Characterizing Convex Elicitable Properties

Introduction. In machine learning, we are often interested in learning some "summary statistic" of a dataset. Data are too high-dimensional and complex for people to intuitively understand in full, so we seek these summary statistics, or *properties*, to help us understand the big picture of the dataset. These properties can be written as a function Γ of a probability distribution p over the possible outcomes, and common examples include the mean, median, mode, and variance. These properties correspond to questions that we can ask about our dataset; *"What is the average rainfall?"* asks about the mean of a distribution, while *"Is this more likely to be a picture of a dog or cat?"* asks about the mode.

We learn these properties by minimizing a numerical punishment for error, called a *loss function*, L. The design of these losses is very important: loss functions should be chosen so that the algorithm's expected loss is minimized by predicting the desired summary statistic. We can write our property $\Gamma : \mathscr{P} \to \mathscr{R}$ as a function of a probability distribution p so that $\Gamma(p)$ yields the optimal report for our question, fixing the data distribution p, as in Equation 1. If a property Γ is the minimizer of an expected loss function, we call Γ *elicitable*, and we say that the loss L *elicits* Γ.

Empirical Risk Minimization (ERM) is the basis for most popular machine learning algorithms. We form a hypothesis $h(x)$ based on the input data x in order to predict the output y, and want the hypothesis that minimizes our expected loss. We then define the hypothesis h^* that minimizes our empirical risk in Equation 2.

$$\Gamma(p) = \arg\min_{r \in reports} \mathbb{E}_{Y \sim p} L(r, Y) \tag{1}$$

$$h^* = \arg\min_{h \in \mathscr{H}} \sum_{(x,y) \in data} L(h(x), y) \tag{2}$$

We can see in Equations 1 and 2 that ERM and property elicitation are closely related. Given the distribution p learned from the training data about the input x, the prediction $\Gamma(p)$ is the hypothesis that minimizes empirical risk.

Why does the loss function matter? Minimizing a loss function helps us answer the desired question about the probability distribution underlying the data. Depending on the choice of loss function, we might form a very different hypothesis about future data by minimizing our loss. One popular loss function, seen in red in Figure 1, is squared loss, which is minimized by predicting the expected value over the probability distribution p. In Figure 1, we see fitting a regression with squared loss instead of absolute loss may lead to a very different prediction for the future input x.

Unfortunately, not every property can be written as the minimizer of the expected loss. To see this, consider the variance of a dataset. A corollary of known results from Osband [Osb85] implies that the variance is not elicitable. Thankfully, we can circumvent this issue by allowing for two predictions; we can *indirectly* elicit the variance. Indirectly learning the variance opens up a realm of open questions in the field of property elicitation: we move from asking "*Is this property elicitable?*" to "*How elicitable is this property?*", which has proven to be a much deeper question.

Why convex losses? Consider the 0-1 loss assigning punishment 0 if the prediction is correct, and 1 if incorrect. This loss elicits the mode, or most likely outcome. However, the optimization

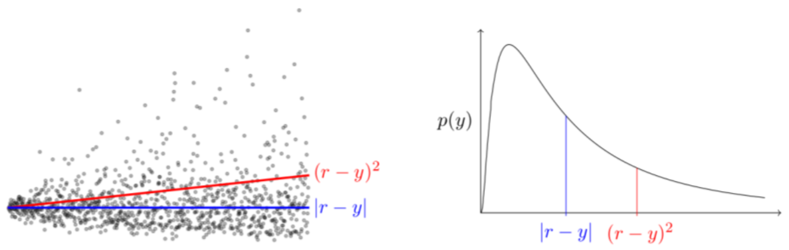

Figure 1: (Left.) Fitting a linear regression to a set of data by using Squared loss and Absolute Loss. (Right.) The probability distribution p generating the data on the left. In red, we see the mean, and in blue, we observe the median.

of most discrete ERM problems, 0-1 loss included, is *NP*-hard. For classification problems, we typically consider continuous ERM problems and threshold our results later. For example, while training a Support Vector Machine, instead of directly labeling a data point, we learn a real-valued model and then threshold results into a discrete model. When a loss function is convex, most optimization algorithms have stronger accuracy and convergence guarantees on the given prediction.

Previous Work. In 2008, Lambert et al. [LPS08] formed a general characterization of directly elicitable, real-valued properties, but these loss functions are not necessarily convex. Recently, my paper [FF18] withmy advisor was accepted to *Neural Information Processing Systems* 2018 as a spotlight, where we worked toward a characterization of which properties are directly elicitable by a convex loss function. To our surprise, we found that essentially any continuous, elicitable property over a finite number of outcomes is elicitable by a convex function. However, the infinite-outcome case (important for regression problems) still remains an open question.

Future Work. Given a property, I want to discover *how (convex) elicitable it is.* One major subtask required to answer this question includes finding some characterization of *vector-valued* properties. That is, when are pairs of properties elicitable together, and can some reports be reused to calculate both properties? While we understand which properties are directly elicitable, we do not know a lower bound on how many reports are needed to indirectly elicit a given property.

Intellectual Merit. Researchers have come up with creative losses to elicit various properties, but we do not understand why these losses work or how they generalize. Because we don't understand why loss functions minimize specific properties, it is rare to know if a given loss is a lower bound to elicit a desired property. Finding a lower bound on the dimension of the optimization problem shows us how tractable a given problem is.

Broader Impacts. As machine learning becomes increasingly utilized for automated decision making in society, we must ask increasingly critical questions of these algorithms. Studying convex elicitation complexity gives us an understanding of which questions we can efficiently answer about a dataset through current machine learning techniques. Understanding *why* we can learn a property of a dataset can shape the narrative of the story machine learning algorithms tell.

References

[FF18] Jessica Finocchiaro and Rafael Frongillo. Convex elicitation of continuous properties. 2018. Accepted as spotlight presentation to NIPS 2018 (top $\sim 4\%$ of submissions).

[LPS08] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, 2008.

[Osb85] Kent Harold Osband. *Providing Incentives for Better Cost Forecasting*. University of California, Berkeley, 1985.