JMLR: Workshop and Conference Proceedings vol 49:1–4, 2016

Open Problem: Property Elicitation and Elicitation Complexity

Rafael Frongillo RAF@COLORADO.EDU
CU Boulder

Ian A. Kash IANKASH@MICROSOFT.COM
Microsoft Research

Stephen Becker STEPHEN.BECKER@COLORADO.EDU
CU Boulder

Abstract

The study of property elicitation is gaining ground in statistics and machine learning as a way to view and reason about the expressive power of emipirical risk minimization (ERM). Yet beyond a widening frontier of special cases, the two most fundamental questions in this area remain open: which statistics are elicitable (computable via ERM), and which loss functions elicit them? Moreover, recent work suggests a complementary line of questioning: given a statistic, how many ERM parameters are needed to compute it? We give concrete instantiations of these important questions, which have numerous applications to machine learning and related fields.

Keywords: Property elicitation, empirical risk minimization

1. Introduction

Over the last several decades, several formalisms for studying empirical risk minimization (ERM) have appeared, using notions such as Fisher consistency, PAC learnability, generalization, etc. A relatively recent but powerful framework is that of *property elicitation*, wherein one asks what ERM computes *as a function of the empirical distribution* or data set. Thus, instead of thinking about minimizing classification error (the number of incorrect label predictions for a classification task), we would ask ERM to compute the mode of the conditional label distribution ("what is the most likely label?"). This framework is in many ways equivalent to previous notions, but develops a more explicit framework for studying ERM; instead of simply saying that two losses behave the same way, we can say that they *elicit* the same statistic,[1] giving us another concrete object with which to work. Moreover, the statistic itself is the central object of study in many cases, such as for financial risk measures (Gneiting, 2011; Fissler and Ziegel, 2015; Frongillo and Kash, 2015c).

Beyond the application to ERM, viewing the statistic and loss separately can enable the study of many problems under one roof. The notion of elicitation arises in several other fields, including statistics, forecasting, economics, and finance, though it goes by other names depending on the context: consistency, calibration, incentive-compatability, and propriety. These different names reflect the many ways in which losses are used: to estimate or predict, to evaluate existing predictions, to detect outliers, and to derive payments or contracts for information. As put by Fissler et al. (2015), "it is generally accepted that elicitability is useful for model selection, estimation, generalized regression, forecast comparison, and forecast ranking." Thus, elicitation gives a very powerful vantage point which extends beyond machine learning.

1. We use the terms "property" and "statistic" interchangeably.

© 2016 R. Frongillo, I.A. Kash & S. Becker.

2. Elicitation

We now give some of the basic definitions needed to state our open problems. We first need three sets: the data or outcomes, the distributions over those outcomes, and a range of values for the property/statistic we are interested in.

- **Outcomes/data** Y, an arbitrary set, usually $Y \subseteq \mathbb{R}$
- **Distributions** $\mathcal{P}$, a (typically convex) set of probability measures over Y
- **Reports** $\mathcal{R}$, the "regression parameters" with which ERM works, usually $\mathcal{R} \subseteq \mathbb{R}^k$

We can now define a property and loss, and what it means for the loss to elicit the property. It is important to note that while we define everything here in terms of probability measures, the framework still applies to ERM, where of course one's data comprises a finite set (cf. Steinwart et al. (2014)).

- **Property** $\Gamma(p)$, a function $\Gamma : \mathcal{P} \to \mathcal{R}$ specifying the correct report $r = \Gamma(p)$ for each dist. p.
- **Loss** $L(r, y)$, a function $L : \mathcal{R} \times Y \to \mathbb{R}$ scoring the inaccuracy of report r given data point y.
- **Elicit**: a loss L elicits property Γ if:

$$\forall\, p \in \mathcal{P}, \;\; \Gamma(p) = \operatorname*{argmin}_{r \in \mathcal{R}} \; \mathbb{E}_{y \sim p}[L(r, y)] \,. \tag{1}$$

  That is, for all p, the report $r = \Gamma(p)$ minimizes the expected loss assuming y is drawn from p.
- **Elicitable**: property Γ is elicitable if some loss elicts it (formerly "ERM-computable").
- **Surrogate**: losses L and L' are (calibrated) surrogates if they elicit the same statistic.

To make these definitions concrete, let us instantiate them for linear regression. Here $Y = \mathcal{R} = \mathbb{R}$, and $\mathcal{P}$ is some convex set of distributions over $\mathcal{R}$. Then the squared loss $L(r, y) = (r - y)^2$ elicits the property $\Gamma(p) = \mathbb{E}_{y \sim p}[y]$, the mean of p. Thus, the mean is elicitable. Squared loss has many surrogates, taking the form of Bregman divergences (Gneiting, 2011; Abernethy and Frongillo, 2012; Frongillo and Kash, 2015a), including for example $L'(r, y) = e^{-r}(1 + r - y)$.

3. Elicitation Complexity

Consider the task of computing the variance of a distribution (data set) via ERM. The natural approach would be to look for a loss function $L(r, y)$, where r is real-valued, such that the variance of p is the minimizer $r = \mathrm{Var}(p)$ of the expected loss $\mathbb{E}_{y \sim p}[L(r, y)]$. One quickly encounters a problem however: **no such loss exists.** This is a corollary of a result that the *level sets* of elicitable properties must be convex (Osband, 1985; Lambert et al., 2008): since mixtures of distributions with the same variance in general have higher variance, the level sets of the variance function are not convex, and thus it cannot be elicitable.

Fortunately, one can easily circumvent this problem by conceding an extra regression parameter: $L(r_1, r_2, y) = e^{-r_2}((r_1 - y)^2 - r_2 - 1)$ elicits both the mean and the variance together. This simple example illustrates the need for a more nuanced notion of elicitability; just because a statistic is not "directly" elicitable does not mean it cannot be computed via ERM. Instead, the right question to ask is *how many variables* it takes to gather enough information to compute the statistic after the fact. This is the notion of *elicitation complexity*, which we now formalize.

While originally introduced in Lambert et al. (2008), we use the formulation of Frongillo and Kash (2015b,c), who define the elicitation complexity $\mathrm{elic}(\Gamma)$ of a property/statistic Γ as the smallest dimension of an elicitable property Γ' from which we may compute Γ via composition with a link function ψ. (Formally, to avoid trivial pathological constructions, we also require Γ' to be *identifiable*; see the references above for details.) In other words, if we allow the indirect computation of Γ by first computing Γ' via ERM and then transforming the result with ψ, i.e. $\Gamma = \psi \circ \Gamma'$, how large (number of ERM parameters) does Γ' have to be? One should think of elicitation complexity as both a refinement and enhancement of the notion of elicitation: not only does it relax the question "is Γ elicitable" to "how elicitable is Γ", but there are many more properties of interest with low elicitation complexity than there are elicitable properties.

The notion of elicitation complexity has close ties to previous work in machine learning and statistics. The concept of a *link function* which transforms the prediction space dates back to the introduction of generalized linear models (Nelder and Wedderburn, 1972; Nelder and Baker, 2004). Such invertible links are widespread in statistics and machine learning, but importantly do not change the elicitation qualities: if ψ is invertible, then $\Gamma = \psi \circ \Gamma'$ is elicitable if Γ' is. The concept of elicitation complexity only becomes interesting for *non-invertible* links, wherein the ERM calculations are in a richer space than the final prediction. This exactly captures the practice of using ERM over continuous spaces for classification tasks, as in support vector machines (SVM), logistic regression, and boosting: rather than solve for the discrete-valued function which minimizes the classification error directly, a continuous-valued function is sought which minimizes a nice convex objective, and only later is this function truncated to answer the original classification problem (Bartlett et al., 2006). In terms of elicitation, the statistic here is the mode, $\Gamma : p \mapsto \mathrm{mode}(p)$, and the link simlply takes the sign of the prediction $\psi(y) = \mathrm{sgn}(y)$, assuming the labels are $\{+1, -1\}$, and Γ' is the continuous ERM minimizer.

4. The Questions

We now formulate our three open questions. Note that these questions are still interesting when regularity conditions are imposed, such as differentiability of the loss function (with respect to r), continuity or differentiability of the property, discrete or finite distributions (e.g. $\mathcal{P} = \Delta_n$), etc.

Question 1 *Which properties $\Gamma : \mathcal{P} \to \mathbb{R}^k$ are elicitable? We ideally would have a geometric characterization in terms of the level sets $L_r = \{p \in \mathcal{P} : \Gamma(p) = r\}$ of Γ, as in (Lambert et al., 2008, Theorem 1).*

Question 2 *Given an elicitable property $\Gamma : \mathcal{P} \to \mathcal{R}$, characterize all loss functions $L : \mathcal{R} \times Y \to \mathbb{R}$ eliciting Γ.*

Question 3 *Characterize the elicitation complexity classes $C_k = \{\Gamma : \mathrm{elic}(\Gamma) = k\}$.*

Currently, Questions 1 and 2 are only well-understood in the case $\mathcal{R} \subseteq \mathbb{R}$ (Lambert, 2011; Steinwart et al., 2014), and the expectation and ratio-of-expectation cases (Frongillo and Kash, 2015a), where $\Gamma(p) = \mathbb{E}_p[X]$ or $\Gamma(p) = \mathbb{E}_p[X]/\mathbb{E}_p[Z]$ where X is a finite-dimensional random variable, and Z is a real-valued random variable. General ties to convex analysis are given in Frongillo and Kash (2014). Very little is known about Question 3, though Frongillo and Kash (2015c) give tight bounds for some properties (the Bayes risks), which establish that $C_k \neq \emptyset$ for all k. For other related conjectures, examples, and counter-examples, see Frongillo and Kash (2015a,c).

References

J. Abernethy and R. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Conference on Learning Theory*, pages 1–27, 2012. URL http://jmlr.csail.mit.edu/proceedings/papers/v23/abernethy12/abernethy12.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.

Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *arXiv:1503.08123 [math, q-fin, stat]*, March 2015. URL http://arxiv.org/abs/1503.08123. arXiv: 1503.08123.

Tobias Fissler, Johanna F. Ziegel, and Tilmann Gneiting. Expected Shortfall is jointly elicitable with Value at Risk - Implications for backtesting. *arXiv:1507.00244 [q-fin]*, July 2015. URL http://arxiv.org/abs/1507.00244. arXiv: 1507.00244.

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

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

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

N.S. Lambert, D.M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

J. A. Nelder and R. J. Baker. Generalized Linear Models. In *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., 2004. ISBN 978-0-471-66719-3. URL http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess0866.pub2/abstract.

J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 0035-9238. doi: 10.2307/2344614. URL http://www.jstor.org/stable/2344614.

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

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