JMLR: Workshop and Conference Proceedings vol 49:1–26, 2016

Multiple-Observation Elicitation

Sebastian Casalaina-Martin CASA@MATH.COLORADO.EDU
CU Boulder

Rafael Frongillo RAF@COLORADO.EDU
CU Boulder

Tom Morgan TDMORGAN@SEAS.HARVARD.EDU
Harvard

Bo Waggoner BWAG@SEAS.UPENN.EDU
UPenn

Abstract

We study loss functions that measure the accuracy of a prediction based on multiple data points simultaneously. To our knowledge, such loss functions have not been studied before in the area of property elicitation or in machine learning more broadly. As compared to traditional loss functions that take only a single data point, these multi-observation loss functions can in some cases drastically reduce the dimensionality of the hypothesis required. In elicitation, this corresponds to requiring many fewer reports; in empirical risk minimization, it corresponds to algorithms on a hypothesis space of much smaller dimension. We explore some examples of the tradeoff between dimensionality and number of observations, give some geometric characterizations and intuition for relating loss functions and the properties that they elicit, and discuss some implications for both elicitation and machine-learning contexts.

Keywords: Property elicitation, loss functions, empirical risk minimization.

1. Introduction

In machine learning and statistics, empirical risk minimization (ERM) is a dominant inference technique, wherein a model is chosen which minimizes some loss function over a data set. As the choice of loss function used in ERM may have a large impact on the model chosen, how should one choose this loss? A growing body of work in *property elicitation* seeks to answer this question, by viewing a loss function as "incentivizing" the prediction of a particular conditional statistic (Lambert et al., 2008; Gneiting, 2011; Steinwart et al., 2014; Frongillo and Kash, 2015a; Agarwal and Agarwal, 2015); for example, it is well-known that squared loss elicits the mean, and hence least-squares regression finds the best fit to the conditional means of the data.[1]

A natural question, which is still open in the vector-valued case, is the following: for which conditional statistics do there exist loss functions which elicit them? Positive examples include the mean, median, other quantiles, moments, and several others. Perhaps surprisingly, however, there are negative examples as well: it is well-known that the variance is not elicitable, meaning there is no loss function for which minimizing the loss will yield the variance of the data or distribution.

1. There are also contributions from microeconomics, and crowdsourcing in particular, where one wishes to incentivize humans rather than algorithms, but the mathematics is the same.

The usual approach to dealing with non-elicitable statistics is called *indirect elicitation*: elicit other conditional statistics from which one can compute the desired statistic. For example, the variance of a distribution can be written as (2nd moment) - (1st moment)^2, and as mentioned above, moments are elicitable. The question of how many such auxiliary statistics are required gives rise to the concept of *elicitation complexity*; since the variance cannot be elicited with one but can with two, we say it is 2-elicitable (Lambert et al., 2008; Frongillo et al., 2015).

In this paper, we explore an alternative approach to dealing with non-elicitable statistics, by allowing the loss function to *depend on multiple data points* simultaneously. In the language of property elicitation, this corresponds to loss functions such as $\ell(r, y_1, y_2)$ which judge the "correctness" of the report r based on two (or more) observations y_1 and y_2. Assuming these observations are drawn independently from the same distribution, this intuitively gives the loss function more power, and could potentially render previously non-elicitable statistics elicitable. In fact, the variance is one such example: if y_1 and y_2 are both drawn i.i.d. from p, it is easy to see that $\frac{1}{2}(y_1 - y_2)^2$ will be an unbiased estimator for the variance of p, whence $\ell(r, y_1, y_2) = (r - \frac{1}{2}(y_1 - y_2)^2)^2$ elicits the variance for the usual reason that squared error elicits expected values.

Beyond the variance, are there other non-elicitable statistics which we can elicit with multiple i.i.d. observations? Moreover, what is the tradeoff between the number of observations and the number of reports? One would expect the elicitation complexity, in the usual number-of-reports sense, to drop as observations are added, but how fast is unclear. Indeed, we will see several examples where the complexity drops dramatically, such a the k-norm of the distribution p. In Section 4 we develop new techniques to prove complexity bounds using algebraic geometry, which show for example that the complexity of the k-norm drops from the support size of p (minus 1) with 1 observation, to 1 with k observations. We call the feasible (# reports, # observations) pairs the *elicitation frontier*, for which the given statistic is elicitable, a concept we explore in Section 5.

Finally, in Section 6 we apply multi-observation elicitation to regression. Traditional elicitation complexity expresses a conditional statistic Γ as a link of other statistics, but as we illustrate, situations can arise where these other statistics have a much more complicated relationship with the covariates than Γ does. We give an example where fitting a model to the conditional variance directly (using nearby data points as proxies for i.i.d. observations) is much better than fitting separate models to the conditional first and second moments and combining these to obtain the variance.

2. Preliminaries

We are interested in a space $\mathcal{Y}$ from which *observations* y are drawn, which will be a finite set unless otherwise specified. We will denote by $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ a set of distributions of interest. We refer to the set $\Delta_{\mathcal{Y}^m}$ of all distributions on m outcomes as the *m-product space*. To capture the assumption that we may collect $m \in \{1, 2, \ldots\}$ observations which are each i.i.d. from the same distribution $p \in \Delta_{\mathcal{Y}}$, we will write $p^m \in \Delta_{\mathcal{Y}^m}$ to denote their joint distribution, $p^m(y_1, \ldots, y_m) = \prod_i p(y_i)$. The set of all such distributions is denoted $\mathcal{P}^m = \{p^m : p \in \mathcal{P}\} \subseteq \Delta_{\mathcal{Y}^m}$, which we will think of as a manifold in the m-product space.

With this notation in hand, we can define the central concepts in elicitation complexity in our context. Properties include any typical statistic,[2] for instance, the mean when $\mathcal{Y} \subseteq \mathbb{R}$ is the property $\Gamma(p) = \sum_y p(y)y$.

Definition 1 (Property) *A* property *is a function* $\Gamma : \mathcal{P} \to \mathbb{R}^d$, $d \geq 1$, *which intuitively represents the information desired about the data or underlying distribution. The range of* Γ *is* $\mathcal{R} = \Gamma(\mathcal{P})$, *sometimes called the "report space".*

The central notion of property elicitation is the relationship between a loss function ℓ and the minimizer of its expected loss. If this minimizer is a particular property Γ, we say ℓ *elicits* Γ. We simply extend this usual definition to allow for multiple observations in the expected loss.

Definition 2 (Loss function, elicits) *An* m-observation loss function *is a function* $\ell : \mathcal{R} \times \mathcal{Y}^m \to \mathbb{R}$, *where* $\ell(r, y_1, \ldots, y_m)$ *is the loss for prediction* $r \in \mathcal{R}$ *scored against realized observations* $y_i \in \mathcal{Y}$. *We say* ℓ (directly) elicits *a property* $\Gamma : \mathcal{P} \to \mathcal{R}$ *if for all* $p \in \mathcal{P}$ *we have* $\{\Gamma(p)\} = \operatorname{argmin}_{r \in \mathcal{R}} \mathbb{E}_{(y_1,\ldots,y_m)\sim p^m}[\ell(r, y_1, \ldots, y_m)]$.

It is useful to consider a property in terms of its *level sets*, the set of distributions sharing the same particular value of the property. For example, when the property is the mean of a distribution on $\{1, 2, 3, 4\}$, both $p = \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right)$ and $p = \left(0, \frac{1}{2}, \frac{1}{2}, 0\right)$ lie in the level set $\Gamma_{2.5}$.

Definition 3 (Level set) *A* level set Γ_r *of a property* $\Gamma : \mathcal{P} \to \mathbb{R}^d$ *is, for* $r \in \mathcal{R}$, *the set of distributions with property* r, *i.e.* $\Gamma_r = \{p \in \mathcal{P} \mid \Gamma(p) = r\}$.

An important technical condition on a property, and one which we will need for the notion of indirect elicitability, is that it be *identifiable*, meaning that its level sets can be described by linear equalities.

Definition 4 (Identifiable) *A property* $\Gamma : \mathcal{P} \to \mathcal{R}$ *is* identifiable with m observations *if for all* $r \in \mathcal{R}$ *we have some* $V_r : \mathcal{Y}^m \to \mathbb{R}^d$ *such that* $\Gamma(p) = r \iff \mathbb{E}_{p^m}[V_r(\mathbf{y})] = 0 \in \mathbb{R}^d$, *where* $\mathbf{y} = (y_1, \ldots, y_m)$. *We also say it is* m-identifiable.

Identifiability is a geometric restriction on properties that is intuitively similar to continuity of the property (cf. Lambert et al. (2008); Steinwart et al. (2014)). Technically, observe that *differentiable* loss functions generally elicit an identifiable property, as any local optimum should have $\sum_i \frac{\partial}{\partial r_i} \ell(r, \mathbf{y}) = 0$, meaning that the gradient of ℓ itself gives an identification function. Following Frongillo and Kash (2015a), we will often assume that properties are identifiable.

Notice that any property can be "indirectly" elicited by using a proper scoring rule, which elicits the entire distribution, and then computing the property from the distribution. But this requires a report of dimension $|\mathcal{Y}| - 1$, whereas to indirectly elicit the variance of y, for example, requires just two reports, e.g. $r_1 = \mathbb{E}y$ and $r_2 = \mathbb{E}y^2$, along with a "link function" $\psi(\mathbf{r}) = r_2 - r_1^2$. The question of *elicitation complexity*, studied by Lambert et al. (2008) and Frongillo et al. (2015), is how many dimensions d are needed to indirectly elicit the property of interest Γ via some elicitable $\hat{\Gamma} : \mathcal{P} \to \mathbb{R}^d$; one hopes that d is much smaller than $|\mathcal{Y}|$. Here we augment this question by another degree of freedom: how many dimensions d, *and observations* m, are needed to indirectly elicit Γ?

2. As defined, statistics like the median would not be included unless restrictions were placed on $\mathcal{P}$ for them to be single-valued (distributions in general may have multiple medians); we may instead extend our definition to include set-valued statistics, which would not substantially alter our results, and in fact we do lift this restriction in Section 3.1.

Definition 5 ((d,m)-elicitable) *A property $\Gamma : \mathcal{P} \to \mathcal{R}$ is (d,m)-elicitable if there exists an m-observation, d-dimensional loss function $\ell : \mathbb{R}^d \times \mathcal{Y}^m \to \mathbb{R}$, an identifiable property $\hat{\Gamma} : \mathcal{P} \to \mathbb{R}^d$, and a "link" function $\psi : \mathbb{R}^d \to \mathcal{R}$ such that*

$$1.\ \ell \text{ directly elicits } \hat{\Gamma}, \text{ and} \quad 2.\ \Gamma(p) = \psi\left(\hat{\Gamma}(p)\right).$$

The elicitation frontier *of Γ is the set of (d,m) such that Γ is (d,m)-elicitable, but neither $(d-1,m)$- nor $(d,m-1)$-elicitable.*

We may say that a property's "report complexity" is d if $(d,1)$ lies on its frontier, and its "observation complexity" is m if $(1,m)$ does.

2.1. Illustrative example

Recall our observation that the variance is not $(1,1)$-elicitable, and the "traditional" fix is to utilize $(2,1)$-elicitability: minimize a loss function over two dimensions (say first and second moments), mapping the result to the variance via a link function. We observed instead that it is possible to utilize $(1,2)$-elicitability: minimize a loss function that takes two observations over a single scalar, the variance itself. Can this tradeoff be more extreme? In particular, are there cases where additional observations drastically decrease the report complexity? Consider the 2-norm of a distribution: $\Gamma(p) = \|p\|_2 = \sqrt{\sum_y p(y)^2}$. We show in Section 5.2 that $\|p\|_2$ has report complexity $|\mathcal{Y}| - 1$ (where $\mathcal{Y}$ is the outcome set) for 1 observation – no single-observation loss function can do better than solving for the entire distribution. However, recall that $\|p\|_2^2 = \sum_y p_y^2 = \Pr[y_1 = y_2]$ for two i.i.d. observations y_1, y_2, or in other words, $\|p\|_2^2 = \mathbb{E}_p \mathbb{1}\{y_1 = y_2\}$. The two-norm is actually elicitable with two observations and a single dimension using e.g. loss function $\ell(r, y_1, y_2) = (r - \mathbb{1}\{y_1 = y_2\})^2$, then simply computing $\|p\|_2 = \sqrt{r}$. In other words, the two-norm's elicitation frontier on $\mathcal{Y}$ consists of the points $(|\mathcal{Y}| - 1, 1)$ and $(1,2)$.

The goal for this paper is to investigate the (algebraic-)geometric reasons underpinning why a property might have low or high observation complexity, as well as providing general results and examples based on these ideas. We next introduce the geometric foundations for this investigation.

3. Geometric Fundamentals

The most basic (yet powerful) lower bound in property elicitation says that elicitable properties' level sets must be convex sets (Lambert et al., 2008). Indeed, this is used to prove the variance is not (1,1)-elicitable; but the variance *is* elicitable with two observations. The geometry is not "broken" here, but merely lives in a higher-dimensional space. When reasoning about eliciting a property $\Gamma : \mathcal{P} \to \mathbb{R}$ using m observations, it often useful to instead think of eliciting the property using a single random draw from a distribution on m-tuples of outcomes. In particular, it suffices to be able to elicit a property on $\Delta_{\mathcal{Y}^m}$ that can be linked to get Γ.

Proposition 6 *Let $\Gamma : \mathcal{P} \to \mathcal{R}$ be a property. If there exists $\Gamma' : \Delta_{\mathcal{Y}^m} \to \mathbb{R}^d$ that is identifiable and directly elicitable with one observation, and there is a link ψ with $\psi(\Gamma'(p^m)) = \Gamma(p)$ for all $p \in \mathcal{P}$, then Γ is (d,m)-elicitable.*

Proof There is a loss function $\ell' : \mathbb{R}^d \times (\mathcal{Y}^m)$ that elicits Γ'. However, we may simply define an m-observation loss function $\ell(r, y_1, \ldots, y_m) = \ell'(r, \mathbf{y})$ where $\mathbf{y} = (y_1, \ldots, y_m)$. If $p \in \mathcal{P}^m \subseteq \Delta_{\mathcal{Y}^m}$,

then $\psi(\Gamma'(p^m)) = \Gamma(p)$, so $\psi\left(\operatorname{argmin} \ell'(r, \mathbf{y})\right) = \psi\left(\operatorname{argmin} \ell(r, y_1, \ldots, y_m)\right) = \Gamma(p)$. ■

This result gives us an alternative for demonstrating that a property is elicitable with m observations. For example, the loss function $\ell(r, a, b) = \left(r - \frac{1}{2}(a-b)^2\right)^2$ elicits the variance with two observations a, b, but if we consider distributions on all of $\mathcal{Y} \times \mathcal{Y}$, including non-i.i.d. distributions, it actually is still a valid loss function eliciting a property that coincides with the variance when a, b are i.i.d. To see this, just note that it still elicits an expectation: $\sum_{a,b} p'(a,b)\frac{1}{2}(a-b)^2$ where p' is a distribution on $\mathbb{R}^2$.

However, considering elicitation on the larger space $\Delta_{\mathcal{Y}^m}$ does not resolve the problem in either the necessary or sufficient directions. First, $\mathcal{P}^m$ is not a convex set for $m > 1$, so conditions on the convexity of level sets do not naturally extend here. An example of this is shown in Figure 1. Second, coming up with an "extended property" may be difficult or non-obvious. For example, it is not so clear whether the above loss function elicits anything *natural* on $\Delta_{\mathcal{Y}^2}$ (it is not the covariance, for instance, which is zero for i.i.d. distributions). More fundamentally, it is not clear whether such extensions should generally exist. (Proving or constructing a counterexample is an interesting open problem.) In general, we hope to be able to accomplish much more by restricting to $\mathcal{P}^m$ because it is only a tiny $|\mathcal{Y}|$-dimensional manifold in a $|\mathcal{Y}|^m$-dimensional space.

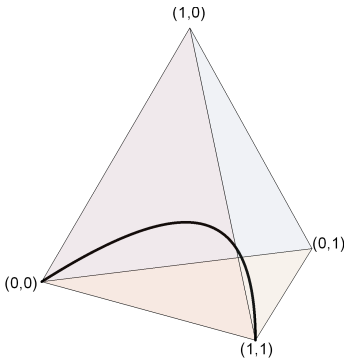

Figure 1: The two outcome, two observation probability simplex $\Delta_{\mathcal{Y}^2}$ where $\mathcal{Y} = \{0, 1\}$. The arc is the space of i.i.d. distributions $(\Delta_{\mathcal{Y}})^2$.

A tighter sufficient condition is given by Frongillo and Kash (2014), which states that essentially all loss functions eliciting a property on any set, such as $\mathcal{P}^m$, also elicit some "extension" of that property on the convex hull of that set. So while the higher-dimensional approach is helpful, it does not preclude reasoning about the space $\mathcal{P}^m$ as a manifold inside $\Delta_{\mathcal{Y}^m}$.

Most significantly, $\mathcal{P}^m$ is not a convex space, which makes lower bounds on elicitation complexity nontrivial as well. However, the result of Frongillo and Kash (2014) shows that it suffices to provide lower bounds for elicitation on the convex hull of $\mathcal{P}^m$, which we will denote $\operatorname{conv}(\mathcal{P}^m)$. Quite naturally then, we explore what leverage we can gain by reasoning about $\operatorname{conv}(\mathcal{P}^m)$.

Theorem 7 *The property $\Gamma : \mathcal{P} \to \mathbb{R}$ is not directly elicitable with m observations if there exists $r_1, r_2 \in \Gamma(\mathcal{P})$, $p_{1,1}, \ldots, p_{1,k_1} \in \Gamma_{r_1}$, $p_{2,1}, \ldots, p_{2,k_2} \in \Gamma_{r_2}$, $\lambda_{1,1}, \ldots, \lambda_{1,k_1} \in [0,1]$ and $\lambda_{2,1}, \ldots, \lambda_{2,k_2} \in [0,1]$ such that $r_1 \neq r_2$, $\sum_{i=1}^{k_1} \lambda_{1,i} = 1$, $\sum_{i=1}^{k_2} \lambda_{2,i} = 1$ and*

$$\sum_{i=1}^{k_1} \lambda_{1,i} p_{1,i}^m = \sum_{i=1}^{k_2} \lambda_{2,i} p_{2,i}^m.$$

In other words, a property is not elicitable if there is a convex combination of one of its level sets in the m-product space that equals a convex combination of another one of its level sets in the m-product space. Theorem 7 allows us to prove for example that the fourth central moment is not directly elicitable with two observations. Consider a Bernoulli random variable $Y \sim p$, then two of the level sets of the fourth central moment $\Gamma(p) = \mathbb{E}_{Y\sim p}[(Y - E_{Y\sim p}[Y])^4]$ are given in Figure 2. When we project these level sets into the 2-product space we can easily find a pair of points from each level set whose connecting lines intersects in $\operatorname{conv}(\Delta_{\mathcal{Y}^m})$. These lines are convex

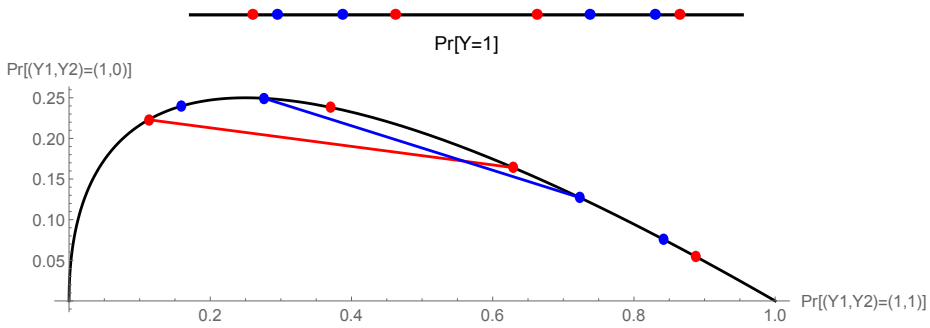

Figure 2: **Top:** The red dots and blue dots are each a level set of the fourth central moment of a Bernoulli random variable $Y \sim p$. These correspond to the distributions with fourth central moments .07 and .08 respectively. **Bottom:** The curve is $\Delta_{\mathcal{Y}^2}$ projected into $\mathbb{R}^2$, and the colored dots are the level sets of the example above projected into this space. The lines demonstrate that there is a point in $\mathrm{conv}(\Delta_{\mathcal{Y}^2})$ that can be written as a convex combination of either of the two level sets.

combinations of points in the same level set, so by Theorem 7 the lines' intersection implies that Γ is not directly elicitable with two observations.

3.1. Finite Properties

Finite properties are those where $\mathcal{R}$, the range of Γ, is a finite set. This corresponds to a "multiple-choice question" (Lambert and Shoham, 2009). In this section, we must allow $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ to be a set-valued function, possibly assigning multiple possible correct reports to a single distribution; this is necessary for "boundary" cases, such as the mode of the uniform distribution on a finite set. (Similarly, we cannot require identifiability.) We have a finite set of outcomes $\mathcal{Y}$, the distributions considered are all $\mathcal{P} = \Delta_{\mathcal{Y}}$, and $\Gamma(p)$ must be nonempty.

We are interested in understanding which finite properties can be elicited with m observations. Previously, this question was studied for the case of one observation by Lambert (2011), who characterized elicitable properties by the shape of their level sets: they are intersections of *Voronoi diagrams* in $\mathbb{R}^{|\mathcal{Y}|}$ with the simplex $\Delta_{\mathcal{Y}}$. In our setting, a Voronoi diagram is specified by a finite set of points $\{x_r : r \in \mathcal{R}\} \subseteq \mathbb{R}^{\mathcal{Y}^m}$, with each *cell* $T_r = \{x : \|x - x_r\| \leq \|x - x_{r'}\| \forall r' \in \mathcal{R}\}$ consisting of those points in $\mathbb{R}^{\mathcal{Y}^m}$ closest in Euclidean distance to x_r.

Using the geometric constructions above, we can simply apply the main result of Lambert (2011) to finite properties in the m-product space; the result is a characterization of elicitable finite properties with m observations.

Corollary 8 *A finite property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is directly elicitable with m samples if and only if there exists a Voronoi diagram in $\mathbb{R}^{\mathcal{Y}^m}$ with $\{x_r : r \in \mathcal{R}\}$ satisfying $\Gamma_r^m = T_r \cap \mathcal{P}^m$. Here $\Gamma_r^m = \{p^m \in \mathcal{P}^m : p \in \Gamma_r\}$.*

Multiple observations afford considerable flexibility in the level sets of such an elicitable Γ. In particular, whereas before the cell boundaries between level sets were restricted to hyperplanes, with m observations these boundaries can be defined by nearly arbitrary m-degree polynomials. We illustrate this flexibility and visualize the cell boundaries in Figure 3. In particular, we show that a classic negative example, where an agent is asked to report whether their belief has low or high variance, is easily elicited with two observations.

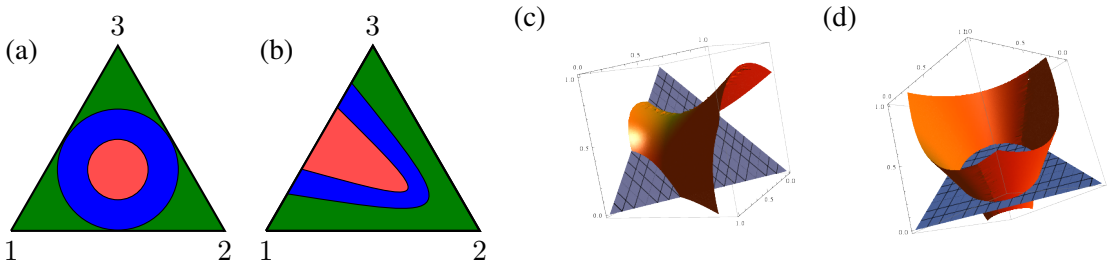

Figure 3: **Examples of finite properties on $\mathcal{Y} = \{1, 2, 3\}$ elicitable with 2 samples.** Pictured is the simplex on 3 outcomes and properties $\Gamma : \Delta_{\{1,2,3\}} \to \{\text{red, green, blue}\}$. The agent reports a color, then is rewarded according to which outcome occurs. (a) The property of "close", "intermediate", and "far" from uniform, as measured by 2-norm. (b) The property of "high", "medium", and "low" variance. (c,d) The boundary between two cells, i.e. a hyperplane in the m-product space "projected" down to $\mathbb{R}^{\mathcal{Y}}$ (in orange) and intersected with the simplex $\Delta_{\mathcal{Y}}$ (in blue/gray); we show the boundary on all of $\mathbb{R}^{\mathcal{Y}}$ to visualize the quadratic surfaces which create these sections.

4. Lower Bounds via Algebraic Geometry

The general formula for showing lower bounds in elicitation complexity introduced by Frongillo et al. (2014) is the following. Let property Γ be indirectly elicited via $\hat{\Gamma}$ and link ψ, so that $\Gamma = \psi \circ \hat{\Gamma}$, and consider the relationship between the level sets of Γ and $\hat{\Gamma}$. If ψ is a bijection, then the level sets must be identical, because all $p \in \hat{\Gamma}_{\hat{r}}$ are mapped to some $r = \psi(\hat{r})$, and only those p are, so $\Gamma_r = \hat{\Gamma}_{\hat{r}}$. As ψ is a surjection onto $\mathcal{R}$ by definition, the only other possibility is that ψ maps multiple $\hat{r}_1, \hat{r}_2$ to the same $r \in \mathcal{R}$. In this case, the level set $\Gamma_r = \cup_{\hat{r}:\psi(\hat{r})=r} \hat{\Gamma}_{\hat{r}}$. In other words, some of the level sets of $\hat{\Gamma}$ are combined to form level sets of Γ. Thus, a necessary condition for Γ to be indirectly elicited via $\hat{\Gamma}$ is that every level set of $\hat{\Gamma}$ is contained in a level set of Γ. Therefore, if one can show that no level set from any $\hat{\Gamma}$, which is m-identifiable and directly elicitable with m observations, can be contained in a particular level set of Γ, then Γ cannot be (d, m)-elicitable.

The above formula was used in Frongillo et al. (2014) to show lower bounds on the report complexity (d) of a property, with $m = 1$. For this section, we will use the same formula to show lower bounds on observation complexity (m), with $d = 1$. Our main tool will be that the level sets of any directly m-observation-elicitable, identifiable $\hat{\Gamma}$ have specific structure as a function of m. This will imply requirements on the structure of the level sets of Γ, which will give a lower bound on m. Specifically, we leverage the structure endowed by identifiability.

Fact 1 *If a property $\hat{\Gamma}(p)$ is m-identifiable, then its level sets are each the set of zeros of a degree-at-most-m polynomial in p.*

Proof The condition $\mathbb{E}_p V_r = 0$ is $\sum_{y_1,\ldots,y_m} p(y_1) \cdots p(y_m) V_r(y_1, \ldots, y_m) = 0$. ■

Combined with the discussion above, Fact 1 tells us that the level sets of indirectly elicitable Γ are unions of zero sets of polynomials. As we are focusing on the $d = 1$ case, however, both Γ and $\hat{\Gamma}$ are real-valued functions, so with enough regularity, their level sets should coincide. This is precisely what the following Lemma gives us.

Lemma 9 *Suppose $\Gamma, \hat{\Gamma}$ are C_∞ functions from $\mathcal{P}$ to $\mathbb{R}$. If Γ_r is a connected set and contains some $\hat{\Gamma}_{\hat{r}}$, then $\Gamma_r = \hat{\Gamma}_{\hat{r}}$.*[3]

Together, Lemma 9 and Fact 1 tell us that the level sets of Γ itself must be zero sets of polynomials.

Corollary 10 *Given C_∞ property $\Gamma : \mathcal{P} \to \mathbb{R}$, if Γ is $(1, m)$-elicitable via identifiable and C_∞ properties $\hat{\Gamma}$, then any connected level set Γ_r is the set of zeros of a polynomial of degree m.*

For example, we can immediately show the existence of properties with *infinite* observation complexity. The proof gives such an example for $|\mathcal{Y}| = 3$, a surprising result given that all properties have report complexity $|\mathcal{Y}| - 1 = 2$ in this case.

Proposition 11 *There are properties that are not $(1, m)$-elicitable for any finite m.*

Proof Take $\mathcal{Y} = \{1, 2, 3\}$ and $\Gamma(p) = p_1 - (1/2)\sin(\pi p_2)$. Here the level sets Γ_r satisfy $r = p_1 - (1/2)\sin(\pi p_2)$, in other words, satisfy the equation $p_1 = (1/2)\sin(\pi p_2) + r$. The level set $\Gamma_0 = \{p \in \Delta_3 : p_1 = (1/2)\sin(\pi p_2)\}$ is simply the graph of $(1/2)\sin(\pi x)$ for $x = 0$ to $1/2$, which is connected, but cannot be the zeros of *any* polynomial because sine is a transcendental function. Corollary 10 now implies that Γ is not $(1, m)$-elicitable for any m. ■

In light of Corollary 10, if we find any connected level set of a property Γ that cannot be the zeros of any degree-m polynomial in $p(y_1), \ldots, p(y_m)$, we can conclude that Γ is not $(1, m)$-elicitable. For example, consider the k-norm or the equivalent problem of eliciting $\Gamma(p) = \sum_y p(y)^k$, whose level sets are L^k-spheres intersected with the simplex. It only remains to show the following.

The level sets of the k-norm of are not all contained in the zero sets of $(k-1)$-degree polynomials. (1)

While (1) may appear obvious, this type of result can be more subtle than might be expected.We next recall standard results in algebraic geometry that allow us to establish (1).

4.1. Reasoning about irreducible polynomials

Let us begin by describing *non*-examples. We know that, for instance, $\Gamma(p) = (\mathbb{E}_y y)^2$ is $(1, 1)$-elicitable, by eliciting $\mathbb{E}_y y$ and using the link $\psi(x) = x^2$. Yet Γ is a degree-2 polynomial in p. To resolve the issue, note that its level sets, the solutions to $\Gamma(p) = r$ for $r \geq 0$, can be written as the solutions to

$$(\mathbb{E}_y y)^2 = r \iff \mathbb{E}_y y = \sqrt{r} \iff \mathbb{E}_y y - \sqrt{r} = 0 \,,$$

so its level sets are also the roots of a degree-one polynomial. An *irreducible polynomial* (over the reals) is one which cannot be factored into non-constant polynomials with real coefficients. Due to cases like the above, we will only be able to show negative results when the polynomials of interest are irreducible of degree m.

A second useful condition will be existence of a *nonsingular zero* of the polynomial f: an x for which $f(x) = 0$ but $\nabla f(x) \neq 0$. This condition is important to rule out "repeated zeros", such as $f(x) = \|x - c\|_2^2$, which has a unique zero at $x = c$ but shares this zero set with $g(x) = x - c$.

3. An omitted detail is that r must be a *regular value* (Γ_r contains no critical points); Corollary B.6 has the full statement.

For f on all of $\mathbb{R}^n$, the (real) Nullstellensatz states that, if f is nonconstant, irreducible, and has a nonsingular zero, then all polynomials which vanish on the zeros of f are multiples of f. In particular, for our purposes, no g of lower degree can have the same zero set. Let $Z(f) = \{x : f(x) = 0\}$. Because we can only distinguish the behavior of these polynomials on the simplex $\Delta\mathcal{Y}$, we need a generalization of this result to subsets of $\mathbb{R}^n$, which we develop in Appendices D and E.

Theorem 12 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a nonconstant, irreducible polynomial, and let $U \subseteq \mathbb{R}^n$ be an open set. If U contains a nonsingular zero of f, then no nonzero polynomial g of degree less than f can have $Z(g) \cap U \supseteq Z(f) \cap U$.*

Note that the simplex is not an open set in $\mathbb{R}^{\mathcal{Y}}$, so to apply the theorem, we will work in $\mathbb{R}^{|\mathcal{Y}|-1}$, i.e., $\hat{f}(x_1, \ldots, x_{n-1}) = f(x_1, \ldots, x_{n-1}, 1 - x_1 - \cdots - x_{n-1})$.

We now have a method to prove (1): show that the k-norm is irreducible as a polynomial over the reals. Establishing irreducibility is not always trivial, however; we give one such technique in Corollary D.3 which computes partial derivatives of the *homogenized* polynomial (wherein powers of a new variable x_0 are added to every monomial until all have the same degree) and verifies that the only solution over the complex vector space $\mathbb{C}^{n+1}$ is 0. We perform this calculation in Example D.1.

Corollary 13 *For integer k, the k-norm of a distribution, $\Gamma(p) = (\sum_y p(y)^k)^{1/k}$, is not $(1, k-1)$-elicitable with respect to the class of C_∞ and identifiable properties.*

5. Examples and Elicitation Frontiers

We now combine our complexity lower bounds with upper bounds to make progress toward determining the elicitation frontiers of some potential properties of interest. We begin with a straightforward, but nonetheless versatile, upper bound.

Lemma 14 *For all $1 \le i \le n$, $1 \le j \le m$, let $f_{ij} : \mathcal{Y} \to \mathbb{R}$ be an arbitrary function such that $\mathbb{E}_p[f_{ij}(Y)]$ exists for all $p \in \mathcal{P}$. Then $\Gamma(p) = \sum_{i=1}^n \prod_{j=1}^m \mathbb{E}_p[f_{ij}(Y)]$ is $(1, m)$-elicitable.*

Proof Using $Y_1, \ldots, Y_m$ which are i.i.d. from p, then $\{f_{i1}(Y_1), \ldots f_{im}(Y_m)\}$ will be independent for all i. Using properties of expectations (linearity and independence), we have

$$\sum_{i=1}^n \prod_{j=1}^m \mathbb{E}[f_{ij}(Y)] = \sum_{i=1}^n \prod_{j=1}^m \mathbb{E}[f_{ij}(Y_j)] = \sum_{i=1}^n \mathbb{E}\left[\prod_{j=1}^m f_{ij}(Y_j)\right] = \mathbb{E}\left[\sum_{i=1}^n \prod_{j=1}^m f_{ij}(Y_j)\right] \quad (2)$$

Now we see that using squared loss (or any loss for the mean) one can leverage these m samples to elicit the desired sum of products, e.g. $\ell(r, y_1, \ldots, y_m) = \left(r - \sum_{i=1}^n \prod_{j=1}^m f_{ij}(y_j)\right)^2$. ■

The proof of Lemma 14 simply constructs an unbiased estimator of the property of interest and elicits the mean of the estimator via squared error. By a very natural extension, this technique also applies to ratios of expectations, as they are elicitable (Gneiting, 2011): construct *two* unbiased estimators, and elicit the ratio of their means. We give two instances of such ratios next, followed by other examples. See Figure 4 for a depiction of the elicitation frontiers described below.

5.1. Ratios of expectations: index of dispersion and Sharpe ratio

The *index of dispersion* of a random variable Y with positive mean is defined to be $\mathrm{Var}(Y)/\mathbb{E}[Y]$ (Cox and Lewis, 1966). The *Sharpe ratio* of a random variable Y, which is a commonly-used measure of the risk-adjusted return of an investment, is defined similarly as $\mathbb{E}[Y]/\sqrt{\mathrm{Var}(Y)}$ (Sharpe, 1966). Both the index of dispersion and the *square* of the Sharpe ratio are $(1,2)$-elicitable by the above discussion: $\mathrm{Var}(Y) = \mathbb{E}_p[\frac{1}{2}(Y_1 - Y_2)^2]$, $\mathbb{E}_p[Y] = \mathbb{E}_p[Y_1]$, and $\mathbb{E}_p[Y]^2 = \mathbb{E}_p[Y_1 Y_2]$, so any ratios of these terms is $(1,2)$-elicitable. (The link function for the Sharpe ratio is thus the square root.) For example, the index of dispersion is elicited by the loss $\ell(r, y_1, y_2) = r(y_1 - y_2)^2 - r^2 y_1$.

To finish describing the elicitation frontiers for these properties, we note that neither is $(1,1)$-elicitable as the level sets are not convex, but both are $(2,1)$-elicitable as we now show. For the index of dispersion, we can take $r_1 = E[Y]$ and $r_2 = E[Y^2]$, both elicitable as means, and then compute the property by $(r_2 - r_1^2)/r_1$. Similarly, for the same r_1, r_2, the Sharpe ratio can be written as $r_1/\sqrt{r_2 - r_1^2}$.

5.2. Norms of distributions

As we have previously discussed, the 2-norm is $(1,2)$ elicitable. For general k, the k-norm is $(1,k)$ elicitable with the following loss function $\ell(r, y_1, \ldots, y_k) = (r - \mathbb{1}\{y_1 = \ldots = y_k\})^2$. (This case also follows from Lemma 14.) This is a tight bound on the observation complexity, as we proved in Section 4.1 that the k-norm is not $(1, k-1)$ elicitable. As it turns out, the report complexity of the k-norm is $|\mathcal{Y}| - 1$, meaning it is as hard to elicit with one observation as the entire distribution. This follows from Theorem 2 of Frongillo et al. (2015), specifically Section 4.2, as $\|p\|_k$ is a convex function of p. An interesting open question, and one that will require additional algebraic tools, is the k-norm's elicitation frontier when we allow multiple dimensions and multiple observations.

5.3. Central Moments

The n^{th} central moment μ_n of a random variable Y is defined as

$$\mu_n = \mathbb{E}[(Y - \mathbb{E}[Y])^n] = \sum_{i=0}^{n} (-1)^i \binom{n}{i} \mathbb{E}[Y]^i \cdot \mathbb{E}[Y^{n-i}] \,, \tag{3}$$

which we see is $(n,1)$-elicitable by simply eliciting $E[Y^i]$ for all $i \in \{1, \ldots, n\}$ and then combining the results. As we will show, μ_n is also $(1,n)$-elicitable, and moreover, we can achieve other dimension-observation tradeoffs in between, such as $(\lfloor\sqrt{n}\rfloor + 1, \lceil\sqrt{n}\rceil)$. The key idea is to partition the binomial sum (3) into k partial sums and factor out the highest power of $\mathbb{E}[Y]$ from each, such that the j^{th} partial sum can be written as

$$\mathbb{E}[Y]^{\frac{j \cdot n}{k}} \sum_{i=0}^{\frac{n}{k}-1} (-1)^i \binom{n}{\frac{j \cdot n}{k} + i} \cdot \mathbb{E}[Y]^i \cdot \mathbb{E}\left[Y^{\frac{(j+1)\cdot n}{k} - 1 - i}\right] \,. \tag{4}$$

Doing so gives the following result.

Theorem 15 *The n^{th} central moment is $(k+1, \lceil n/k \rceil)$- elicitable; $0 < k \leq n$*

Proof Consider the partial sum (4) without the $\mathbb{E}[Y]^{j \cdot n/k}$ factor; by Lemma 14 each such factored sum is $(1, \lceil n/k \rceil)$-elicitable, as the maximum number of terms in any product is $\lceil n/k \rceil$. Since we have k such factored sums, and need to additionally elicit the mean $\mathbb{E}[Y]$ to compute their factors, the entire sum can be elicited using $\lceil n/k \rceil$ observations and $k+1$ dimensions. ■

When $k = 0$, we can do much better than $m = \infty$: by Lemma 14, as the maximum number of terms in any product of (3) is n, the term $(\mathbb{E}[Y])^n$, we have than μ_n is $(1, n)$-elicitable. For lower bounds, little is known beyond μ_n not being $(1, 1)$-elicitable (Frongillo and Kash, 2015b).

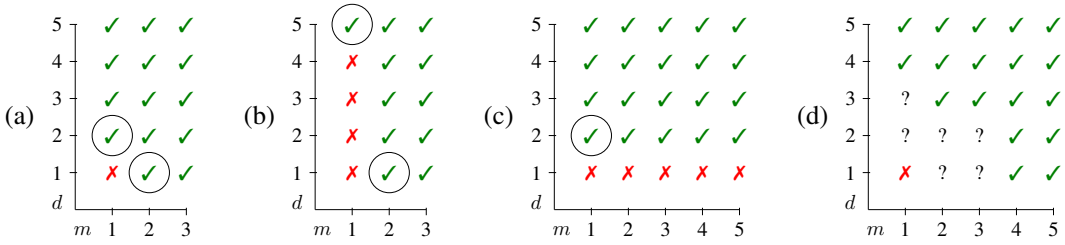

Figure 4: The elicitation frontiers for various properties: (a) the variance, Sharpe ratio, and index of dispersion; (b) the 2-norm when $|\mathcal{Y}| = 5$, with respect to C_∞ properties; (c) $\Gamma(p) = p_1 - (1/2)\sin(p_2\pi)$ from Proposition 11; (d) the 4th central moment, which is not fully known.

6. Multi-Observation Regression

One of the earliest problems in modern statistics was the estimation of biodiversity in a geographic region (Fisher et al., 1943). One scalar measure of diversity of a distribution is the (inverse of the) 2-norm, which we will take here as an example.[4] Consider a dataset of species samples: pairs (x, y) where x gives the features of the geographic region and y is a categorical giving the species to which this sample belongs. Suppose we wish to regress the diversity of species against geographic features such as climate. The single-observation approach would require a surrogate loss function $\ell(\hat{f}(x), y)$ and a link $f(x) = \psi(\hat{f}(x))$. We claim that any single-observation loss function $\ell(f(x), y)$ is poorly suited for this task. For the 2-norm, lower bounds on report complexity show that the best possible approach has dimensionality $\hat{f} : x \to \mathbb{R}^{d-1}$ where d is the number of unique species in the dataset (which may have a very long tail). So this approach requires, in essence, fitting $\hat{f}$ to the entire distribution over species as a function of geographic region, a task of immense idiosyncrasy and complexity compared to the end goal of e.g. estimating a scalar measure of diversity as a function of rainfall level.

On the other hand, a two-observation loss function $\ell(f(x), y_1, y_2)$ can be used to directly learn an f estimating the desired diversity measure, e.g. 2-norm, as a function of geographic features. One can then use empirical risk minimization to directly learn relationships between, e.g. rainfall level and this measure of species diversity.

Multi-observation regression does introduce an additional challenge, however: risk in this context is naturally defined as $\mathbb{E}_{x,\mathbf{y}}\ell(f(x), \mathbf{y})$ where $\mathbf{y} = (y_1, \ldots, y_m)$ is a set of observations drawn i.i.d. conditioned on x, but our data points are of the form (x, y). If e.g. x comes from a continuous

4. A similar intuition will hold for most if not all elicitable measures of diversity.

space, we may not have *any* sets of m samples $y_1, \ldots, y_m$ belonging to the same x. One natural setting where this poses no concern is in active learning where we may choose to re-draw the label for a given x. In a more standard regression framework, we propose to leverage the intuition that the distribution of y conditioned on x generally changes gradually as a function of x.[5] Pragmatically, with dense enough data points, we can simply group together nearby x values and "merge" them into a data point of the form $(\bar{x}, y_1, \ldots, y_m)$ where $\bar{x}$ is an average and the y_i are drawn independently and *approximately* identically from *approximately* the distribution of $\mathcal{Y}$ conditioned on $\bar{x}$. For this paper, we demonstrate the idea in simulations below and give a basic proof-of-concept theoretical result in Appendix C, leaving a more thorough investigation to future work.

In general, the cases where the multi-observation approach can be useful are those where the property of interest is believed to follow a simple functional form, but the conditional statistics given by the indirect elicitation approach are expected to follow unknown or complicated trends as a function of features. For another example, one could imagine learning the noise (e.g. variance) of a medical test, e.g. white blood cell count, as a function of patient features, in order to improve the test. The indirect elicitation approach suggests first fitting a model for estimating the mean of the test's outcome as a function of patient data, then fitting the expected square of the statistic, and then computing an estimate for the variance by combining them. In general, these prediction problems may be highly complex and nonlinear even when the *noise* in the test might follow some simple linear relationship with e.g. height or age. The multi-observation approach allows direct regression of the noise versus features. Formally, we show a basic extension of classic risk guarantees in Appendix C, under the assumption that x is distributed uniformly on $[0, 1]$ and a closeness condition on the conditional distribution of Y given X.

6.1. Simulation

Here we describe some simulations run as a proof of concept of multi-observation regression. Our data points are of the form $(x, y) \in \mathbb{R} \times \mathbb{R}$ where x is drawn uniformly at random from the interval $[0, 1]$. Given x, $y = a\sin(4\pi x) + Z$, where a is a constant and $Z \sim N(0, 1)$ is drawn independently for each sample, we wish to learn $\mathrm{Var}(Y|X)$.

Our multi-observation loss function here is $\ell(f(x), y_1, y_2) = (f(x) - \frac{1}{2}(y_1 - y_2)^2)^2$. We approximate (x, y_1, y_2) samples by sorting the (x_i, y_i) pairs by x_i, and making samples of the form $(\frac{1}{2}(x_i + x_{i+1}), y_i, y_{i+1})$. We compare to the single observation approach, in which we estimate $\mathbb{E}[Y|X]$ and $\mathbb{E}[Y^2|X]$ and then combine them to estimate $\mathrm{Var}(Y|X)$.

The point of these simulations is to demonstrate that multi-observation regression can greatly outperform single observation regression in the case when the function is in a known concept class, and the statistics needed to indirectly elicit it with a single observation are not in a known concept class. As such, our multi-observation regression fits a linear function to $\mathrm{Var}(Y|X)$, and our single observation regression fits linear functions to $\mathbb{E}[Y|X]$ and $\mathbb{E}[Y^2|X]$. The true $\mathrm{Var}(Y|X) = 1$ is indeed a linear function, while the true moment functions $\mathbb{E}[Y|X = x] = a\sin(x)$ and $\mathbb{E}[Y^2|X = x] = a^2\sin^2(x) + 1$ are very far from linear.

Figure 5 gives the results for $a = 1$ and $a = 10$. Both plots show the mean squared error of the variance functions reported by the two regression methods (averaged over 4000 simulations) as a

5. Phrased differently, at least it seems reasonable to parameterize the rate of change and expect learning bounds to depend on this parameter.

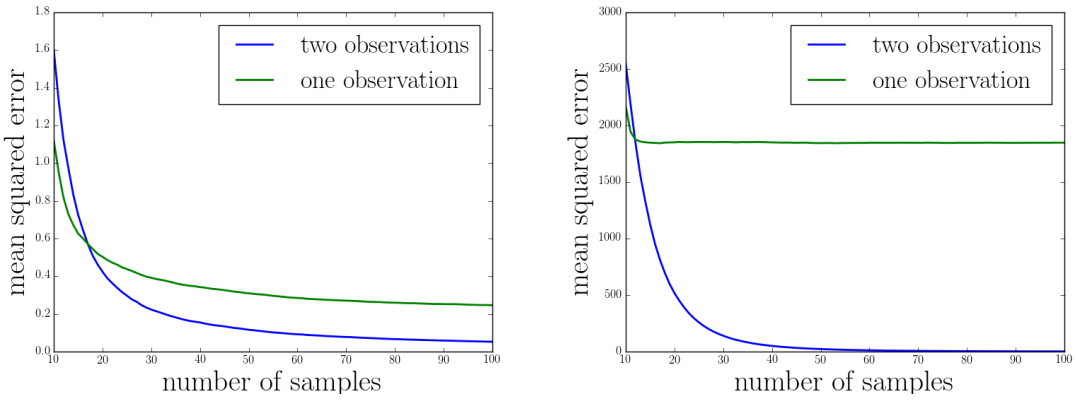

Figure 5: The mean squared error of the two regression strategies for estimating $\mathrm{Var}(y|x)$, where $x \sim \mathrm{Unif}(0,1)$ and $y \sim a\sin(4\pi x) + N(0,1)$, for $a = 1$ (left) and $a = 10$ (right). The single-observation loss function approach fails because it tries to fit to the complex underlying model of $y|x$, while the two-observation loss approach is able to directly model the simple relationship between $\mathrm{Var}(y)$ and x.

function of the number of samples. In both cases we see that for sufficiently many samples, the two observation regression significantly outperforms the single observation regression.

7. Conclusion and Future Work

An immediate host of directions is the proving of upper and lower bounds on elicitation frontiers for various properties. In particular, our lower bounds here focus on techniques for lower-bounding observation complexity (the $(1, m)$ case), leaving open approaches for lower bounds on (d, m) complexity for $d \geq 2$. Another direction is to formalize learning guarantees for multi-observation regression under suitable assumptions on slow-changing conditional distributions.

Acknowledgments

We thank Karthik Kannan for contributing the upper bound for central moments.

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.

M. F. Atiyah and I. G. Macdonald. *Introduction to commutative algebra*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1969.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1998. ISBN 3-540-64663-9. doi: 10.1007/978-3-662-03718-8.

URL http://dx.doi.org/10.1007/978-3-662-03718-8. Translated from the 1987 French original, Revised by the authors.

David R Cox and Peter AW Lewis. The statistical analysis of series of events. *Monographs on Applied Probability and Statistics*, 1966.

Ronald A. Fisher, A. Steven Corbet, and Carrington B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42–58, 1943.

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, 2015a.

Rafael Frongillo and Ian A. Kash. On Elicitation Complexity and Conditional Elicitation. *arXiv preprint arXiv:1506.07212*, 2015b. URL http://arxiv.org/abs/1506.07212.

Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. Elicitation for Aggregation. In *NIPS Workshop on Crowdsourcing and Machine Learning*, 2014.

Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. Elicitation for Aggregation. *AAAI*, 2015.

William Fulton. *Intersection theory*, volume 2 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, second edition, 1998. ISBN 3-540-62046-X; 0-387-98549-2. doi: 10.1007/978-1-4612-1700-8. URL http://dx.doi.org/10.1007/978-1-4612-1700-8.

T. Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

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.

N.S. Lambert. Elicitation and Evaluation of Statistical Forecasts. *Preprint*, 2011.

Anthony P Morse. The behavior of a function on its critical set. *Annals of Mathematics*, pages 62–70, 1939.

Arthur Sard et al. The measure of the critical values of differentiable maps. *Bull. Amer. Math. Soc*, 48(12):883–890, 1942.

William F Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, 1966.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and Identification of Properties. In *Proceedings of The 27th Conference on Learning Theory*, pages 482–526, 2014.

Loring W Tu. *An introduction to manifolds*. Springer Science & Business Media, 2010.

Appendix A. Overlapping Level Sets: Proof of Theorem 7

Theorem 7 states that a property is not elicitable if there is a convex combination of one of its level sets in the m-product space that equals a convex combination of another one of its level sets in the m-product space. To reason about these level sets we will need the following theorem.

Theorem A.1 (Theorem 3.5, Frongillo and Kash (2014)) *The property $\Gamma : \mathcal{P}' \to \mathbb{R}$ (where $\mathcal{P}' \subseteq \Delta_{\mathcal{Y}'}$) is directly elicitable by the loss function ℓ if and only if there exists some convex $G : \text{conv}(\mathcal{P}') \to \bar{\mathbb{R}}$ with $G(\mathcal{P}') \subseteq \mathbb{R}$, some $D \subseteq \delta G$, and some bijection $\phi : \Gamma(\mathcal{P}') \to D$ with $\Gamma(p) = \phi^{-1}(D \cap \delta G_p)$, such that for all $r \in \mathbb{R}$ and $y \in \mathcal{Y}'$,*

$$\ell(r, y) = \phi(r)(p_r - y) - G(p_r),$$

where $\{p_r\} \subseteq \mathcal{P}'$ satisfies $\hat{r} = \Gamma(p_{\hat{r}})$ for all $\hat{r}$.

Here δG_r is the set of subgradients to G at r.

Proof of Theorem 7

Let $\mathcal{Y}' = \mathcal{Y}^m$ and $\mathcal{P}' = \mathcal{P}^m$. Let ℓ be a loss function that elicits Γ of the form given by Theorem A.1, and let $G, \{p_r\}$ and ϕ be the corresponding values defined in Theorem A.1. We will let $\Gamma' : \mathcal{P}^m \rightrightarrows \mathbb{R}$ be the property that is elicited by ℓ on $\text{conv}(\mathcal{P}^m)$.

Note that Γ' is not necessarily single-valued everywhere on $\text{conv}(\mathcal{P}^m)$. This is because we cannot guarantee that there is a unique value that minimizes the loss function for distributions in the interior of $\text{conv}(\mathcal{P}^m)$. However, we can show that whenever $q \in \text{conv}(P^m)$ can be written as a convex combination of points on $\mathcal{P}^m$ that all have property value r then $\mathbb{E}_{y\sim q}\ell(r, y)$ is uniquely minimized at r, thus r is the unique property value of Γ' at q. This implies the theorem, as if q can be written as a convex combination of two separate level sets of Γ then there must not be an ℓ of the form specified in Theorem A.1 which elicits it.

If $q = \sum_{i=1}^k \lambda_i p_i^m$ for $p \in \Gamma_{r^*}$, $\lambda_1, \ldots, \lambda_k \in [0, 1]$ and $\sum_{i=1}^k \lambda_i = 1$ then

$$\begin{aligned}
\mathbb{E}_{y\sim q}\ell(r, y) &= \phi(r)(q_r - q) - G(q_r) \\
&= \phi(r)\left(q_r - \sum_{i=1}^k \lambda_i p_i^m\right) - G(q_r) \\
&= \sum_{i=1}^k \lambda_i \left(\phi(r)(q_r - p_i^m) - G(q_r)\right) \\
&= \sum_{i=1}^k \lambda_i \mathbb{E}_{y\sim p_i^m} L(r, y).
\end{aligned}$$

We know that each term of the final sum is uniquely minimized by $r = r^*$, thus $\mathbb{E}_{y\sim q}\ell(r, y)$ is uniquely minimized by r^*. ■

Appendix B. Manifolds and Level Sets

Theorem B.1 (Tu (2010) Theorem 9.10) *Let $\Gamma : M \to \mathbb{R}$ be a C^∞ function on the k-dimensional manifold M. Let r be a regular value of Γ such that the level set $S = \Gamma_r$ is nonempty. Then S is a regular submanifold of M of dimension $k-1$.*

Definition B.2 (Tu (2010) Proposition 8.20) *Let $\Gamma : M \to \mathbb{R}$ be a map between the k-dimensional manifold M and the reals. $x \in M$ is a* critical point *if and only if for some chart $(U, x^1, \ldots, x^n)$ containing p, for all $i \in \{1, \ldots, n\}$,*

$$\frac{\partial \Gamma}{\partial x^i}(p) = 0.$$

Definition B.3 (Tu (2010) Definition 8.19) *A point $r \in \mathbb{R}$ is a* critical value *of the map $\Gamma : M \to \mathbb{R}$ if and only if there exists a point $x \in \Gamma_r$ which is a critical point of Γ. Otherwise, r is a* regular value *of Γ.*

Theorem B.4 (Morse (1939); Sard et al. (1942)) *Given a k-dimensional manifold M, let $\Gamma : M \to \mathbb{R}$ be a C^k function. The set of critical values of Γ has Lebesgue measure 0 in $\mathbb{R}$.*

Lemma B.5 *Let $M \subseteq \Delta_{\mathcal{Y}^m}$ be a connected manifold of dimension $n-1$, where n is the dimension of $\Delta_{\mathcal{Y}^m}$. Let $\Gamma : \Delta_{\mathcal{Y}^m} \to \mathbb{R}$ be a C^∞ function. There exists an $r \in \Gamma(M)$ for which one of the following holds:*

- $\Gamma_r = M$ *or*
- $\Gamma_r \not\subseteq M$.

Proof This lemma is fairly immediate if $\Gamma|_M$ is constant. In this case there exists $r \in \mathbb{R}$ such that for all $m \in M$, $\Gamma(m) = r$. This gives us that $M \subseteq \Gamma_r$, implying the theorem.

Now let us assume that $\Gamma|_M$ is not constant. If we can find an $r \in \mathbb{R}$ such that $\Gamma_r \cap M$ is nonempty, and r is a regular value of both Γ and $\Gamma|_M$, then Theorem B.1 will imply the lemma. This is because we would have that Γ_r is an $n-1$ dimensional submanifold, and that $\Gamma_r \cap M$ is an $n-2$ dimensional manifold so clearly $\Gamma_r \not\subseteq M$.

Since M is connected and $\Gamma|_M$ is continuous and not constant, then by the intermediate value theorem $\Gamma(M)$ is a non-empty interval and thus has Lebesgue measure greater than 0. By Theorem B.4 and the fact that Γ and $\Gamma|_M$ are C^∞ functions, the sets of critical values of Γ and $\Gamma|_M$ are both of Lebesgue measure 0. Therefore, the subset of $\Gamma(M)$ whose points are regular values of Γ and $\Gamma|_M$ has Lebesgue measure greater than 0 and thus is nonempty. ■

Corollary B.6 *Let $\Gamma : \Delta_{\mathcal{Y}^m} \to \mathbb{R}$ and $\hat{\Gamma} : \Delta_{\mathcal{Y}^m} \to \mathbb{R}$ be C^∞ functions. Let $r \in \Gamma(\Delta_{\mathcal{Y}^m})$ be a regular value of Γ such that Γ_r is connected and there exists $\hat{r} \in \hat{\Gamma}(\Delta_{\mathcal{Y}^m})$ such that $\hat{\Gamma}_{\hat{r}} \subseteq \Gamma_r$, then $\Gamma_r = \hat{\Gamma}_{\hat{r}}$.*

Appendix C. Regression

In this section, we give a proof-of-concept showing that classic risk bounds for ERM can go through with only slight modification with multi-observation loss functions, under a natural assumption.

Regression can be naturally formulated in the multi-observation setting as follows: Given a hypothesis class $\mathcal{F} : \mathcal{X} \to \mathcal{R}$ and loss function $\ell : \mathcal{R} \times \mathcal{Y}^m \to \mathbb{R}$, given access to an unknown distribution $\mathcal{D}$ on $\mathcal{X}$ and conditional distributions $\{\mathcal{D}_x \in \Delta_{\mathcal{Y}} : x \in \mathcal{X}\}$, approximately minimize

$$\text{Risk}(f) = \mathbb{E}_{x\sim\mathcal{D},\mathbf{y}\sim\mathcal{D}_x}\ell(f(x), y_1, \ldots, y_m).$$

The central challenge that arises, new to the multi-observation setting, is that the data we are given is of the form $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \sim \mathcal{D}$ and $y_i \sim \mathcal{D}_{x_i}$ i.i.d. We may only obtain a single y for any given x. In this section, we give an example of how this obstacle can be overcome under natural assumptions.

For simplicity, let us suppose that $\mathcal{X} \subseteq \mathbb{R}^d$ (in this section, d is not being used for dimensionality of the report space). The key idea is that, if the distribution $\mathcal{D}_x$ changes slowly as a function of x, then with enough samples, then a set of m close neighbors $x_1, \ldots, x_m$ can be viewed as approximating a single x with m "almost i.i.d." conditional draws $y_1, \ldots, y_m$. We formalize this intuition here using a Lipschitz condition on the total variation distance:

$$D_{TV}(\mathcal{D}_x, \mathcal{D}_{x'}) \leq K\|x - x'\|_2.$$

However, the exact formalization is less important than the general idea, and we expect that future work will be able to prove similar results with a variety of similar assumptions.

Our approach will be to cluster the data into groups of size m having nearby xs, then treat each group as a single sample of the form $(x^*, y_1, \ldots, y_m)$ with each y_i *approximately* i.i.d. from $\mathcal{D}_{x^*}$. We then have n' "samples" of this form, where n' is the number of clusters. Of course, for this approach, it is necessary that that m be small compared to the total number of samples $n \approx n'm$; we are often interested in the $m = 2$ case where our theory and simulations already show dramatic differences from the traditional case of $m = 1$.

A classic risk bound translated into our setting is the following, where R_n denotes the *Rademacher complexity* of a hypothesis class.

Theorem C.1 (Bartlett and Mendelson (2002)) *Suppose ℓ is L-Lipschitz in its first argument and bounded by c, $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a distribution $\mathcal{D}$, and each $\mathbf{y}_i$ is drawn independently from $\mathcal{D}_{x_i}$. Then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,*

$$\text{Risk}(f) \leq \text{Risk}_{\text{emp}}(f, \{x_i, \mathbf{y}_i\}_{i=1}^n) \;+\; 2LR_n(\mathcal{F}) \;+\; c\sqrt{\frac{\log 1/\delta}{2n}}.$$

Here the probability is over the randomness in $\{x_i, \mathbf{y}_i\}$.

In other words, if we could actually sample a set $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,m})$ from $\mathcal{D}_{x_i}$ i.i.d., we would reduce to the standard setting. This theorem is leveraged to prove specific ERM risk bounds depending on $\mathcal{F}$. Here we just show that this bound changes only slightly in the multi-observation case, with an increase in sample complexity.

Our "cluster-points" algorithm roughly functions as follows: draw n i.i.d. data points $x_1^*, \ldots, x_n^*$ and $n' = \Omega(n(m + \log(n/\delta))/\epsilon)$ "scatter points" of the form (x, y). Assign to each x_i^* a set $\mathbf{y}_i^*$ of size m where for each y_{ij}^*, its corresponding x has $\|x - x_i^*\|_2 \leq \epsilon$. We first show that this is possible with probability $1 - \delta$, in two lemmas.

Lemma C.2 *Given $x \in [0,1]$, $\epsilon < 1$ and $\Omega((m+\log(1/\delta'))/\epsilon)$ i.i.d. from the uniform distribution over $[0,1]$, with probability at least $1-\delta'$, at least m of the samples fall within ϵ of x.*

Proof The probability that a given sample falls within ϵ of x is at least ϵ. If we take s samples, then by a standard Chernoff bound we have that the probability of fewer than m samples falling within ϵ of x is upper bounded by

$$e^{-\left(1-\frac{m}{\epsilon s}\right)^2 \epsilon s/2}.$$

Solving for s when this is δ' gives us the Lemma. ■

Lemma C.3 *Let $\mathcal{D}$ be the uniform distribution on $[0,1]$. $n' = O(n(m+\log(n/\delta))/\epsilon)$ samples of the form (x,y) where $x \sim \mathcal{D}$ and $y \sim D_x$ are sufficient to find, with probability at least $1-\delta$, a set of n independent samples of the form $(x^*, y_1^*, \ldots, y_m^*)$ where $x^* \sim \mathcal{D}$ and the y_i^*s are independent and of the form $y_i^* \sim \mathcal{D}_{x'}$ for $|x' - x^*| \leq \epsilon$.*

Proof First we take m samples and use there x values as our m x^*s. For each x^*, we take a new set of $n'/m = O((m+\log(n/\delta))/\epsilon)$ samples $(x_1,y_1),\ldots,(x_{n'/m},y_{n'/m})$. Let $j_1,\ldots,j_m$ be m distinct indices such that for all i, $|x_{j_i} - x^*| \leq \epsilon$. By Lemma C.2 (setting $\delta' = \delta/n$) such a set will exist with probability at least $1-\delta/n$. We then construct the sample

$$(x^*, y_1^*, \ldots, y_m^*) = (x^*, y_{j_1}, \ldots, y_{j_m}).$$

By a union bound, this algorithm will succeed with probability at least $1-\delta$, and the produced samples trivially fulfill the distributional requirements of the Lemma. ■

Now we obtain the desired result. Note that we can choose ϵ as small as desired, e.g. $\epsilon = 1/n^2$, with a blowup of $1/\epsilon$ in the sample complexity. However, a more sophisticated bound would preferably use higher-powered concentration inequalities or a more carefully tailored assumption in order to get a bound holding with higher probability.

Theorem C.4 *Suppose ℓ is L-Lipschitz in its first argument and bounded by c, $\mathcal{D}$ is uniform on $[0,1]$, and $\{x_i^*, \mathbf{y}_i\}_{i=1}^n$ are drawn according to our cluster-points algorithm, taking $n' = O((m+\log(n/\delta))/\epsilon)$ total samples. Then with probability at least $1-2\delta-mnK\epsilon$, for all $f \in \mathcal{F}$,*

$$\mathrm{Risk}(f) \leq \mathrm{Risk}_{\mathrm{emp}}(f, \{x_i^*, \mathbf{y}_i\}_{i=1}^n) \;+\; 2LR_n(\mathcal{F}) \;+\; c\sqrt{\frac{\log 1/\delta}{2n}}.$$

Again the probability is over the randomness in $\{x_i^*, \mathbf{y}_i\}$.

Proof With probability $1-\delta$, our "cluster-points" algorithm succeeds in finding $\{x_i^*\}_{i=1}^n$ drawn i.i.d. and $\{\mathbf{y}_i\}_{i=1}^n$ drawn from ϵ-close points. We wish to consider $\mathrm{Risk}_{\mathrm{emp}}(f, \{x_i^*, \mathbf{y}_i\}_{i=1}^n)$, where each $\mathbf{y}_i$ is $Km\epsilon$-close in total variation distance to $\mathbf{y}_i^*$, as each member is $K\epsilon$ close. So the whole quantity, by the properties of total variation distance, is $mnK\epsilon$-close to $\mathrm{Risk}_{\mathrm{emp}}(f, \{x_i, \mathbf{y}_i\}_{i=1}^n)$, and we apply Theorem C.1. ■

Appendix D. Zero sets of polynomials over the real numbers

Consider a polynomial $f(x_1, \ldots, x_n)$ in the set $\mathbb{R}[x_1, \ldots, x_n]$ of polynomials in n variables with real coefficients. The *zero set* of $f(x_1, \ldots, x_n)$ is by definition the set

$$Z(f(x_1, \ldots, x_n)) := \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n : f(\alpha_1, \ldots, \alpha_n) = 0\} \subseteq \mathbb{R}^n.$$

Recall that a nonconstant polynomial $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ is said to be *irreducible* if it cannot be written as the product of two polynomials in $\mathbb{R}[x_1, \ldots, x_n]$ of strictly lower degree. Recall also that a subset $U \subseteq \mathbb{R}^n$ is said to be *open in the Euclidean topology* if for every $\alpha = (\alpha_1, \ldots, \alpha_n) \in U$, there exists a real number $\epsilon_\alpha > 0$, depending on α, such that the ball of radius ϵ_α centered at α, $B_{\epsilon_\alpha}(\alpha_1, \ldots, \alpha_n)$, is contained in U:

$$B_{\epsilon_\alpha}(\alpha_1, \ldots, \alpha_n) := \left\{(\beta_1, \ldots, \beta_n) \in \mathbb{R}^n : \sqrt{(\beta_1 - \alpha_1)^2 + \cdots + (\beta_n - \alpha_n)^2} < \epsilon_\alpha\right\} \subseteq U.$$

With this terminology, we can state the following theorem:

Theorem D.1 *Suppose that $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ is a nonconstant irreducible polynomial, and $U \subseteq \mathbb{R}^n$ is an open subset in the Euclidean topology. If there is a point*

$$(\alpha_1, \ldots, \alpha_n) \in Z(f(x_1, \ldots, x_n)) \cap U \subseteq \mathbb{R}^n$$

such that

$$\left(\frac{\partial f}{\partial x_1}(\alpha_1, \ldots, \alpha_n), \ldots, \frac{\partial f}{\partial x_n}(\alpha_1, \ldots, \alpha_n)\right) \neq (0, \ldots, 0) \in \mathbb{R}^n, \tag{5}$$

then there are no nonzero polynomials of degree less than the degree of $f(x_1, \ldots, x_n)$ that vanish at every point of the zero set $Z(f(x_1, \ldots, x_n)) \cap U$.

We expect the theorem is well known; for instance, the case where $U = \mathbb{R}^n$ is a special case of (Bochnak et al., 1998, Thm. 4.5.1). The proof of (Bochnak et al., 1998, Thm. 4.5.1) easily generalizes to our situation. For the convenience of the reader, in Theorem E.1 below we include a generalization of (Bochnak et al., 1998, Thm. 4.5.1) that impiles Theorem D.1.

Remark D.2 (Checking the conditions of Theorem D.1) *There are many techniques for checking that a polynomial $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ is irreducible and satisfies the condition (5) for all $(\alpha_1, \ldots, \alpha_n) \in Z(f(x_1, \ldots, x_n)) \cap U$, and therefore satisfies the hypotheses of Theorem D.1. For $n \geq 2$, we recall the following elementary condition that suffices. Suppose $f(x_1, \ldots, x_n)$ is a nonconstant polynomial of degree d. The homogenization of $f(x_1, \ldots, x_n)$ is the degree d homogeneous (all monomials of degree d) polynomial $F(X_0, X_1, \ldots, X_n) \in \mathbb{R}[X_0, \ldots, X_n]$ that is obtained from $f(x_1, \ldots, x_n)$ by replacing x_i with X_i for $i = 1, \ldots, n$, and then multiplying each monomial by a power of X_0 until it is of degree d. For instance, if $f(x_1, x_2) = x_1^2 + 2x_2 + 3$, then $F(X_0, X_1, X_2) = X_1^2 + 2X_0X_2 + 3X_0^2$. If the complex zero set*

$$\left\{(\alpha_0, \ldots, \alpha_n) \in \mathbb{C}^{n+1} : \frac{\partial F}{\partial X_0}(\alpha_0, \ldots, \alpha_n) = \cdots = \frac{\partial F}{\partial X_n}(\alpha_0, \ldots, \alpha_n) = (0, \ldots, 0)\right\} \subseteq \mathbb{C}^{n+1} \tag{6}$$

is equal to $\{(0, \ldots, 0)\}$ or $\emptyset$, then $f(x_1, \ldots, x_n)$ is irreducible and satisfies (5) for all $(\alpha_1, \ldots, \alpha_n) \in Z(f(x_1, \ldots, x_n))$. This is by no means a necessary condition for $f(x_1, \ldots, x_n)$ to satisfy the conditions of Theorem D.1, but it is easy to implement in examples. There are a number of other techniques that can be used, including using computer algebra systems.

Using the technique outlined in the remark, and standard results in algebraic geometry, it is elementary to establish the following corollary:

Corollary D.3 *Let $n \geq 2$, let $U \subseteq \mathbb{R}^n$ be a nonempty open subset in the Euclidean topology, let $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$, and for each $c \in \mathbb{R}$ define*

$$f_c(x_1, \ldots, x_n) := f(x_1, \ldots, x_n) + c.$$

Let $F_c(X_0, \cdots, X_n)$ be the homogenization of $f_c(x_1, \ldots, x_n)$.

If for some $c_0 \in \mathbb{R}$ the complex zero set (6) for $F_{c_0}(X_0, \ldots, X_n)$ is equal to $\{(0, \ldots, 0)\} \subseteq \mathbb{C}^{n+1}$ or $\emptyset$, then there is a nonempty open subset $B \subseteq \mathbb{R}$ in the Euclidean topology such that for all $c \in B$, there are no nonzero polynomials of degree less than d that vanish at every point of the zero set $Z(f_c(x_1, \ldots, x_n)) \cap U$.

As a consequence:

Example D.1 *For a given pair of natural numbers n and d with $n \geq 2$, suppose that:*

- *For $c \in \mathbb{R}$, we set $f_c(x_1, \ldots, x_n) := x_1^d + \cdots + x_n^d + (1 - x_1 - \cdots - x_n)^d + c$.*
- *$U := \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n : \alpha_1, \ldots, \alpha_n > 0,\ \sum_{i=1}^n \alpha_i < 1\}$.*

We can confirm using the approach in Remark D.2 that f_c is irreducible for all $c \leq 0$:

$$\begin{aligned} F_c &= cX_0^d + X_1^d + \cdots + X_n^d + (X_0 - X_1 - \cdots - X_n)^d \\ \partial_{X_0} F_c &= cdX_0^{d-1} + d(X_0 - X_1 - \cdots - X_n)^{d-1}, \\ \partial_{X_1} F_c &= dX_1^{d-1} - d(X_0 - X_1 - \cdots - X_n)^{d-1}, \\ &\vdots \\ \partial_{X_n} F_c &= dX_n^{d-1} - d(X_0 - X_1 - \cdots - X_n)^{d-1}. \end{aligned}$$

To use Corollary D.3, we need to consider the complex zero set (6):

$$\left\{(\alpha_0, \ldots, \alpha_n) \in \mathbb{C}^{n+1} : \partial_{X_0} F_c(\alpha_0, \ldots, \alpha_n) = \cdots = \partial_{X_n} F_c(\alpha_0, \ldots, \alpha_n) = (0, \ldots, 0)\right\} \subseteq \mathbb{C}^{n+1},$$

and show that for some $c \in \mathbb{R}$ it is either empty or equal to $\{(0, \ldots, 0)\}$. We consider the case $c = 0$, for simplicity. Under this assumption, we have

$$0 = \partial_{X_0} F_c = cdX_0^{d-1} + d(X_0 - X_1 - \cdots - X_n)^{d-1} \iff (X_0 - X_1 - \cdots - X_n) = 0.$$

Then, assuming $X_0 - X_1 - \cdots - X_n = 0$, we have for $i = 1, \ldots, n$ that

$$0 = \partial_{X_i} F_c = dX_i^{d-1} - d(X_0 - X_1 - \cdots - X_n)^{d-1} \iff X_i = 0.$$

With this new information, returning to $\partial_{X_0} F_c$, we see that we also must have

$$X_0 = 0.$$

In other words, the complex zero set is $\{(0, \ldots, 0)\} \subseteq \mathbb{C}^n$, so that our example satisfies the conditions of Corollary D.3.

In fact, there exists a nonempty open subset $B \subseteq \mathbb{R}$ in the Euclidean topology such that for all $c \in B$, there are no nonzero polynomials of degree less than d that vanish at every point of $Z(f_c(x_1, \ldots, x_n)) \cap U$. For $n \geq 3$, we may take $B = \{c \in \mathbb{R} : Z(f_c(x_1, \ldots, x_n)) \cap U \neq \emptyset\}$.

Remark D.4 *Most polynomials $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$, $n \geq 2$, satisfy the hypotheses of Corollary D.3. More precisely, there is a dense open subset (the complement of linear subspace) of an $\binom{n+d}{d}$-dimensional real vector space that parameterizes degree-d polynomials in n variables. That subset contains a dense open subset Ω (the complement of the discriminant locus; see e.g., Fulton (1998)) such that every $f(x_1, \ldots, x_n) \in \Omega$ satisfies the hypotheses of the corollary; i.e., there is some $c_0 \in \mathbb{R}$ (for instance $c_0 = 0$) such that the complex zero set (6) for $F_{c_0}(X_0, \ldots, X_n)$ is equal to $\{(0, \ldots, 0)\} \subseteq \mathbb{C}^{n+1}$ or $\emptyset$. On the other hand, as described in Example D.2 below, it is easy to find polynomials $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ of degree $d \geq 2$, and nonempty open subsets $U \subseteq \mathbb{R}^n$, so that for every $c \in \mathbb{R}$ there exist nonzero polynomials of degree less than d that vanish at every point of the zero set $Z(f_c(x_1, \ldots, x_n)) \cap U$.*

Example D.2 *Consider the polynomial $f(x_1, \ldots, x_n) = x_1^2$, and take $U = \mathbb{R}_{>0} \times \mathbb{R}^{n-1}$. Then for every $c \in \mathbb{R}$ there is a linear polynomial that vanishes at every point of $Z(f_c(x_1, \ldots, x_n)) \cap U$; for $c > 0$, we can take any linear polynomial, and for $c \leq 0$, we can take $x_1 - \sqrt{-c}$.*

We can construct many more similar examples in the following way. Let $h(x_1) \in \mathbb{R}[x_1]$ be a polynomial of degree at least 2. We have for every $c \in \mathbb{R}$ that $h(x_1) + c$ factors in $\mathbb{R}[x_1]$ as a product of linear terms and a product of quadratic terms each having no real root. For simplicity, let us assume that for all $c \neq 0$, the polynomial $h(x_1) + c$ has a root that is not real; e.g., $h(x_1) = x_1^m$ for some natural number $m \geq 3$. Let $g(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ be any *nonconstant polynomial. Let λ be a real root of $h(x_1)$ (if there is one), and let U be the complement of the zero set of $g(x_1, \ldots, x_n) - \lambda$, or simply $\mathbb{R}^n$ if there is no real root. Then $f(x_1, \ldots, x_n) := h(g(x_1, \ldots, x_n))$ has the property that for every $c \in \mathbb{R}$, there is a nonzero polynomial of degree less than the degree of $f(x_1, \ldots, x_n)$ that vanishes at every point of the zero set $Z(f_c(x_1, \ldots, x_n)) \cap U$.*

Remark D.5 *Theorem D.1 and Corollary D.3 are not interesting in the case $n = 1$. For $f(x) \in \mathbb{R}[x]$ (irreducible or not) there are no nonzero polynomials of degree less than d that vanish at every point of the zero set $Z(f(x)) \cap U$ if and only if* all of the roots of $f(x)$ are real, distinct, and lie in U. *There are standard techniques to check this condition (e.g., (Bochnak et al., 1998, pp.12–14)). In Example D.1 with $n = 1$, by inspection one finds that for $d = 1, 2$ the condition holds if and only if $-1 < c < 1$, and for $d = 3, 4$ the condition does not hold for any c.*

Appendix E. The real Nullstellenstatz for principal ideals and open sets

The main goal of this section is to prove the following theorem generalizing the well known real Nullstellenstatz for principal ideals (e.g., (Bochnak et al., 1998, Thm. 4.5.1)) to allow for Euclidean open sets.

Theorem E.1 *Let $\mathbb{K}$ be a real closed field (e.g., $\mathbb{K} = \mathbb{R}$). Let $f(x_1, \ldots, x_n) \in \mathbb{K}[x_1, \ldots, x_n]$, and let $U \subseteq \mathbb{K}^n$ be an open subset in the Euclidean topology. Suppose that*

$$f(x_1, \ldots, x_n) = f_1(x_1, \ldots, x_n)^{m_1} \cdots f_r(x_1, \ldots, x_n)^{m_r} \tag{7}$$

is a factorization into powers of distinct nonconstant irreducible polynomials. The following are equivalent:

1. $(f) = I(Z(f) \cap U)$.

2. $m_1 = \cdots = m_r = 1$ *and for each* $i = 1, \ldots, r$ *there is a point* $\alpha^{(i)} \in Z(f_i) \cap U$ *with*
$$(\partial_{x_1} f_i(\alpha^{(i)}), \ldots, \partial_{x_n} f_i(\alpha^{(i)})) \neq 0 \in \mathbb{K}^n.$$
For $\mathbb{K} = \mathbb{R}$*, this is equivalent to having for each* i *that* $Z(f_i) \cap U$ *is a smooth* $(n-1)$*-dimensional submanifold of an open neighborhood of* $\alpha^{(i)}$*.*

3. $m_1 = \cdots = m_r = 1$ *and for each* $i = 1, \ldots, r$ *the sign of the polynomial* f_i *changes on an open ball in* U *(i.e., for* $i = 1, \ldots, n$ *there is an open ball* $B_\epsilon^{(i)} \subseteq U$ *and points* $\alpha^{(i)}, \beta^{(i)} \in B_\epsilon^{(i)}$ *such that* $f_i(\alpha^{(i)}) f_i(\beta^{(i)}) < 0$*).*

4. $m_1 = \cdots = m_r = 1$ *and for each* $i = 1, \ldots, r$ *the semi-algebraic Krull dimension of the topological space* $Z(f_i) \cap U$ *(i.e., the Krull dimension of the ring* $\mathbb{K}[x_1, \ldots, x_n]/I(Z(f_i) \cap U)$*) satisfies*
$$\dim(Z(f_i) \cap U) = n - 1.$$

We expect this result is known to the experts (the case where f is irreducible and $U = \mathbb{R}^n$ is (Bochnak et al., 1998, Thm. 4.5.1)), but for lack of a reference we provide a proof in §E.6. See §E.1 for an explanation of the notation.

Remark E.2 *The case* $n = 1$ *is elementary and has the following simple interpretation:* we have $(f(x)) = I(Z(f(x)) \cap U)$ if and only if all of the roots of $f(x)$ in an algebraic closure $\overline{\mathbb{K}}$ are distinct, and lie in $U \subseteq \mathbb{K}$. *There are standard techniques to check this condition (e.g., (Bochnak et al., 1998, pp.12–14)).*

Remark E.3 *If* $f(x_1, \ldots, x_n)$ *is given as in* (7)*, then* $\sqrt{(f)} = (f_1 \cdots f_r)$*. Thus Theorem E.1 also gives conditions for when there is an equality* $\sqrt{(f)} = I(Z(f) \cap U)$*.*

E.1. Notation and conventions

Let K be a field. Given an ideal $I \subseteq K[x_1, \ldots, x_n]$ we will be interested in both the closed subscheme
$$V(I) \subseteq \mathbb{A}^n_K,$$
as well as the zero set
$$V(I)(\operatorname{Spec} K) \simeq Z_K(I) := \{\alpha \in K^n : f(\alpha) = 0, \text{ for all } f \in I\} \subseteq K^n.$$
If the field is clear from the context, we will write $Z(I) = Z_K(I)$. For a subset $S \subseteq K^n$, we denote as usual the ideal of polynomials vanishing on S as
$$I(S) := \{g(x_1, \ldots, x_n) \in K[x_1, \ldots, x_n] : g(s) = 0 \text{ for all } s \in S\}.$$

We refer the reader to (Bochnak et al., 1998, Def. 1.1.9, Def. 1.2.1) for a review of the definition of a real closed field. In particular, such a field $\mathbb{K}$ is of characteristic 0 and is an ordered field; the Euclidean topology on $\mathbb{K}^n$ then has a basis given by the open balls
$$B_\epsilon(\alpha) := \{\beta \in \mathbb{K}^n : \sum_{i=1}^n (\beta_i - \alpha_i)^2 < \epsilon^2\}$$
for all $\alpha \in \mathbb{K}^n$ and all $\epsilon \in \mathbb{K}$.

E.2. The principal Nullstellensatz

For an ideal $I \subseteq K[x_1, \ldots, x_n]$, there is a natural inclusion

$$\sqrt{I} \subseteq I(Z(I)). \tag{8}$$

Hilbert's Nullstellensatz asserts that over an algebraically closed field $\overline{K} = K$, this inclusion is an equality. Focusing on principal ideals, this reads

$$\sqrt{(f)} = I(Z(f)), \quad (K = \overline{K}); \tag{9}$$

in other words $(f) = I(Z(f))$ whenever f is reduced and $K = \overline{K}$ is algebraically closed.

Over nonalgebraically closed fields (9) clearly fails; i.e., one may have

$$\sqrt{(f)} \subsetneq I(Z(f)).$$

For instance, trivially, one has in $\mathbb{Q}[x]$ that $\sqrt{(x^2+1)} = (x^2+1) \subsetneq \mathbb{Q}[x] = I(\emptyset) = I(Z(x^2+1))$. The following example is a little more interesting:

Example E.1 *Consider $f(x,y) = x^2 + y^2 - x^3 \in \mathbb{R}[x,y]$, and the zero set $Z(f) \subseteq \mathbb{R}^2$. It is a cubic plane curve with an isolated point at $(0,0) \in \mathbb{R}^2$. In particular, if we take $U = B_\epsilon(0,0)$ to be a small ball around $(0,0)$ in $\mathbb{R}^2$, then $(x^2+y^2-x^3) \neq (x,y) = I(Z(x^2+y^2-x^3) \cap U)$. On the other hand, it is true that $(x^2+y^2-x^3) = I(Z(x^2+y^2-x^3))$.*

E.3. The connection with dimension

Proposition E.4 *Let $f(x_1, \ldots, x_n) \in K[x_1, \ldots, x_n]$ be a nonconstant irreducible polynomial, and let $U \subseteq K^n$ be* any *subset. The following are equivalent:*

1. $(f) = I(Z(f) \cap U)$.
2. *The semi-algebraic Krull dimension of the topological space $Z(f) \cap U$ (i.e., the Krull dimension of the ring $K[x_1, \ldots, x_n]/I(Z(f) \cap U)$) satisfies*

$$\dim(Z(f) \cap U) = n - 1.$$

Proof (1) $\implies$ (2). By assumption we have $(f) = I(Z(f) \cap U)$. Now the Krull dimension of $K[x_1, \ldots, x_n]$ is n (e.g., (Atiyah and Macdonald, 1969, Exe. 11.7)). Consequently, since f is neither a zero divisor nor a unit, we have that the Krull dimension of $K[x_1, \ldots, x_n]/(f)$ is $(n-1)$ (e.g., (Atiyah and Macdonald, 1969, Cor. 11.7); using that f is irreducible, this is even easier). Note that this direction does not require that f be irreducible.

(2) $\implies$ (1). We have inclusions

$$(f) \subseteq I(Z(f) \cap U) \subseteq K[x_1, \ldots, x_n]. \tag{10}$$

As above, since f is neither a zero divisor nor a unit, we have that the Krull dimension of the ring $K[x_1, \ldots, x_n]/(f)$ is $(n-1)$. By assumption, the Krull dimension of $K[x_1, \ldots, x_n]/I(Z(f) \cap U)$ is also $(n-1)$. Now since (f) is prime (finally using that f is irreducible), and has the same Krull dimension as the ideal $I(Z(f) \cap U)$, it follows from the containment (10) and the definition of Krull dimension that the two ideals are equal. ■

E.4. The connection with smoothness

We say a zero set $Z(I) \subseteq K^n$ is smooth at a point $\alpha \in Z(I)$ if the associated scheme $V(I) \subseteq \mathbb{A}^n_K$ is smooth at the point $(x_1 - \alpha_1, \ldots, x_n - \alpha_n) \in V(I)$. We will also simply say that $V(I)$ is smooth at α. Recall that if $I = (f)$ is principal, and $\alpha \in Z(f)$, then $V(f)$ is smooth at $(x_1 - \alpha_1, \ldots, x_n - \alpha_n)$ if and only if $(\partial_{x_1} f(\alpha), \ldots, \partial_{x_n} f(\alpha)) \neq 0 \in K^n$.

Lemma E.5 *Suppose* $\mathrm{char}(K) = 0$. *Let* $f(x_1, \ldots, x_n) \in K[x_1, \ldots, x_n]$ *be a nonconstant polynomial, and let* $U \subseteq K^n$ *be any subset. Then:*

1. $(f) = I(Z(f) \cap U)$.

implies

(2) *There is a point* $\alpha^{(0)} \in Z(f) \cap U$ *with*

$$(\partial_{x_1} f(\alpha^{(0)}), \ldots, \partial_{x_n} f(\alpha^{(0)})) \neq 0 \in K^n.$$

In other words, there is a point in U *at which* $V(f)$ *is a smooth scheme.*

Proof Suppose that (2) fails. This means that $\partial_{x_1} f, \ldots, \partial_{x_n} f \in I(Z(f) \cap U)$. But since f is nonconstant, and $\mathrm{char}(K) = 0$, there is an i such that $\partial_{x_i} f$ is nonzero. Since $\partial_{x_i} f$ is of degree less than the degree of f, it cannot be a multiple of f, and therefore is not in (f). Thus $(f) \subsetneq I(Z(f) \cap U)$, and (1) fails. ■

The following example shows why there is the characteristic 0 hypothesis on K in Lemma E.5.

Example E.2 *Let* $k = \overline{K}$ *be an algebraically closed field of characteristic* 2, *and consider*

$$f(x_1, x_2) = x_1^2 + x_2^2 - 1 \in k[x_1, x_2].$$

This polynomial is irreducible, so by the Nullstellensatz we have $(f) = I(Z(f))$. *However, we also have that* $\frac{\partial}{\partial x_1} f(x_1, x_2) = \frac{\partial}{\partial x_2} f(x_1, x_2) = 0 \in k[x_1, x_2]$.

The following example shows that the converse to Lemma E.5 need not hold.

Example E.3 *Let* $K = \mathbb{Q}$ *and let* $f(x_1, x_2) = x_1^3 + x_2^3 - 1$. *Then* $Z(f) \subseteq \mathbb{Q}^2$ *is a finite set of points, and in particular one can show that* $(f) \subsetneq I(Z(f))$. *On the other hand, at the point say* $(1, 0) \in Z(f)$, *one has* $(\partial_{x_1} f(1, 0), \partial_{x_2} f(1, 0)) = (3, 0) \neq 0 \in \mathbb{Q}^n$.

Nevertheless, a converse to Lemma E.5 does hold over the real and complex numbers. This is essentially because the implicit function theorem asserts that condition (2) implies that the zero set is an $(n-1)$-dimensional manifold in a neighborhood of the given point. In fact, one can also establish the converse over real closed fields:

Lemma E.6 *Suppose* $K = \mathbb{K}$ *is real closed or equal to* $\mathbb{C}$. *Let* $f(x_1, \ldots, x_n) \in \mathbb{K}[x_1, \ldots, x_n]$ *be a nonconstant irreducible polynomial, and let* $U \subseteq \mathbb{K}^n$ *be an open subset in the Euclidean topology. Then:*

1. $(f) = I(Z(f) \cap U)$.

is implied by

(2) *There is a point* $\alpha^{(0)} \in Z(f) \cap U$ *with*

$$(\partial_{x_1} f(\alpha^{(0)}), \ldots, \partial_{x_n} f(\alpha^{(0)})) \neq 0 \in \mathbb{K}^n.$$

In other words, there is a point in U *at which* $V(f)$ *is a smooth scheme.*

Proof When $K = \mathbb{C}$, we can simply use the Nullstellensatz. In fact, if $K = \mathbb{C}$, condition (2) follows from the fact that f is irreducible (so long as $Z(f) \cap U$ is nonemepty), and (1) follows independently from (2), via the Nullstellensatz.

Consider now the case $K = \mathbb{K}$ is real closed. Let $\overline{(Z(f) \cap U)}^{\text{Zar}} \subseteq \mathbb{K}^n$ be the closure in the Zariski topology. Now using condition (2), and (iii) $\implies$ (ii) of (Bochnak et al., 1998, Prop. 3.3.10), we have that $\dim \mathbb{K}[x_1, \ldots, x_n]/I(\overline{(Z(f) \cap U)}^{\text{Zar}}) = n - 1$. (We are applying (Bochnak et al., 1998, Prop. 3.3.10) with $V = \overline{(Z(f) \cap U)}^{\text{Zar}}$ and $P_1 = f$.) Now we observe that $I(Z(f) \cap U) = I(\overline{(Z(f) \cap U)}^{\text{Zar}})$, and conclude that $\dim(Z(f) \cap U) = n - 1$. Note that so far we did not use that f was irreducible, as this is not required in (Bochnak et al., 1998, Prop. 3.3.10). To conclude (1), we use Proposition E.4, and the assumption that f is irreducible. ■

E.5. The connection with the sign of the polynomial

Lemma E.7 *Suppose* $K = \mathbb{K}$ *is real closed. Let* $f(x_1, \ldots, x_n) \in \mathbb{K}[x_1, \ldots, x_n]$ *be a nonconstant irreducible polynomial, and let* $U \subseteq \mathbb{K}^n$ *be an open subset in the Euclidean topology. Then the following are equivalent:*

1. $(f) = I(Z(f) \cap U)$.
2. *The sign of the polynomial* f *changes on an open ball in* U *(i.e., there is an open ball* $B_\epsilon \subseteq U$ *such that* $f(\alpha)f(\beta) < 0$ *for some* $\alpha, \beta \in B_\epsilon$*).*

Proof (1) $\implies$ (2). Assuming (1), then from Lemma E.5, there is a point $\alpha^{(0)} \in Z(f) \cap U$ with $(\partial_{x_1} f(\alpha^{(0)}), \ldots, \partial_{x_n} f(\alpha^{(0)})) \neq 0 \in \mathbb{K}^n$. In other words, there is an i such that $\partial_{x_i} f(\alpha^{(0)}) \neq 0$. Then consider the polynomial in one variable

$$\phi(x_i) := f(\alpha_1^{(0)}, \ldots, x_i, \ldots, \alpha_n^{(0)}).$$

We have $\phi(\alpha_i^{(0)}) = 0$. But since $\phi'(\alpha_i^{(0)}) = \partial_{x_i} f(\alpha_i^{(0)})$ is non-zero, the function $\phi(x_i)$ is monotone in a real interval around $\alpha_i^{(0)}$, and so it changes sign (Bochnak et al., 1998, Cor. 1.2.7). Therefore f changes sign. (Note that we did not use that f was irreducible.)

(2) $\implies$ (1). (Bochnak et al., 1998, Lem. 4.5.2) states the following: *Let* $B_\epsilon \subseteq \mathbb{K}^n$ *be an open ball (including the case where* $B_\epsilon = \mathbb{K}^n$*) and let* U_1 *and* U_2 *be two disjoint nonempty semi-algebraic open subsets of* B_ϵ*. Then we have* $\dim(B_\epsilon - (U_1 \cup U_2)) \geq n - 1$. Now apply this in our situation, with

$$U_1 = \{\alpha \in B_\epsilon : f(\alpha) > 0\} \text{ and } U_2 = \{\alpha \in B_\epsilon : f(\alpha) < 0\},$$

so that $B_\epsilon - (U_1 \cup U_2) = Z(f) \cap B_\epsilon$. Then

$$n - 1 = \dim Z(f) \geq \dim(Z(f) \cap U) \geq \dim(Z(f) \cap B_\epsilon) \geq n - 1.$$

As mentioned above, we have the equality $\dim Z(f) = n - 1$ on the left since f is neither a zero divisor nor a unit. Note that so far we did not use that f was irreducible. To conclude (1), we use Proposition E.4, and the assumption that f is irreducible. ■

E.6. Proof of Theorem E.1

Proof [Proof of Theorem E.1]

We have now proved the theorem under the hypothesis that f is irreducible (Proposition E.4, Lemma E.5, Lemma E.6, Lemma E.7). We now reduce to this case.

First, it is clear that (2) $\iff$ (3) $\iff$ (4), from the irreducible case. Also, it is clear that if (1) holds (i.e., $(f) = I(Z(f) \cap U)$), we must have that $m_1 = m_2 = \cdots = m_r = 1$. Indeed, if say $m_1 > 1$, then $f_1 f_2^{m_2} \cdots f_r^{m_r} \in I(Z(f) \cap U)$, but for degree reasons $f_1 f_2^{m_2} \cdots f_r^{m_r}$ is not a multiple of $f = f_1^{m_1} \cdots f_r^{m_r}$ and thus (1) fails. So from here on, we assume $m_1 = m_2 = \cdots = m_r = 1$.

(1) $\implies$ (2). Suppose that (2) fails. Then there is some i, j so that $\partial_{x_j} f_i \in I(Z(f_i) \cap U)$ and is nonzero. Therefore $f_1 \cdots \partial_{x_j} f_i \cdots f_r \in I(Z(f) \cap U))$ and is nonzero. But for degree reasons, it is not a multiple of $f = f_1 \cdots f_i \cdots f_r$ and thus (1) fails.

(2) $\implies$ (1). This follows from the fact that

$$\begin{aligned}
\bigcap_{i=1}^r (f_i) &= \prod_{i=1}^r (f_i) && (\mathbb{K}[x_1, \ldots, x_n] \text{ is a UFD}) \\
&= (f) \\
&\subseteq I(Z(f) \cap U)) \\
&= I\left(\bigcup_{i=1}^r (Z(f_i) \cap U) \right) \\
&= \bigcap_{i=1}^r I(Z(f_i) \cap U),
\end{aligned}$$

since, assuming (2) and the special case Theorem E.1 for irreducible polynomials, then for all i, we have $(f_i) = I(Z(f_i) \cap U)$, forcing the containment above to be an equality. ■