functions for such discrete tasks. The dimensionality of such surrogate prediction space often scales linearly with the number of outcomes unless sacrifices are made in terms of consistency if one is not careful. This linear growth of prediction dimension (in the number of outcomes) leads to computationally expensive or intractable optimization problems, but it is unknown when it is possible to design a surrogate in a lower-dimensional prediction space for many tasks that have large outcome spaces.

We do know that for some tasks, it is possible to reduce the dimension of this surrogate prediction space without sacrificing consistency. Consider the high-confidence classification presented by the abstain target loss (Ramaswamy and Agarwal, 2012; Ramaswamy et al.,

2018). When trying to predict one of n outcomes, the abstain loss is minimized by predicting the most likely outcome if it is likely enough, and "abstaining" from predicting otherwise. Ramaswamy and Agarwal (2016) present a convex surrogate, called the BEP loss, for the abstain target loss that takes as input a prediction logarithmic in the number of outcomes. The BEP loss is a high-dimensional generalization of the hinge loss so that outcome predictions are "embedded" into the corners of the ± 1 hypercube, and abstaining is embedded to the origin. This BEP loss is consistent with respect to the target 0-1 modification $\ell(r, y) = \mathbf{I}\{r \notin \{y, \perp\}\} + (1/2)\mathbf{I}\{r = \perp\}$, and takes $\Theta(\log n)$-dimensional surrogate predictions, instead of $\Theta(n)$, as was required of previous consistent surrogates.[JF: Some of this repetitive]

In its own line of literature, financial institutions have sought scoring rules to learn banks' estimates of financial risk in their investments.[JF: cite] Given the desired measure of financial risk, such as the variance or Conditional Value at Risk (CVaR), auditing institutions are tasked with designing scoring rules that, directly or indirectly, incentivize truthfully estimating the desired statistic. For example, while there is no loss function that enables one to directly learn the variance, we can learn the first and second central moments [JF: how much detail to share?] by proper scoring rules, which yields an unbiased estimator of the variance. However, asking a bank for a large number of predictions may be infeasible by their own knowledge of the financial risk distribution. That is, institutions and individuals may only be expected to articulate predictions of summary statistics of their beliefs, rather than their entire belief. To this end, learning desired statistics with a minimal number of reports also emerges as an important question in finance. [JF: Raf, helppp]

The two examples above are seemingly unrelated given how different the settings are; in the abstain problem, we have a finite prediction problem where we want to design a consistent convex surrogate for a given target loss. In risk estimation, on the other hand, we aim to estimate a continuous variable based on the desired summary statistic instead of target loss. These two types of problems have been historically studied with different mathematical tools because of these differences; however, in this paper, we show formal connections between consistent surrogates and property elicitation, and leverage these to understand new bounds on prediction dimension for consistent surrogates that spans both settings.

2. Background and Related work

We consider supervised learning problems in the space $\mathcal{X} \times \mathcal{Y}$, for some *feature space* $\mathcal{X}$ and a *label space* $\mathcal{Y}$ of size n, with data drawn from a distribution D over $\mathcal{X} \times \mathcal{Y}$. The task is to produce a hypothesis $f : \mathcal{X} \to \mathcal{R}$, for some *prediction space* $\mathcal{R}$, which may be different from $\mathcal{Y}$. For example, in ranking problems, $\mathcal{R}$ may be all $n!$ permutations over the n labels forming $\mathcal{Y}$. As we focus heavily on conditional distributions $p := D_x = \Pr[Y|X = x]$ over $\mathcal{Y}$ given some $x \in X$, we often abstract away x, working directly with distributions over outcomes $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, through tools such as calibration (Steinwart and Christmann, 2008, Chapter 3).

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 $u \in \mathbb{R}^d$. We write $\mathcal{L}_d$ for the set of $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{Y}$-measurable and lower semi-continuous surrogates

$L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}_p L(u, Y) < \infty$ for all $u \in \mathbb{R}^d, p \in \mathcal{P}$, and are minimizable, in the sense that $\arg\min_u \mathbb{E}_p L(u, Y)$ is nonempty for all $p \in \mathcal{P}$. [JF: Sufficient condition for $\mathcal{A}$-normal convex integrand: L is lower semi-continuous and $\mathcal{B}(\mathcal{A}) \otimes \mathcal{Y}$-measurable; $\mathbb{E}_p L(u, Y)$ finite for all u, p, there exists a u_0 for each p so that $\mathbb{E}_p L(u_0, Y)$ is finite and continuous for Rockefellar's corollary, though it's stricter than Ioffe and Tokhimorov IIRC] Moreover, we write $\mathcal{L} = \cup_{d \in \mathbb{N}} \mathcal{L}_d$. A loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ is *discrete* if $\mathcal{R}$ is a finite set. A surrogate $L : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ is *convex* if $L(\cdot, y)$ is convex in u almost surely in $\mathcal{Y}$. For a given $p \in \mathcal{P}$, the (conditional) *regret*, or excess risk, of a loss L is given by $R_L(u, p) := \mathbb{E}_p L(u, Y) - \inf_{u^*} \mathbb{E}_p L(u^*, Y)$. Typically, we notate finite report sets $\mathcal{R}$ (and $\mathcal{R}'$ if a second finite set is needed) and continuous prediction spaces $\mathcal{U} \subseteq \mathbb{R}^d$.

2.1. Consistency

[JF: Looking at this now, sections 2.1-4 feel like a lot...] [JF: I think the story we want to tell is as follows: there are these two types of problems where [convex] prediction dimension is important. Typically, these problems are studied with different tools; here, we present a tool to bound prediction dimension in both settings. This tool relies on property elicitation, which we show (somehow for the first time to our knowledge) can be formally connected to consistency.] [JF: Could split into high level in previous work; ie previous work and setting are distinct.] [JF: Top of 2 in setting] [JF: Want a clear distinction between what came before and what we're adding.] [JF: Can defer Lemma 8 to appendix? Maybe]

A basic requirement of surrogate losses $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is consistency, which roughly means that minimizing L-loss corresponds to solving the target problem of interest. We consider two kinds of consistency. In the first, we are given a *target loss* ℓ, and we roughly define L to be consistent if minimizing L and applying a link, minimizes ℓ (Definition 6). This definition follows much of the machine learning literature (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Steinwart, 2007; Ramaswamy and Agarwal, 2016). In the second notion of consistency, we are given a *target statistic*, such as the conditional variance or quantile, as in classical statistics, but no target loss.[JF: citation for stats def] Here we will define L to be consistent if minimizing L and applying a link function yields estimates converging to the correct value, by some measure (Definition 7).

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 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. [JF: How does this paragraph change?]

The goal of this paper is to give lower bounds on the dimension d of consistent surrogate losses. A priori, it is not necessarily clear that compatible definitions of consistency could be given for both target statistics and target losses. We observe that, in fact, target losses are a special case of target statistics (§ 3), which suggests property elicitation (see § 2.3) as being uniquely [JF: not nec. uniquely?] suited to studying general lower bounds. In prior work, most research on this problem focuses on the quadrant of target losses and discrete predictions (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Ramaswamy et al., 2015; Ramaswamy and Agarwal, 2016; Ramaswamy et al., 2018). In particular, as definitions of consistency are relatively intractable to apply directly, the literature often

focuses on a weaker condition called calibration, which only applies when given a discrete target loss. [JF: Target doesn't have to be discrete, but $\mathcal{Y}$ finite, right?] [JF: Make this section more concise?]

2.2. 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 (2008, Chapter 3), which aligns with the definition in Tewari and Bartlett (2007) for multiclass classification, and its generalizations to arbitrary discrete target losses (Agarwal and Agarwal, 2015; Ramaswamy and Agarwal, 2016). Calibration is more tractable and weaker than consistency, yet the two are equivalent under suitable assumptions Tewari and Bartlett (2007); Ramaswamy and Agarwal (2016). Intuitively, calibration says one cannot achieve the optimal surrogate loss while linking to a suboptimal target prediction.

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

$$\forall p \in \Delta_{\mathcal{Y}} : \inf_{u \in \mathcal{U} : \psi(u) \notin \arg\min_r \mathbb{E}_p \ell(r,Y)} \mathbb{E}_p L(u,Y) > \inf_{u \in \mathcal{U}} \mathbb{E}_p L(u,Y) \ . \tag{1}$$

Typically, we take $\mathcal{U} = \mathbb{R}^d$ for some $d \in \mathbb{N}$. BTW: We could also replace $\mathcal{U}$ by $\mathbb{R}^d$ throughout, but it seemed distracting in this section

Many works characterize calibrated surrogates for specific discrete target losses (Zhang, 2004; Lin, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007), including the canonical 0-1 loss for binary and multiclass classification. In Appendix A, we give a definition of calibration similar to that of Steinwart and Christmann (2008), which is equivalent to Definition 1 in discrete prediction settings, but generalizes to continuous estimation settings.

2.3. Property elicitation

Arising from the statistics and economics literature, property elicitation is similar to calibration, but only places a restriction on the exact minimizers of a surrogate (Savage, 1971; Osband and Reichelstein, 1985; Lambert et al., 2008; Lambert and Shoham, 2009; Lambert, 2018; Frongillo and Kash, 2015, 2014). Specifically, given a statistic or *property* Γ of interest, which maps a distribution p over $\mathcal{Y}$ to the set of desired or correct predictions, the minimizers of L should precisely coincide with Γ. Intuitively, $p = \Pr[Y|X = x]$ is a conditional distribution, though the definition is also applied to point prediction settings.

RF: FYI Jessie: I like this line but D_x has not been defined yet [JF: I added D in the notation section and introduced it there. Let me know if it really doesn't mesh with you, but I think it fit somewhat naturally.]

Definition 2 (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*

$$\forall p \in \Delta_{\mathcal{Y}}, \quad \Gamma(p) = \arg\min_{u \in \mathcal{R}} \mathbb{E}_p L(u,Y) \ . \tag{2}$$

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. A property is *single-valued* if $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$.

When $L \in \mathcal{L}$, we use $\Gamma := \text{prop}[L]$ to denote the unique property elicited by L from eq. (2). [JF: Small detail: probably want to specify the domain in this notation at least once, since we move away from $\mathcal{P} = \Delta_\mathcal{Y}$] Finally, we typically denote the target property by γ, and the surrogate by Γ.

To relate property elicitation to consistency, we need to allow for a link function, which gives rise to the notion of *indirect* elicitation. For single-valued properties, this definition reduces to the natural requirement $\gamma = \psi \circ \Gamma$.

Definition 3 (Indirect Elicitation) *A loss and link (L, ψ)* indirectly elicit *a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ if L elicits a property $\Gamma : \mathcal{P} \rightrightarrows \mathcal{U}$ such that for all $u \in \mathcal{U}$, we have $\Gamma_u \subseteq \gamma_{\psi(u)}$. We say L* indirectly elicits *γ if such a link ψ exists. BTW: interesting discussion of set-valued properties commented out; revive later!*

The close connection between indirect elicitation and calibration was first explored by Agarwal and Agarwal (2015). In particular, calibration of $L \in \mathcal{L}$ with respect to ℓ implies indirect elicitation quite directly: take $u \in \mathcal{U}$ and $p \in \Gamma_u$, implying $u \in \Gamma(p)$. From eq. (2), $\mathbb{E}_p L(u, Y) = \inf_{u' \in \mathcal{U}} \mathbb{E}_p L(u', Y)$, so we must have $\psi(u) \in \gamma(p)$ from eq. (1), as desired. As a result, indirect elicitation in a necessary condition for consistency in this case, and in fact, we show this is true for all four quadrants mentioned above (§ 3). [JF: quadrant speak vs "in both cases" (discrete target loss vs continuous estimation)?]

An important caveat to the above definitions is that, since $\Gamma = \text{prop}[L]$ is nonempty everywhere, we must have $L \in \mathcal{L}$, meaning that $\mathbb{E}_p L(\cdot, Y)$ always achieves a minimum. This restriction is also implicit in e.g. (Agarwal and Agarwal, 2015). While some popular surrogates such as logistic and exponential loss are not minimizable, these losses are still covered in Corollary 13 and Theorem 17 as $\Gamma(p) \neq \emptyset$ when $p \in \text{relint}(\Delta_\mathcal{Y})$; 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), thereby ensuring a minimum is always achieved for convex losses, would allow us to lift the minimizable restriction entirely. [RF: Check me!] BTW: Some refs / related work commented out here. I think not as relevant as elicitation complexity.

2.4. Prediction 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 (Steinwart and Christmann, 2008; Agarwal and Agarwal, 2015; Ramaswamy and Agarwal, 2016) and property elicitation (Frongillo and Kash, 2015; Fissler et al., 2016; Frongillo and Kash, 2018). For discrete target losses, Ramaswamy and Agarwal (2016) introduce *convex calibration dimension*, the minimum prediction dimension yielding a convex calibrated surrogate. Their results has led to the design of consistent convex surrogates for discrete prediction problems such as hierarchical classification (Ramaswamy et al., 2015) and classification with an abstain option (Ramaswamy et al., 2018).

RF: Nice! Love "proxies" here.

Definition 4 (Convex Calibration Dimension) *Given a target discrete loss ℓ, its* convex calibration dimension $\text{cc}\dim(\ell)$ *is the minimum dimension d such that there is a convex surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and link ψ such that (L, ψ) is calibrated with respect to ℓ. BTW:*

RF: FYI Jessie: Two general points. (1) I spent the majority of my time in this section just on dependencies; e.g. you had "similar to elicitation complexity mentioned above" but elicitation complexity was not mentioned above. Next time it would be more efficient if you made sure to balance local edits with global scans to catch these sorts of things, as well as making sure we

Maybe take out of definition environment, but (a) I'd worry that people wouldn't be able to find the definition when we use it later, and (b) I'm hoping we can make the distinction very clear, which now is achieved by the parallel definitions

In the case of a target statistic/property γ, Lambert et al. (2008) similarly introduce the notion of *elicitation complexity*, later generalized by Frongillo and Kash (2018), which captures the lowest prediction dimension of a surrogate which indirectly elicits γ. This notion is more general as it extends to continuous estimation settings and does not inherently depend on a target loss being given.

Definition 5 (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 convex surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ indirectly eliciting γ.*

In particular, $\mathrm{elic}_{\mathrm{cvx}} \leq \mathrm{cc\,dim}$ when restricting to minimizable surrogates ($\mathcal{L}$).

In related work, Agarwal and Agarwal (2015, Corollary 10) provide a necessary condition for the direct convex elicitation of single-valued properties, yielding bounds on the dimensionality of level sets. Moreover, Finocchiaro et al. (2019) study surrogate losses which *embed* a discrete loss, which is a special case of indirect elicitation. They construct a calibrated link functions from embeddings, which together with observations about indirect elicitation, imply that for polyhedral (piecewise linear and convex) losses, calibration and indirect elicitation are equivalent. Finocchiaro et al. (2020) further introduce the notion of *embedding dimension*, which is a lower bound on both convex elicitation complexity of discrete properties and convex calibration dimension of discrete losses.

3. Consistency implies indirect elicitation

[JF: Simplify; merge with Sec 4. Jessie take a pass at proof sketch of Theorem 1] [JF: Narrative here reads to me as "hey this general result through calibration is our main contribution" which I'm not sure if we particularly want. Also want to double check how different our calibration definition is from Steinwart]

In this section, we show that indirect elicitation is necessary for consistency (Theorem 9) for losses in $\mathcal{L}$, while simultaneously applying to all of the above settings (our four quadrants). [JF: Tagging quadrant speak] For the case of a given discrete target loss, it is well-known that a surrogate is consistent if and only if it is calibrated (e.g. (Bartlett et al., 2006, Theorem 1, part 3)). For this case, we can show Theorem 9 via calibration, and even extend the definition of calibration and proof approach to continuous prediction spaces. As we are also interested in the other two quadrants, when given a target statistic instead of as target loss, we delegate this proof to Appendix A and directly prove the general result for all four quadrants. [JF: Quadrant speak]

Since indirect elicitation is implied by both consistency and calibration of a surrogate $L \in \mathcal{L}$, it might seem a very weak necessary condition for consistency. Yet, as we show in the following sections, it gives state-of-the-art lower bounds on the prediction dimension of consistent convex surrogates. In particular, our bounds imply those given via feasible subspace dimension (Ramaswamy and Agarwal, 2016, Theorem 16) in Corollary 16.

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 implies applying a link ψ optimizes ℓ (Definition 6). 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 7). We then observe that Definition 6 is subsumed by Definition 7, and use this to show consistency implies L indirectly elicits prop$[\ell]$ or γ respectively.

Definition 6 (Consistent: loss) *A loss and link (L, ψ) are consistent with respect to a target loss ℓ if, for all distributions D over input and label spaces $\mathcal{X} \times \mathcal{Y}$, and for all sequences of measurable hypothesis functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

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

RF: I think okay as written – if ψ is restrictive, you won't achieve consistency. And typically f is much more expressive anyway, e.g. $\mathbb{R}^{\mathcal{Y}}$ for classification.

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 (Györfi et al., 2006; Fan and Yao, 1998; Ruppert et al., 1997), 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 $\mu(r, p) = 0 \iff r \in \gamma(p)$. BTW: Bo: Would be nice to give some natural special cases: for a finite property, $\mu(r, p) = \mathbf{I}\{r \in \gamma(p)\}$, and for single-valued properties with a distance metric on $\mathcal{R}$, $\mu(r, p) = \text{dist}(r, \gamma(p))$.

Definition 7 (Consistent: property) *Suppose we are given a loss $L : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$, link function $\psi : \mathcal{U} \to \mathcal{R}'$, and property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}'$. Moreover, let $\mu : \mathcal{R}' \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ be any function satisfying $\mu(r, p) = 0 \iff r \in \gamma(p)$. We say (L, ψ) is* μ-consistent with respect to γ *if, for all distributions D over $\mathcal{X} \times \mathcal{Y}$, and for all sequences of measurable functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \ , \tag{3}$$

where $D_x = \Pr[Y|X = x]$. We say (L, ψ) is consistent with respect to γ if there is such a μ so that (L, ψ) is μ-consistent with respect to γ.

If γ is a property elicited by a target loss ℓ, then taking μ to be the ℓ-regret makes the two definitions are equivalent.

Lemma 8 *Given a surrogate loss L, link ψ, and target loss ℓ, set $\mu(r, p) := R_\ell(r, p)$. Then (L, ψ) is consistent with respect to ℓ if and only if (L, ψ) is μ-consistent with respect to $\gamma := \text{prop}[\ell]$. BTW: Note: don't actually need γ nonempty here.*

[JF: Defer proofs to appendix from this section]

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 6,

$$\mathbb{E}_D \ell(\psi \circ f_m(X), Y) \to \inf_f \mathbb{E}_D \ell(\psi \circ f(X), Y) \tag{4}$$
$$\iff \mathbb{E}_X R_\ell(\psi \circ f_m(X), D_X) \to 0$$
$$\iff \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \; . \tag{5}$$

Therefore, $\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y)$ implies (4) if and only if it implies (5). ■

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. This points to indirect elicitation as a natural necessary condition for consistency, as formalized in Theorem 9.

Theorem 9 *For a convex surrogate $L \in \mathcal{L}$, if the pair (L, ψ) is consistent with respect to a property γ or a loss ℓ eliciting γ, then (L, ψ) indirectly elicits γ.*

Proof By Lemma 8, it suffices to show the result for consistency with respect to a property γ, setting $\gamma := \text{prop}[\ell]$ if ℓ is given instead. We show the contrapositive; suppose (L, ψ) does not indirectly elicit γ, meaning we have some $p \in \Delta_\mathcal{Y}$ so that $u \in \Gamma(p)$ but $\psi(u) \notin \gamma(p)$, where $\Gamma := \text{prop}[L]$. Observe that we use the fact $\Gamma(p) \neq \emptyset$. Consider the constant sequence $\{f_m\}$ with $f_m(x) = u$ for all m, x, and take D with full support on some $x \in \mathcal{X}$, and let $D_x = p$. Since $u \in \Gamma(p)$, we observe $\mathbb{E}_D L(f_m(X), Y) = \inf_f L(f(X), Y)$ for all m; in particular $\mathbb{E}_D L(f_m(X), Y) \to \inf_f L(f(X), Y)$. However, we have $\mathbb{E}_X \mu(\psi \circ f_m(X), D_X) = \mu(\psi(u_m), p) = \mu(\psi(u), p) \neq 0$, since $\psi(u) \notin \gamma(p)$. Thus, we observe (L, ψ) is not consistent with respect to γ (Definition 7). ■

4. Prediction Dimension of Consistent Convex Surrogates

[JF: Feels like this section got cut a lot for NeurIPS, but seems like some more narrative in this section might be helpful.]

We now turn to the question of bounding the prediction dimension of a consistent convex surrogate. From Theorem 9, given a target property γ or loss ℓ with $\gamma = \text{prop}[\ell]$, this task reduces to lower bounding the prediction dimension of a convex surrogate indirectly eliciting γ. We now explore a tool, Theorem 14, for proving such convex-elicitation lower bounds. The key idea, crystallized from the proofs of Ramaswamy and Agarwal (2016, Theorem 16) and Agarwal and Agarwal (2015, Theorem 9), is to consider a particular distribution p and surrogate prediction $u \in \mathbb{R}^d$ with is optimal for p. Theorem 14 will show that if d is small, then the level set $\{p : u \in \arg\min_{u'} \mathbb{E}_p L(u', Y)\}$ must be large; in fact, it must roughly contain a high-dimensional affine subspace, or *flat*. By definition of indirect elicitation, there is some level set γ_r (where u is linked to r) containing this flat as well. The use of this result is to leverage the contrapositive: if γ has a level set intricate enough to not contain any high-dimensional flats, then γ cannot have a low-dimensional consistent surrogate.

Recall $\text{affhull}(\Delta_\mathcal{Y}) = \{\sum_{i=1}^n \alpha_i p_i : p_i \in \Delta_\mathcal{Y}, \alpha_i \in \mathbb{R}, \sum_{i=1}^n \alpha_i = 1\}$, which has dimension $n - 1$.

Definition 10 (Flat) *For $d \in \mathbb{N}$, a d-*flat*, or simply* flat *is a nonempty set $F = \ker_{\mathcal{P}} W = \{q \in \mathcal{P} : \mathbb{E}_q W = \mathbf{0}\}$ for some $W : \mathcal{Y} \to \mathbb{R}^d$.*

We now state our elicitation lower bound, which when combined with Theorem 9, implies consistency bounds. [JF: Be explicit about corollary; add corollary after Thm 11.] A similar result is Agarwal and Agarwal (2015, Theorem 9), which bounds the dimension of level sets of a single-valued prop$[L]$. Corollaries 12 and 13 instead bound the dimension of flats contained in the level sets, an additional power which we leverage in our examples. Corollaries 12 and 13 also give a lower bound for any given target ℓ or γ, rather than a given surrogate L. [JF: Does AA15 do this? Seems out of the blue without that context.] [RF: Cut the last bits here; I bet AA15 thought of their result as being about consistency, though I could be wrong]

Lemma 11 *Let Γ be (directly) elicited by a convex $L \in \mathcal{L}_d$ for some $d \in \mathbb{N}$. For all $u \in \operatorname{range} \Gamma$ and $p \in \Gamma_u$, there is some $V_{u,p} : \mathcal{Y} \to \mathbb{R}^d$ such that $p \in \ker_{\mathcal{P}} V_{u,p} \subseteq \Gamma_u$.*

Proof [JF: Direct] Let $\Gamma := \text{prop}[L]$. As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \mathbf{0} \in \partial \mathbb{E}_p L(u, Y)$. We proceed in two cases, depending on $|\mathcal{Y}|$.

Finite $\mathcal{Y}$: If $\mathcal{Y}$ is finite, this is additionally equivalent to $\mathbf{0} \in \oplus_y p_y \partial L(u, y)$, where $\oplus$ denotes the Minkowski sum (Hiriart-Urruty and Lemaréchal, 2012, Theorem 4.1.1).[1] Expanding, we have $\oplus_y p_y \partial L(u, y) = \{\sum_{y \in \mathcal{Y}} p_y x_y : x_y \in \partial L(u, y) \ \forall y \in \mathcal{Y}\}$, and thus $Wp = \sum_y p_y x_y = \mathbf{0}$ where $W = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$; cf. (Ramaswamy and Agarwal, 2016, $\mathbf{A}^m$ 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.

RF: Simplified prev line

Infinite $\mathcal{Y}$: $L \in \mathcal{L}_d$ and convex implies L satisfies the assumptions of Ioffe and Tikhomirov (1969), we can interchange subdifferentiation and expectation to observe $\partial \mathbb{E}_p L(u, Y) = \mathbb{E}_p \partial L(u, Y) := \{\int V(y) dp(y) : V \text{measurable}, V(y) \in \partial L(u, y) \text{a.s. in } p\}$. [JF: What happens if $\mathbf{0}$ not in the set, but is in the closure][JF: Cite Ioffe and Tikhomirov (1969) instead; they don't use closure and it's the same result. (Instead of (Rockafellar and Wets, 1982, Cor 1))] We can now take $V_{u,p}$ to be such a $V : \mathcal{Y} \to \mathbb{R}^d$, which exists as $\mathbf{0} \in \partial \mathbb{E}_p L(u, Y)$, and therefore, the set is nonempty.

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 $q \in \ker_{\mathcal{P}} V_{u,p} \implies \mathbf{0} \in \partial \mathbb{E}_q L(u, Y) \implies u \in \Gamma(q) \implies q \in \Gamma_u$. ■

Corollary 12 *Let target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ and $d \in \mathbb{N}$ be given. Let $p \in \mathcal{P}$ with $|\gamma(p)| = 1$, and take $\gamma(p) = \{r\}$. If there is no flat F that is a d-representation with $p \in F \subseteq \gamma_r$, then no $L \in \mathcal{L}_d$ indirectly elicits γ, and in particular, there is no convex surrogate $L \in \mathcal{L}_d$ consistent with respect to γ.*

Proof Let (L, ψ) indirectly elicit γ and the convex function L and elicit Γ. As Γ is non-empty, there is some $u \in \Gamma(p)$. Since γ is single-valued at p, we have $r = \psi(u)$; by Lemma 11, we know there is a flat $F = \ker_{\mathcal{P}} V_{u,p}$ so that $p \in F \subseteq \Gamma_u$. Moreover, by construction of such a

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

flat, we know it has a d-representation. By definition of indirect elicitation, we additionally have $\Gamma_u \subseteq \gamma_r$. Thus, we have $p \in F \subseteq \gamma_r$.

Now, if we do not have a flat satisfying the above conditions, then we do not have any L so that prop$[L]$ indirectly elicits γ, and in turn do not have a link such that L is consistent with respect to γ by Theorem 9. ■

Corollary 13 *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 $p \in \text{relint}(\mathcal{P})$. If there is no flat F of that has a d-representation with $p \in F \subseteq \gamma_r$, then no $L \in \mathcal{L}_d$ indirectly elicits γ, and in particular, there is no convex surrogate $L \in \mathcal{L}_d$ consistent with respect to γ.*

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 Lemma 11. If $r = r'$, then $p \in F \subseteq \Gamma_u \subseteq \gamma_r$ by indirect elicitation. Otherwise, by Lemma 26, 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 we do not have a flat satisfying the above conditions, then we do not have any L so that prop$[L]$ indirectly elicits γ, and in turn do not have a link such that $L \in \mathcal{L}_d$ is consistent with respect to γ by Theorem 9. ■

BTW: Jessie: Suggest delete - this is the original version

Theorem 14 *Let a property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ be given, let $p \in \Delta_\mathcal{Y}$ such that either (i) $|\gamma(p)| = 1$ or (ii) γ is elicitable and $p \in \text{relint}(\Delta_\mathcal{Y})$, and let $r \in \gamma(p)$. Then if a convex $L \in \mathcal{L}_d$ indirectly elicits γ, there is a flat F with $p \in F \cap \Delta_\mathcal{Y} \subseteq \gamma_r$ and $\text{codim}(F) \leq d$.*

Proof

Let $\Gamma := \text{prop}[L]$ and suppose (L, ψ) indirectly elicits γ. By definition of a property, there is some $u \in \Gamma(p)$. We will first show $F \cap \Delta_\mathcal{Y} \subseteq \Gamma_u$.

As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \mathbf{0} \in \partial \mathbb{E}_p L(u, Y) \iff \mathbf{0} \in \oplus_y p_y \partial L(u, y)$, where $\oplus$ denotes the Minkowski sum (Hiriart-Urruty and Lemaréchal, 2012, Theorem 4.1.1). Moreover, ∂ represents the subdifferential $\partial f(x) = \{z : f(x') - f(x) \geq \langle z, x' - x\rangle \ \forall x'\}$. Expanding, we have $\oplus_y p_y \partial L(u, y) = \{\sum_{y \in \mathcal{Y}} p_y x_y : x_y \in \partial L(u, y) \ \forall y \in \mathcal{Y}\}$, and thus $Wp = \sum_y p_y x_y = \mathbf{0}$ where $W = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$; cf. (Ramaswamy and Agarwal, 2016, $\mathbf{A}^m$ in Theorem 16). Now taking $F := \ker(W) \cap \text{affhull}(\Delta_\mathcal{Y})$, we have $\text{codim}(F) = \text{rank}(W) \leq d$ by Definition **??**, and $p \in F$ by construction. To see $F \cap \Delta_\mathcal{Y} \subseteq \Gamma_u$, from the chain of equivalences above, we have for any $q \in \Delta_\mathcal{Y}$ that $q \in \ker W \implies u \in \Gamma(q) \implies q \in \Gamma_u$.

Now, we show $\Gamma_u \subseteq \gamma_r$, which will complete the proof. Let $r' = \psi(u)$; by definition of indirect elicitation, we have $\gamma_{r'} \supseteq \Gamma_u \supseteq F \cap \Delta_\mathcal{Y}$. If $|\gamma(p)| = 1$, then $r' = r$, so we are done. Otherwise, we have γ elictable and $p \in \text{relint}(\Delta_\mathcal{Y})$. Apply Lemma 26, which states that for elicitable properties with $p \in \gamma_r \cap \gamma_{r'}$, so $p \in F \cap \Delta_\mathcal{Y} \subseteq \gamma_{r'} \iff p \in F \cap \Delta_\mathcal{Y} \subseteq \gamma_r$. ■

5. Discrete-valued predictions

The main known technique for lower bounds on surrogate dimensions is given by Ramaswamy and Agarwal (2016) for the quadrant of a 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 $\mathcal{L}$, we show their bound follows relatively directly from Corollary 13. (We conjecture that the minimizability restriction to $\mathcal{L}$ can be lifted; see § 7.) Ramaswamy and Agarwal (2016) construct what they call the subspace of feasible dimensions and give bounds in terms of its dimension.

Definition 15 (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 (2016) gives a lower bound on the dimensionality of all calibrated convex surrogates, i.e. $\operatorname{cc}\dim(\ell) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1$ 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 Lemma 11. So, by making a slight restriction to the class of minimizable convex surrogates $\mathcal{L}$, we can derive this lower bound from our general technique in a way that we find shorter and simpler. Due to the infimum in the definition of calibration, the original proof requires careful taking of sequences and the approximate subdifferentials, while indirect elicitation allows us to work with exact minimizers of the surrogate.

Corollary 16 (Ramaswamy and Agarwal (2016) Theorem 18) *Let* $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ *be a discrete loss eliciting* $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$*, and let* $L \in \mathcal{L}_d$ *be a minimizable consistent convex surrogate for* ℓ*. Then for all* $p \in \Delta_{\mathcal{Y}}$ *and* $r \in \gamma(p)$*,*

$$d \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{6}$$

Proof [Proof sketch] [JF: Redid this citing Lemma 25 instead] First, take $p \in \operatorname{relint}(\Delta_{\mathcal{Y}})$ for intuition, so $\|p\|_0 = n$. Lemma 11 guarantees the existence of a flat F with a d-representation so that $p \in F \subseteq \gamma_r$. Moreover, since $p \in \operatorname{relint}(\Delta_{\mathcal{Y}})$, the flat F satisfies the requirements of Lemma 25, and the bound $\dim(\mathcal{S}_{\gamma_r}(p)) \geq \operatorname{span}(F - \{p\}) \geq n - d - 1$ is given by chaining parts (2.) and (3.) of the Lemma together.

When $p \notin \operatorname{relint}(\Delta_{\mathcal{Y}})$, we can loosely "project down" to the subsimplex on the support of p and modify L and ℓ accordingly. Now p is in the relative interior of this subsimplex, so the above result gives $\dim(\mathcal{S}_{\gamma_r}(p)) \geq \|p\|_0 - d - 1$, where the feasible subspace dimension is relative to $\mathbb{R}^{\operatorname{supp}(p)}$. Finally, we observe that 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 $\Delta_{\mathcal{Y}}$. ■

There are some cases where the bound provided by Corollaries 12 and 13 is strictly tighter than the bound provided by feasible subspace dimension in Corollary 16. For an example of how Corollary 12 applies to a discrete property for which there is no target loss – a non-elicitable property, i.e. a quadrant not considered by Ramaswamy et al. (2018) – we refer the reader to Appendix C.

Example: High-confidence classification. When given a target loss, we can consider the *abstain property* of Ramaswamy et al. (2018) over 3 outcomes where one wishes to predict the most likely outcome y if $Pr[Y = y|x] \geq 1/2$ and "abstain" by predicting $\perp$ otherwise. This is elicited by the target loss $\ell^{abs}(r, y) := \mathbf{I}\{r \notin \{y, \perp\}\} + (1/2)\mathbf{I}\{r = \perp\}$. 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 to be the uniform distribution (black dot in Figures 1 and 2). In this case, the bound by feasible subspace dimension yields $\operatorname{cc}\dim(\ell^{abs}) \geq 3 - 2 - 1 = 0$, as the feasible subspace dimension is 2 since we are on the relative interior of the level set and simplex, as shown in Figure 1.

RF: This is a bit of a strawman, since the FSD technique as a whole gives ≥ 1 here, for e.g. $p = (1/4, 1/4, 1/2)$; let's maybe discuss both p's for both techniques?

JF: Added both distributions.

However, consider any flat of with a 1-representation containing p. When intersected with the simplex, one can see that any line in the simplex through p also leaves the cell $\gamma_\perp$, which contains p. See Figure 2 for intuition; a flat with a 1-representation would be a line in such a figure. Therefore, we have no flat with a 1-representation containing p staying in $\gamma_\perp$, so we obtain better lower bound, $\operatorname{cc}\dim(\ell^{abs}) \geq 3 - 0 - 1 = 2$.

For a case where our bounds match that of (Ramaswamy and Agarwal, 2016), consider the distribution $q = (1/4, 1/4, 1/2)$, shown in blue in Figures 1 and 2. The feasible subspace dimension (of both $\gamma_\perp$ and γ_3) 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 affhull($\Delta_\mathcal{Y}$). This yields $\operatorname{cc}\dim(\ell^{abs}) \geq 3 - 1 - 1 = 1$. The same line segment defines a flat contained in both $\gamma_\perp$ and γ_3, so we have $\operatorname{cc}\dim(\ell^{abs}) \geq 3 - 1 - 1 = 1$ by Corollary 13, matching the feasible subspace dimension bound.

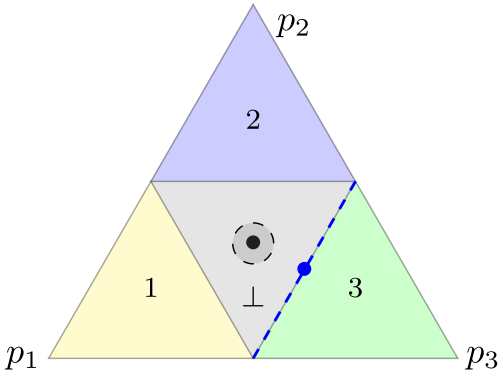

Figure 1: $\dim(\mathcal{S}_{\gamma_\perp}(\bullet)) = 2$ and $\dim(\mathcal{S}_{\gamma_\perp}(\bullet)) = 1$

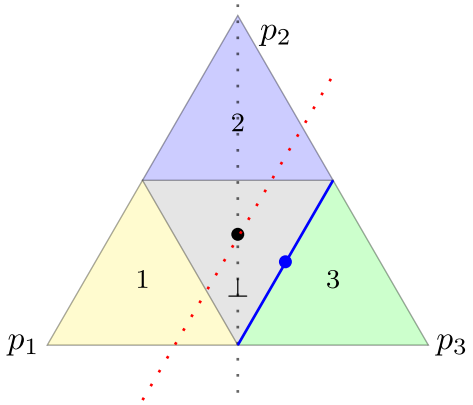

Figure 2: Any flat F with a 1-representation through $\bullet$ also leaves the cell $\gamma_\perp$.

6. Continuous-valued predictions

In continuous estimation problems, often one is not given a target loss, but instead a target (conditional) statistic of the data one wishes to estimate, such as the mean or variance. In this setting, Theorem 14 [JF: Change ref] gives lower bounds on the prediction dimension of convex losses with a link to the desired conditional statistic, i.e., the convex elicitation

complexity. In particular, Theorem 17 below yields new bounds on the convex elicitation complexity of statistics which quantify risk or uncertainty such as variance, entropy, or financial risk measures.

These bounds address an open question of Frongillo and Kash (2020), that of developing a theory of elicitation complexity for with respect to convex-elicitable properties. The lower bounds of that work are essentially all with respect to identifiable properties, that is, properties γ for which there is some d where γ_r is a flat with a d-representation for all $r \in \mathcal{R}$. In contrast, properties elicited by non-smooth convex losses are generally not identifiable. For example, the properties elicited by hinge loss and the abstain surrogate are not identifiable, as their level sets are not flats (see Figure 1). Establishing a theory of elicitation complexity for convex-elicitable properties therefore seemed to require entirely new ideas. Surprisingly, we show that the central techniques of Frongillo and Kash (2020) also apply to convex-elicitable properties, via [RF: MAIN THM]. In particular, we can recover their main lower bound for the large class of Bayes risks.

Theorem 17 *Let L elicit some $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$. Let $p \in \operatorname{relint}(\Delta_\mathcal{Y})$ and let Γ_r be some level set of Γ such that (i) $p \in \Gamma_r$, (ii) $\operatorname{codim}(\operatorname{affhull}(\Gamma_r)) = d$, and (iii) either Γ_r is a singleton or $\underline{L} := \inf_u \mathbb{E}_p L(u, Y)$ is nonconstant on Γ_r. Then $\operatorname{elic}_{\text{cvx}}(\underline{L}) \geq \min(d+1, n-1)$.*

Proof [Proof sketch; full proof in Appendix B] We roughly follow the argument of Frongillo and Kash (2018, Corollary 7). To indirectly elicit $\underline{L}$, we must link from a loss $\hat{L} : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$. By Frongillo and Kash (2018, Theorem 4), the property $\hat{\Gamma} = \operatorname{prop}[\hat{L}]$ refines Γ, in the sense that every level set of $\hat{\Gamma}$ is contained in a level set of Γ. If $\underline{L}$ is non-constant (the interesting case) on the level set Γ_r containing p, then Γ_r must strictly contain some level set $\hat{\Gamma}_{\hat{r}}$ containing p. But by Theorem 2, there is a flat $\hat{F}$ containing p of codimension at most k, yet strictly contained in $\operatorname{affhull}(\Gamma_r)$ of codimension d. So $k \geq d+1$. ■

Example: Variance. [JF: Keep this] As a warm-up, let us see how to show $\operatorname{elic}_{\text{cvx}}(\text{Var}) = 2$ whenever $|\mathcal{Y}| \geq 3$, meaning the lowest dimension of a convex loss to estimate conditional variance is 2 (Osband, 1985; Lambert, 2018). It is interesting to note that, to the best of our knowledge, even this simple bound is novel. While intuitively obvious, the lower bound of 2 is not trivial. In particular, the well-known fact that the variance is not elicitable does not yield this lower bound, as it does not rule out the variance being a link of a real-valued convex-elicitable property; cf. Frongillo and Kash (2018, Remark 1).

Let $\mathcal{Y} \subseteq \mathbb{R}$ be a finite set with $n = |\mathcal{Y}|$. Observe that the variance of Y is the Bayes risk of squared loss $L(r, y) = (r - y)^2$; as L elicits the mean $\Gamma(p) = \{\mathbb{E}_p[Y]\}$, we have $\text{Var}_p[Y] = \{\mathbb{E}_p(\mathbb{E}_p Y - Y)^2\} = \{\min_{r \in \mathbb{R}} \mathbb{E}_p L(r, Y)\}$. To apply Theorem 17, we take this L and Γ and $d = 1$, and must choose an appropriate level set Γ_r; we choose $p = \frac{1}{n}\mathbb{1}$ to be the uniform distribution and let $r = \mathbb{E}_p Y$. We summarize three conditions here, leaving the full details to Appendix B.4: (i) $p \in \Gamma_r$ by construction. (ii) Letting $v \in \mathbb{R}^\mathcal{Y}$ with $v_y = y - r$, the flat $F = \ker W \cap \operatorname{affhull}(\Delta_\mathcal{Y})$ for $W = [v] \in \mathbb{R}^{1 \times n}$ contains Γ_r, and has $\operatorname{codim}(\operatorname{affhull}(\Gamma_r)) = \operatorname{rank}(W) = 1 = d$. (iii) $\Gamma_r = \{p\}$ is a singleton for $n \leq 2 = d + 1$, and for $n \geq 3$ there are enough degrees of freedom in $\Delta_\mathcal{Y}$ for there to be two distributions with mean r and different variances.

Applying Theorem 17 gives $\operatorname{elic}_{\text{cvx}}(\text{Var}) \geq \min(2, n-1)$, as desired. In fact, this bound is tight for all n. For $n = 1$ there is only one distribution and we have complexity 0; for

$n = 2$ the mean itself, elicited by squared loss, determines the distribution; for $n \geq 3$ we may elicit the first two moments via the convex $L(r, y) = (r_1 - y)^2 + (r_2 - y^2)^2$, and recover the variance via $\psi(r) = r_2 - r_1^2$.

Example: Entropy and Norms. [JF: Move to Appendix?] For another application, let us see how Theorem 17 implies $\text{elic}_{\text{cvx}}(G) = n - 1$ for any strictly concave function $G : \Delta_{\mathcal{Y}} \to \mathbb{R}$ of the distribution, including most entropy functions. In other words, there is no convex loss function allowing one to consistently estimate conditional entropy using fewer dimensions than required to estimate the conditional distribution itself. This observation also extends to norms; given any $k > 1$, $G(p) = \|p\|_k^k$ is strictly convex, and hence $\text{elic}_{\text{cvx}}(\|\cdot\|_k) = n - 1$, as otherwise we could link to G via $\psi(r) = r^k$. These results illustrate the power of our technique.

To show the bound, recall that G is the Bayes risk of a proper loss defined by $L(p, y) = G(p) + s_p \cdot (\delta_y - p)$, where $\delta_y(y') = \mathbb{1}\{y = y'\}$ is the point distribution on y and $-s_p \in \partial(-G)(p)$ is a subgradient of $-G$ at p (Gneiting and Raftery, 2007; Reid and Williamson, 2009; Frongillo and Kash, 2014). Intuitively, since L elicits the identity property $\Gamma(p) = p$, Theorem 17 should therefore give us a maximal lower bound $(n - 1)$, as the level sets of are Γ have codimension $n - 1$. For technical reasons, however, we need to drop a dimension from the prediction space, e.g. via the bijection $\varphi(p) = (p_1, \ldots, p_{n-1})$, defining $L'(p', y) = L(\varphi^{-1}(p'), y)$ which elicits $\Gamma'(p) = \varphi(p)$ but still has Bayes risk G. Now the conditions of Theorem 17 are easily checked, where again $p \in \text{relint}(\Delta_{\mathcal{Y}})$ is the uniform distribution and $r = \varphi(p)$: (i) is trivial, (ii) $\text{codim}(\text{affhull}(\Gamma'_r)) = \text{codim}(\{p\}) = n - 1$, and (iii) Γ'_r is a singleton. As with the variance example, the lower bound is easily matched, e.g. by $L_2(p', y) = \|\varphi^{-1}(p') - \delta_y\|_2^2$, which is convex in p', and the link $\psi(p') = G(\varphi^{-1}(p'))$.

7. Conclusions and future work

In this work, we show that indirect property elicitation can be a powerful necessary condition for the existence of a consistent surrogate loss (Theorem 9). Furthermore, we introduce a new lower bound (Corollaries 12 and 13) on the dimension of a consistent convex loss that is generally applicable and extends previous results from both the discrete and continuous estimation settings.

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. For continuous settings, it would be especially interesting to extend to the case $\mathcal{Y} = \mathbb{R}$, which would require a more careful analysis in Theorem 17.[JF: Flagging this nc we take care of it I hope] Finally, of course, we would like to characterize $\text{cc}\dim$ and elic_{cvx} and develop a general framework for constructing surrogates achieving the best possible prediction dimension. [JF: for neurips, we had commented out everything below this for space] One might be able to further tighten these bounds by property elicitation by studying monotonicity and adjacency of level sets. In discrete predictions, these bounds might also be tightened if the equivalence of convex calibration dimension and embedding dimension of Finocchiaro et al. (2020) is shown; the current embedding dimension bounds are not tight either, but additional structure is imposed by considering the embedding framework. Moreover, the practical reason why consistency is desired is to ensure the

guarantee of empirical risk minimization (ERM) rates; however, the relationship between ERM rates and property elicitation has not been studied.

References

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.

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.

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.

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

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

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

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

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

Rafael Frongillo and Ian A Kash. Elicitation complexity of statistical properties. *arXiv preprint arXiv:1506.07212v2*, 2018.

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.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

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

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

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

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

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.

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.

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

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.

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

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

Harish G Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2078–2086. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4528-classification-calibration-dimension-for-general-multiclass-losses.pdf.

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

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.

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

R. T. Rockafellar and R. J. B. Wets. On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics*, 7(3):173–182, 1982. doi: 10.1080/17442508208833217. URL https://doi.org/10.1080/17442508208833217.

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.

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

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

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

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.

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

Appendix A. A general notion of calibration

For general settings, we introduce a notion of calibration that is a special case of calibration as introduced by (Steinwart, 2007, Definition 2.7) and Steinwart and Christmann (2008, Chapter 3). [JF: I think we need to be careful how we present this; I'm not sure it's a new definition, though we seem to prove new results about it.] We will show that in discrete prediction settings, it is equivalent to the more commonplace definition given in Definition 1. Therefore, we use this more general definition of calibration when proving statements about the relationship between consistency, calibration, and indirect elicitation.

Definition 18 (Calibrated) *A loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is* calibrated *with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting the property γ [JF: γ isn't used here...?] if there is a link $\psi : \mathbb{R}^d \to \mathcal{R}$ such that, for all distributions $p \in \Delta_\mathcal{Y}$, 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*

$$\ell(\psi(u); p) - \underline{\ell}(p) \leq \zeta\left(\mathbb{E}_p L(u, Y) - \underline{L}(p)\right) \ . \tag{7}$$

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 $p \in \Delta_\mathcal{Y}$ 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 18, and we show $A \iff B$ in Lemma 20. Lemma 21 shows calibration if and only if D, which yields a condition equivalent to calibration without dependence the function ζ.

Proposition 19 *When $\mathcal{R}$ and $\mathcal{Y}$ are finite, a continuous loss and link (L, ψ) are calibrated with respect to a target loss ℓ via Definition 18 if and only if they are calibrated via Definition 1. BTW: Okay if Γ is empty*

Proof $\implies$ We prove the contrapositive; if (L, ψ) is not calibrated with respect to ℓ by Definition 1, then it is not calibrated via Definition 18 either. If (L, ψ) are not calibrated with respect to ℓ by Definition 1, then there is a $p \in \Delta_\mathcal{Y}$ so that $\inf_{u:\psi(u)\notin\gamma(p)} \mathbb{E}_p L(u, Y) = \inf_u \mathbb{E}_p L(u, Y)$. Thus there is a sequence $\{u_m\}$ so that $\lim_{m\to\infty} \psi(u_m) \notin \gamma(p)$ 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 21, we contradict calibration by Definition 18.

$\impliedby$ Suppose there was a function ζ satisfying the bound in Equation (7) for a fixed distribution $p \in \Delta_\mathcal{Y}$. Observe the bound in Equation (1) can be written as $R_L(u, p) > 0$ for all $p \in \Delta_\mathcal{Y}$ and u such that $\psi(u)$ is bounded away from $\gamma(p)$. [JF: details correct??]

By Equation (7), 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 Equation (7) since $R_\ell(\psi(u), p) \not\to 0$. Therefore $R_L(u_m, p) \not\to 0$; thus, the strict inequality holds. ■

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 21

Lemma 20 *A function $\zeta : \mathbb{R} \to \mathbb{R}$ is continuous at 0 and $\zeta(0) = 0$ if and only if the sequence $\{u_m\} \to 0 \implies \zeta(u_m) \to 0$. [JF: A $\iff$ B]*

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 $\zeta(u_m) \to 0$.

$\impliedby$ 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 $\zeta(u_m) \not\to 0$. ■

Lemma 21 now gives a condition equivalent to calibration without requiring one to already have a function ζ in mind.

Lemma 21 *A continuous surrogate and link (L, ψ) are calibrated (via definition 18) with respect to ℓ if and only if, for all $p \in \Delta_\mathcal{Y}$ and sequences $\{u_m\}$ so that $R_L(u_m; p) \to 0$, we have $R_\ell(\psi(u_m); p) \to 0$. [JF: $(A \wedge C) \iff D$]*

Proof [JF: $(A \wedge C) \implies D$] $\implies$ 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 (7) is satisfied for all $u \in \mathbb{R}^d$ by assumption, we observe

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

$\impliedby$ [JF: $D \implies (A \wedge C)$] Fix $p \in \Delta_\mathcal{Y}$, 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 (7) is satisfied for all $u \in \mathbb{R}^d$ and apply Lemma 20 to observe that if there is a sequence $\{\epsilon_m\} \to 0$ so that $\zeta(\epsilon_m) \not\to 0$, it is because $R_L(u_m, p) \not\to 0 \not\iff R_\ell(\psi(u_m), p) \to 0$.

[JF: D $\implies$ C] Now, we observe that the bound in Equation (7) 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.

[JF: Not B leads to contradiction of D.] 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, BTW: really just need lower semi-continuity and boundedness from below since we can take the constant sequence at the (attained) infimum. ■

A.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 22 *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 21. 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. ■

Moreover, we have calibration implying indirect elicitation.

Lemma 23 *If a surrogate and link (L, ψ) are calibrated with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$, then L indirectly elicits the property $\gamma := \text{prop}[\ell]$.*

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. ■

Combining the two results, we can observe the result of Theorem 9 another way: *through calibration.*

Appendix B. Omitted Proofs

B.1. Coping with links for elicitable set-valued properties

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

Lemma 24 *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 BTW: Bo: the proof holds as written for the case where the level sets have empty intersection. 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 $\gamma_r \cap H = \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'}$. ■

Lemma 25 *Let the d-flat $F \subseteq \mathcal{P}$ contain some $p \in \text{relint}(\mathcal{P})$. Then*

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

(ii) $\dim(\text{span}(F - \{p\})) \geq \dim(\text{span}(\mathcal{P} - \{p\})) - d$;

(iii) Given a discrete elicitable property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ and $r \in \gamma(p)$, we have $F - \{p\}$ contained in $\mathcal{S}_{\gamma_r}(p)$.

Proof

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

(ii) Let $p \in F$, and define $F_p := F - \{p\}$, $\mathcal{P}_p := \mathcal{P} - \{p\}$ and $T_W(v) = \mathbb{E}_{Y \sim v} W(Y)$ for the function $W : \mathcal{Y} \to \mathbb{R}^d$ such that $F = \ker_{\mathcal{P}} W$. (Such a function must exist by F being a d-representation.)

Now let $v \in \ker(T_W) = \{v \in \text{span}(\mathcal{P}_p) : T_W(v) = \mathbf{0}\}$. There must exist an $\epsilon > 0$ so that $\epsilon v \in F_p$ and $-\epsilon v \in F_p$ $p \in \text{relint}(\mathcal{P})$ implies $\mathbf{0} \in \text{relint}(F_p)$ by (1.). Moreover, we have $\pm\epsilon v \in \ker(T_W)$ by linearity of expectation, so $v \in \text{relint}(\ker(T_W))$. This allows us to conclude $\text{span}(F_p) = \ker(T_W)$.[JF: Right?]

We want to show $F_p = \mathcal{P}_p \cap \ker(T_W)$, then derive the result from there. [JF: I don't see why this is necessary.]

In order to see this, consider $v \in F_p \iff v + p \in F$, and similarly for $\mathcal{P}_p$. Now, as $F \subseteq \mathcal{P}$ by construction, we have $F_p \subseteq \mathcal{P}_p$, and since $\text{span}(F_p) = \ker(T_W)$, we additionally have $F_p \subseteq \ker(T_W)$, yielding $F_p \subseteq \mathcal{P}_p \cap \ker(T_W)$. Now consider $v \in \mathcal{P}_p \cap \ker(T_W)$; we know that $\mathbb{E}_v W = \mathbf{0}$, and therefore, $\mathbb{E}_q W = \mathbf{0}$, so $v \in F_p$, thus $F_p = \mathcal{P}_p \cap \ker(T_W)$.

Since $\text{span}(F_p) = \ker(T_W)$, we have $\dim(\text{span}(F_p)) = \dim(\ker(T_W)) \geq \dim(\text{span}(\mathcal{P}_p)) - d$ by [JF: isomorphism theorem] since $F_p = \ker(T_W) \cap \mathcal{P}_p$. [JF: ??]

(iii) We aim to show $v \in F_p \implies v \in \mathcal{S}_{\gamma_r}(p)$. By (i), we know $p \in \text{relint}(F)$, and therefore, $\mathbf{0} \in \text{relint}(F_p)$. Consider $v \in F_p$. By F_p being convex and $\mathbf{0} \in \text{relint}(F_p)$, we know there exists an $\alpha > 0$ so that $-\alpha v \in F_p$. This implies $p - \alpha v \in F \implies p - \alpha v \in \gamma_r \implies v \in \mathcal{S}_{\gamma_r}(p)$ since F and $\mathcal{S}_{\gamma_r}(p)$ are convex and by construction of F as the kernel space of a subgradient of a convex loss This yields the desired result, $F_p \subseteq \mathcal{S}_{\gamma_r}(p)$. ■

JF: We need the loss setup here and the loss to be convex... makes me think this last part should be pulled out into its own statement.

Lemma 26 *Suppose we are given an elicitable property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, where $\mathcal{Y}$ is finite, and distribution $p \in \text{relint}(\mathcal{P})$ 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 24 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 show $F \subseteq \gamma_{r'}$. Let $q \in F$. We claim that for small enough ϵ, the point $q' = p - \epsilon(q - p)$ is in F as well. Containment in F follows because both p and q are in F and it is a flat, while containment in $\mathcal{P}$ follows because $p \in \text{relint}(\mathcal{P})$. [JF: Does relint($F$) actually come in here?]

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. ■

B.2. Reconstructing the result of Ramaswamy and Agarwal (2016, Theorem 16)

The next result helps us generalize Lemma 25 to any p; not just to $p \in \text{relint}(\mathcal{P})$.[JF: Just for their corollary though...]

Lemma 27 *For any $p \in \mathcal{P}$ and r such that $p \in \gamma_r$, take $\mathcal{Y}' := \text{supp}(p)$. Define $\gamma' : \mathcal{P} \rightrightarrows \mathcal{R}$ with $\gamma' : q \mapsto \gamma(q) \cap \Delta_{\mathcal{Y}'}$. Then $\mathcal{S}_{\gamma_r}(p) = \mathcal{S}_{\gamma'_r}(p)$.*

Proof Consider the ambient space of both $\mathcal{S}_{\gamma_r}(p)$ and $\mathcal{S}_{\gamma'_r}(p)$ is $\mathbb{R}^{\mathcal{Y}}$. We trivially have $\dim(\mathcal{S}_{\gamma_r}(p)) \geq \dim(\mathcal{S}_{\gamma'_r}(p))$ since γ' is simply γ projected down to an affine subspace of $\mathbb{R}^{\mathcal{Y}}$.

Now to see $\dim(\mathcal{S}_{\gamma_r}(p)) \leq \dim(\mathcal{S}_{\gamma'_r}(p))$, it suffices to show subset inclusion. Take some $v \notin \mathcal{S}_{\gamma'_r}(p)$. Observe that $\gamma'_r = \gamma_r \cap \Delta_{\mathcal{Y}'}$, so if $q^{\pm} := p \pm \epsilon v \notin \gamma'_r$ for all $\epsilon > 0$ (i.e. $p \notin \text{relint}(\Delta_{\mathcal{Y}'})$), it is either because one of q^+ or q^- is not in γ_r or because either $q^{\pm} \notin \Delta_{\mathcal{Y}'}$. The first case can be seen easily by the definition of γ', and the latter can be seen because leaving $\Delta_{\mathcal{Y}'}$ means one of $q^{\pm}$ also is not in $\mathcal{P}$, and therefore not in γ_r as $\gamma_r \subseteq \mathcal{P}$. Thus $\mathcal{S}_{\gamma_r}(p) = \mathcal{S}_{\gamma'_r}(p)$. ■

Corollary 28 (Ramaswamy and Agarwal (2016) Theorem 18) *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete loss eliciting $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, and let $L \in \mathcal{L}_d$ be a minimizable consistent convex surrogate for ℓ. Then for all $p \in \Delta_{\mathcal{Y}}$ and $r \in \gamma(p)$,*

$$d \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{6}$$

Proof Let $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ be a calibrated surrogate for ℓ, and consider $\mathcal{Y}' := \text{supp}(p)$ and what happens when we restrict L and ℓ to only the outcomes in $\mathcal{Y}'$. Take $L' := L|_{\mathcal{Y}'}$ and $\ell' := \ell|_{\mathcal{Y}'}$.

First, observe L' (eliciting Γ) indirectly elicits $\gamma' := \text{prop}[\ell']$ since, for all $p \in \Delta_{\mathcal{Y}}$, and therefore all $p \in \Delta_{\mathcal{Y}'}$, we have $p \in \Gamma_u \implies p \in \gamma_{\psi(u)}$, and $\Gamma(p) = \Gamma'(p)$ and $\gamma(p) = \gamma'(p)$

for $p \in \Delta_{\mathcal{Y}'}$. This same observation can be used to observe that L' is also calibrated with respect to ℓ' as the calibration bound holds for all $p \in \Delta_{\mathcal{Y}}$, and therefore for all $p \in \Delta_{\mathcal{Y}'}$ by equality of γ and γ' for $p \in \Delta_{\mathcal{Y}'}$.

As L' is calibrated (and consistent) with respect to ℓ' and indirectly elicits $\gamma' := \text{prop}[\ell']$, then by Corollary 13, we know there exists a flat F with $p \in F \subseteq \gamma'_r$ that is a d-representation.

When we substitute $\mathcal{P} = \Delta_{\mathcal{Y}'}$ into Lemma 25 part (2.), we yield $\dim(\text{span}(F_p)) \geq \|p\|_0 - 1 - d$. Moreover, we can observe that $F_p \subseteq \mathcal{S}_{\gamma'_r}(p)$ since $p \in \text{relint}(F)$ by Lemma 25 part (3.). By Lemma 27, we additionally have $\mathcal{S}_{\gamma_r}(p) = \mathcal{S}_{\gamma'_r}(p)$. Chaining these results, we obtain

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

■

B.3. Proof of Theorem 17

Lemma 29 ((Frongillo and Kash, 2018)) *Suppose the loss L elicits a single-valued property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. Let $\underline{L}$ be the Bayes risk of L. Then for any $p, p' \in \Delta_{\mathcal{Y}}$ 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 30 *Let $C \subseteq \mathbb{R}^m$ be convex. Then for any affine subspace $F \subseteq \text{affhull}(C)$ with $F \cap \text{relint}(C) \neq \emptyset$, we have $\text{affhull}(F \cap C) = F$.*

Proof As $\text{affhull}(F) = F$ and $F \cap C \subseteq F$, the inclusion $\text{affhull}(F \cap C) \subseteq F$ is clear. For the reverse, let $p \in F \cap \text{relint}(C)$ and let $B \subseteq C$ be a relatively open set containing p. For any $q \in F \subseteq \text{affhull}(C)$, we thus have $q' = p + \epsilon(q - p) \in B$ for sufficiently small $\epsilon > 0$. As $q' = (1-\epsilon)p + \epsilon q$, we have $q' \in \text{affhull}(F) \cap C = F \cap C$. As $q = (1 - 1/\epsilon)p + (1/\epsilon)q'$, we thus have $q \in \text{affhull}(F \cap C)$. ■

Theorem 17 *Let L elicit some $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}^d$. Let $p \in \text{relint}(\Delta_{\mathcal{Y}})$ and let Γ_r be some level set of Γ such that (i) $p \in \Gamma_r$, (ii) $\text{codim}(\text{affhull}(\Gamma_r)) = d$, and (iii) either Γ_r is a singleton or $\underline{L} := \inf_u \mathbb{E}_p L(u, Y)$ is nonconstant on Γ_r. Then $\text{elic}_{\text{cvx}}(\underline{L}) \geq \min(d+1, n-1)$.*

Proof Suppose that we have some convex loss $\hat{L} : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ eliciting a property $\hat{\Gamma} : \Delta_{\mathcal{Y}} \to \mathbb{R}^k$, and link $\psi : \mathbb{R}^k \to \mathbb{R}$ such that $\underline{L} = \psi \circ \hat{\Gamma}$. The proof of Frongillo and Kash (2018, Theorem 4) argues that $\hat{\Gamma}$ must refine Γ, in the sense that every level set of $\hat{\Gamma}$ is contained in a level set of Γ; for completeness we give the argument here. Suppose for a contradiction that we have p, p' with $\hat{\Gamma}(p) = \hat{\Gamma}(p')$ but $\Gamma(p) \neq \Gamma(p')$. As $\underline{L} = \psi \circ \hat{\Gamma}$, we also have $\underline{L}(p) = \underline{L}(p')$. Letting $p'' = \frac{1}{2}p + \frac{1}{2}p'$, Lemma 29 would then give us $\underline{L}(p'') > \frac{1}{2}\underline{L}(p) + \frac{1}{2}\underline{L}(p') = \underline{L}(p)$. By Osband (1985), the level sets $\hat{\Gamma}_{\hat{r}}$ are convex, giving $\hat{\Gamma}(p'') = \hat{\Gamma}(p)$, which would imply $\underline{L}(p'') = \underline{L}(p)$, contradicting $\underline{L} = \psi \circ \hat{\Gamma}$. We conclude $\hat{\Gamma}$ must refine Γ, and thus $\hat{L}$ actually indirectly elicits Γ through some other link function.

In particular, we have $p \in \hat{\Gamma}_{\hat{r}} \subseteq \Gamma_r$ for some $\hat{r}, r$. By Theorem 14, there is a flat $\hat{F} \subseteq \text{affhull}(\Delta_{\mathcal{Y}})$ containing p such that $\text{codim}(\hat{F}) \leq k$ and $S := \hat{F} \cap \Delta_{\mathcal{Y}} \subseteq \hat{\Gamma}_{\hat{r}} \subseteq \Gamma_r$. As $p \in \hat{F} \cap \text{relint}(\Delta_{\mathcal{Y}})$, Lemma 30 gives $\text{affhull}(S) = \hat{F}$. Let $F = \text{affhull}(\Gamma_r)$ and recall

$\operatorname{codim}(F) = d$. Now as $S \subseteq \Gamma_r$, we also have $\hat{F} = \operatorname{affhull}(S) \subseteq \operatorname{affhull}(\Gamma_r) = F$, implying $\operatorname{codim}(\hat{F}) \geq \operatorname{codim}(F)$.

If $\Gamma_r = \{p\}$ is a singleton, then $F = \operatorname{affhull}(\Gamma_r) = \{p\}$, and in particular $n - 1 = \operatorname{codim}(F) \leq \operatorname{codim}(\hat{F})$, so we must have $k = d = n - 1$. If Γ_r is not a singleton, then $\underline{L}$ is non constant on Γ_r. But by definition, $\underline{L}$ is constant on $\hat{\Gamma}_{\hat{r}}$, so the containment must be strict, in particular, $S \subsetneq \Gamma_r$. So $\hat{F} \subsetneq F$, and both are flats, so $\operatorname{codim}(\hat{F}) > \operatorname{codim}(F)$. In other words, $k \geq d - 1$. ■

B.4. Variance example

We justify the three conditions of Theorem 17:

(i) We have $p \in \Gamma_r$ by construction.

(ii) Let $\{y_1, \ldots, y_n\} = \mathcal{Y}$ be the n distinct outcome/label values. Letting $v \in \mathbb{R}^n$ with $v_i = y_i - r$, define $F = \ker W \cap \operatorname{affhull}(\Delta_{\mathcal{Y}})$ for $W = [v] \in \mathbb{R}^{1 \times n}$. Note that $\Gamma_r \subseteq F$, and $\operatorname{rank}(W) = 1$ as the y_i are distinct. As $p \in F$, Lemma 30 gives $\operatorname{affhull}(\Gamma_r) = F$ and thus $\operatorname{codim}(\operatorname{affhull}(\Gamma_r)) = \operatorname{rank}(W) = 1 = d$.

(iii) For $n \leq 2 = d + 1$, $\Gamma_r = \{p\}$ is a singleton and we are done; otherwise $n \geq 3$. If $\operatorname{Var}[Y]$ were constant in Γ_r, then we would have some $c \in \mathbb{R}$ such that $c = \operatorname{Var}_{p'}[Y] = \mathbb{E}_{p'}[Y^2] - r^2$ for all $p' \in \Gamma_r$. Letting $W' = [v; v'] \in \mathbb{R}^{2 \times n}$ where $v'_y = y^2 - r^2 - c$, and $F' = \ker W' \cap \operatorname{affhull}(\Delta_{\mathcal{Y}})$, this would imply $\Gamma_r = F' \cap \Delta_{\mathcal{Y}}$ as well. Lemma 30 applies again to show $F' = \operatorname{affhull}(\Gamma_r) = F$. Yet as the y values are distinct, $\operatorname{rank}(W') = 2$ (for any $\alpha \in \mathbb{R}$ there are at most two solutions to $y - r = \alpha(y^2 - r^2 - c)$), contradicting $\operatorname{codim}(F) = 1$. Thus $\operatorname{Var}[Y]$ cannot be constant on Γ_r.

Appendix C. Omitted examples

Discrete problem with no target loss. Consider the following scenario where someone is deciding how to dress for the weather based on a meteorologist's forecast. Consider the three outcomes $\mathcal{Y} = \{\text{sunny, snowy, rainy}\}$, and we suppose we want to have some bias towards health and safety, so the meteorologist should only predict sunny weather if $Pr[\text{sunny} \text{ — weather data}] \geq 3/4$. Otherwise, they should predict whatever is more likely given the weather data: rain or snow.

We can now model this problem by a property with the reports $\mathcal{R} = \mathcal{Y}$, and have

$$\gamma(p) = \begin{cases} \text{sunny} & p_{\text{sunny}} \geq 3/4 \\ \text{rainy} & p_{\text{sunny}} \leq 3/4 \wedge p_{\text{rainy}} \geq p_{\text{snowy}} \\ \text{snowy} & p_{\text{sunny}} \leq 3/4 \wedge p_{\text{snowy}} \geq p_{\text{rainy}} \end{cases},$$

shown in Figure 3. Since the cells of elicitable properties in the simplex form a power diagram (Lambert and Shoham, 2009), we know that there is actually *no* target loss that directly elicits this problem. Constructing a consistent surrogate for this task is ill-defined without Definition 7. The function $\mu(r, p) = \mathbb{1}\{r \notin \gamma(p)\}$, which satisfies the requirements of μ, now allows us to use Definition 7 to think about consistent surrogates for this task.

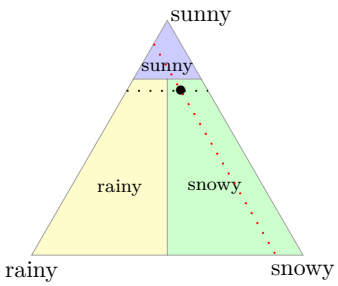

Figure 3: Our meteorology example with a bias towards citizen safety.

Intuitively, since the feasible subspace dimension bound would be lowest at the distribution $p = (1/8, 3/4, 1/8)$, we might want to test our Theorem 14 at p. However, we cannot apply Theorem 14 at p since $\gamma(p) = \{\text{rainy, snowy, sunny}\}$. Ramaswamy and Agarwal (2016, Theorem 16) cannot draw any conclusions about this property for two reasons that go hand in hand: first, we are given a target property instead of a target loss. Second, since the property is not elicitable (hence why there can be no target loss), we observe $\dim(\mathcal{S}_{\gamma_{\text{rainy}}}(p)) \neq \dim(\mathcal{S}_{\gamma_{\text{sunny}}}(p))$, so their result cannot be applied as Ramaswamy and Agarwal (2016, Lemma 23) does not hold.

However, our bounds from Theorem 14 on the distribution $q = (1/8, 3/4 - \epsilon, 1/8 + \epsilon)$ for a small enough $\epsilon > 0$, which we can apply since $\gamma(q) = \{\text{snowy}\}$, suggest that the convex elicitation complexity $\text{elic}_{\text{cvx}}(\gamma) \geq 3 - 0 - 1 = 2$, since there is no way to draw a line through q while staying in just one level set on the simplex.

This example, although seemingly contrived, also extends to other decision-tree-like properties that do not have an explicit or easily constructed target loss.

Appendix D. Extensions to infinite $\mathcal{Y}$

The proof for Theorem 14 only requires slight modifications for infinite $\mathcal{Y}$; we restate such definitions here and generalize the proof.

Definition 31 (Flat; infinite $\mathcal{Y}$) *A flat $F \subseteq \text{span}(\mathcal{P})$ is a nonempty set $F := \{p \in \text{span}(\mathcal{P}) : \int \nu(y) d(py) = 0\}$, for some measurable $\nu : \mathcal{Y} \to \mathbb{R}^d$. The codimension of F is given by $\text{codim}(F) = \dim(\text{span}(P)/F) = \dim(\{p + F : p \in \text{span}(\mathcal{P})\}) \leq d$ by design of ν.*

We additionally have to generalize the definition of $\mathbb{E}_p \partial L(u, Y) := \mathbf{cl}(\{\int \nu(y) p(dy) : \nu \in \mathcal{L}_d^1(\mathcal{A}), \nu(y) \in \partial L(u, Y) \text{ a.s.}\})$, where $\mathcal{L}_d^1(\mathcal{A})$ is the set of Lebesgue measurable functions defined on the σ-algebra $\mathcal{A}$.

Theorem 32 (Theorem 14 with infinite $\mathcal{Y}$) *Let a single-valued property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ be given, and let $p \in \mathcal{P}$, and $r \in \gamma(p)$. Let $L \in \mathcal{L}_d$ be a convex loss[2] such that such that $\mathbb{E}_p L(\cdot, Y)$ is finite for all $u \in \mathbb{R}^d$ and $p \in \mathcal{P}$. If L indirectly elicits γ, there is a flat F with $p \in F \cap \mathcal{P} \subseteq \gamma_r$ with $\text{codim}(F) \leq d$.*

2. $\mathbb{E}_p L(u, Y)$ must be a $\mathcal{A}$-convex normal integrand for all $p \in P$

Proof Consider the probability space $(\mathcal{Y}, \mathcal{A}, p)$ so that $\mathbb{E}_p L(u, Y) = \int L(u, y) d(py)$ is well-defined and finite for all $u \in \mathbb{R}^d$, and $p \in \mathcal{P}$. Moreover, let $\Gamma := \text{prop}[L]$ and suppose (L, ψ) indirectly elicits γ. By definition of properties, there is some $u \in \Gamma(p)$. [JF: Address if we get rid of minnability]

We first construct the flat F and show $F \cap \mathcal{P} \subseteq \Gamma_u$, which implies $F \cap \mathcal{P} \subseteq \gamma_r$ since we must have $u = \psi(r)$ by indirect elicitation and γ being single-valued.

As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \mathbf{0} \in \partial \mathbb{E}_p L(u, Y)$. Since, for each $p \in \mathcal{P}$, we have L satisfying the assumptions of (Rockafellar and Wets, 1982, Corollary 1), we can interchange subdifferentiation and expectation to observe $\mathbf{0} \in \partial \mathbb{E}_p L(u, Y) \iff \mathbf{0} \in \mathbb{E}_p \partial L(u, Y)$. As $\mathbf{0}$ is in the expectation of the subgradient, such a ν exists, so we take $F := \{p \in \text{span}(\mathcal{P}) : \int \nu(y) d(py) = \mathbf{0}\} \cap \mathcal{P}$. [JF: Want F to be an affine space; intersect with an affine hull]

By construction we have $p \in F \implies \int \nu(y) d(py) = \mathbf{0} \implies \mathbf{0} \in \mathbb{E}_p \partial L(u, Y) \implies p \in \Gamma_u$. Moreover, we $\text{codim}(F) \leq d$ by definition. ■