# Machine Learning for Evaluating and Improving Theories

DREW FUDENBERG
MIT
and
ANNIE LIANG
University of Pennsylvania

We summarize our recent work that uses machine learning techniques as a complement to theoretical modeling, rather than a substitute for it. The key concepts are those of the *completeness* and *restrictiveness* of a model. A theory's completeness is how much it improves predictions over a naive baseline, relative to how much improvement is possible. When a theory is relatively incomplete, machine learning algorithms can help reveal regularities that the theory doesn't capture, and thus lead to the construction of theories that make more accurate predictions. Restrictiveness measures a theory's ability to match arbitrary hypothetical data: A very unrestrictive theory will be complete on almost any data, so the fact that it is complete on the actual data is not very instructive. We algorithmically quantify restrictiveness by measuring how well the theory approximates randomly generated behaviors. Finally, we propose "algorithmic experimental design" as a method to help select which experiments to run.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Economics, Experimentation, Measurement, Theory

Additional Key Words and Phrases: machine learning, economic theory, modeling, prediction

## 1. INTRODUCTION

This survey summarizes our recent and ongoing work [Fudenberg and Liang 2019; Fudenberg et al. 2019; Fudenberg et al. 2020] on how to use machine learning techniques to evaluate and then improve theories. Black-box algorithms can generate better predictions than parametric theories, but direct application of these methods generally does not yield an improved understanding into the behavior of interest. We demonstrate how black-box algorithms can nevertheless contribute to this latter objective when used as a complement to traditional modeling.

In Section 2, we define the "completeness" of a theory to be the fraction of *achievable* prediction that it attains, benchmarked against the performance of a fully nonparametric black box. We show by example how studying cases where a machine learning algorithm predicts well, but the theory does not, can allow us to identify new regularities that the theory has not yet captured. A theory that is very complete for prediction of the actual data captures most of the important regularities in the observed behavior. But if a theory can approximate *most* patterns of behavior, then its ability to fit the actual data doesn't speak to its relevance. In Section 3, we quantify the "restrictiveness" of a model by measuring how well it

Authors' addresses: drewf@mit.edu, anliang@upenn.edu

approximates arbitrary behaviors. We illustrate our ideas with two classic prediction problems from experimental economics—predicting certainty equivalents for binary lotteries and predicting initial play in matrix games—and evaluate models in these domains from the dual perspectives of completeness and restrictiveness.

Both of our proposed measures depend on the domain of "test cases" used to evaluate the model, which is generally a choice variable of the experimenter the collecting data. In Section 4, we show how algorithms can help in designing new instances for data collection, for example finding cases in which a given theory is likely to fail. Our work here shows that machine learning techniques are useful not only for identifying structure in given data, but can also be useful to experimenters in figuring out what new data to acquire.

### 1.1 Prediction Problems

Let $x$ be an observable *feature vector* taking values in a set $X$, and let $y$ be an *outcome* of interest taking values in $Y$. An analyst observes pairs $z_i = (x_i, y_i)$, where each $x_i$ takes on a finite set of values that were selected by the analyst, e.g. which games or lotteries to use in a laboratory experiment. We call any function $f : X \to Y$ a *predictive mapping* or simply *mapping*. We are interested in parametric models $\mathcal{F}_\Theta = \{f_\theta\}_{\theta\in\Theta}$, where $\Theta$ is a finite-dimensional, closed, and compact set and $f$ is continuous in $\theta$.

We evaluate predictions with a *loss function*, $\ell : Y \times Y \to \mathbb{R}$, where $\ell(y', y)$ is the error assigned to prediction of $y'$ when the realized outcome is $y$. The commonly used loss functions *mean-squared error* and *classification loss* correspond to $\ell(y', y) = (y' - y)^2$ and $\ell(y', y) = \mathbb{1}(y' \neq y)$.

*Definition* 1.1. Let $P$ denote the joint distribution of $(x, y)$. The *(expected) prediction error for model $f$* is the expected error on a new test case: $\mathcal{E}_P(f) = \mathbb{E}_P[\ell(f(x), y)]$. The *prediction error for a model $\mathcal{F}_\Theta$* is $\mathcal{E}_P(f^*_\Theta)$, where

$$f^*_\Theta = \arg\min_{f\in\mathcal{F}_\Theta} \mathcal{E}_P(f)$$

is the error-minimizing prediction rule from $\mathcal{F}_\Theta$.

Typically, the distribution $P$ is not known, so these quantities need be estimated from the data. For example, to evaluate the error $\mathcal{E}_P(f^*_\Theta)$ for a model $\mathcal{F}_\Theta$, we might estimate its economic parameter $\theta$ from training data, and test the trained mapping $f_\theta$ on new observations. We put aside details of estimation for this survey, and refer interested readers to our papers.

### 1.2 Examples

We illustrate our methodologies using two examples from the economics literature.

*Example* 1 *Risk Preferences.* We consider the problem of predicting *certainty equivalents* for lotteries, i.e. the certain payment that an individual considers equivalent to the lottery's random payment. We use a data set from Bruhin et al. [2010] of the reported certainty equivalents (across different subjects) for a set of 25 binary lotteries over positive prizes. The feature space $X$ is the set of 25 unique tuples $x = (\overline{z}, \underline{z}, p)$ describing the binary lotteries, where $\overline{z} > \underline{z} \geq 0$ are the two prizes, and $p$ is the probability of $\overline{z}$. The outcome to be predicted is a *given* subject's certainty

equivalent for a given lottery, so $Y = \mathbb{R}$. We use mean-squared error as the loss function, so the optimal prediction is the average certainty equivalent in the data.

The economic model that we evaluate is the three-parameter version of *Cumulative Prospect Theory* suggested by Goldstein and Einhorn [1987] and Lattimore et al. [1992]. The parameter vector here is $\theta = (\alpha, \delta, \gamma)$, and the associated model is $f_\theta(\overline{z}, \underline{z}, p) = w(p)v(\overline{z}) + (1 - w(p))v(\underline{z})$, where $w(p) = (\delta p^\gamma) / (\delta p^\gamma + (1-p)^\gamma)$ with $\delta \geq 0$ and $\gamma \geq 0$ is a nonlinear probability weighting function, and $v(z) = z^\alpha$ with $\alpha \geq 0$ is a value function for money.

*Example* 2 *Predicting Play in Games.* Our second example is predicting how people will play the first time they encounter a new simultaneous-move game. We use a data set of play in $3 \times 3$ normal-form games constructed by Wright and Leyton-Brown [2014] from six previous papers. The feature space $X$ is the set of 86 unique payoff matrices $x \in \mathbb{R}^{18}$. The outcome to be predicted is the action that is chosen by the row player in a given instance of play, so $Y = \{a_1, a_2, a_3\}$. We use the misclassification rate as our loss function, so the optimal prediction is the modal action.

The economic model that we evaluate is the *Poisson Cognitive Hierarchy Model* (PCHM), which supposes that there is a distribution over players of differing levels of sophistication: The *level-0* player randomizes uniformly over his available actions, while the *level-1* player best responds to level-0 play [Stahl and Wilson 1994; 1995; Nagel 1995]. Camerer et al. [2004] defines the play of level-$k$ players, $k \geq 2$, to be the best response to a perceived distribution over (lower) opponent levels, which is a Poisson distribution with rate parameter $\tau$ (truncated at $k$ and re-normalized). The parameter $\tau$ is the only free parameter in this model.

## 2. COMPLETENESS

In Fudenberg et al. [2019], we define the "completeness" of a model as the amount that it improves predictions over a naive rule, compared to the best achievable improvement given the available features. We normalize in this way because in many cases there is residual variation in the outcome $y$ after conditioning on the features $x$, and so perfect prediction is not achievable by any mapping that makes predictions using the feature set $X$. The mapping from $X$ to $Y$ that minimizes prediction error is

$$f^*(x) = \arg\min_{y' \in Y} \mathbb{E}_P \left[ \ell(y', y) \mid x \right]. \tag{1}$$

For example, if the outcome $y$ is real-valued, and the loss function is mean-squared error, then $f^*$ assigns to each feature vector $x$ its conditional mean.

To interpret the prediction error of a model, it is useful to distinguish between two sources of error. The **irreducible error** in the prediction problem is the error $\mathcal{E}_P(f^*) = \mathbb{E}_P \left[ \ell(f^*(x), y) \right]$ of the ideal rule on a new test observation. This is a bound on how well any mapping could perform. In addition, there can be error due to the specification of the class: If $\mathcal{F}_\Theta$ leaves out an important regularity, then the prediction error of the best mapping from this class, $\mathcal{E}_P(f^*_\Theta)$ may be substantially higher than the irreducible error, $\mathcal{E}_P(f^*)$.

These two sources of prediction error have very different implications for how to generate better predictions. If the model's prediction error is substantially higher

than the irreducible error, it may be possible to identify new regularities and incorporate them into new models that improve prediction given the same feature set. Conversely, if the model's prediction error is close to the irreducible error for the current feature set, the priority should be to identify additional features that will allow for better predictions.

We define the **completeness** of a model to be the ratio of the reduction in prediction error (over a selected **naive mapping** $f_n$) that it achieves compared to the best possible reduction, which is to the irreducible error. We set the naïve prediction for a lottery's certainty equivalent to be its expected value, and we set the naïve prediction of initial play to be a uniform distribution over the available actions.

*Definition* 2.1. The **completeness of model $\mathcal{F}_\Theta$ is**

$$\frac{\mathcal{E}_P(f_n) - \mathcal{E}_P(f^*_\Theta)}{\mathcal{E}_P(f_n) - \mathcal{E}_P(f^*)}. \tag{2}$$

Table I reports completeness measures for the two economic models and the corresponding prediction tasks described in Section 1.1.

| | Risk Preferences | | Initial Play | |
|---|---|---|---|---|
| | Error | Completeness | Error | Completeness |
| Naive Benchmark | 98.32 | 0% | 0.66 | 0% |
| | (4.00) | | (0.02) | |
| Economic Model | 64.92 | 91% | 0.40 | 76% |
| | (4.49) | | (0.02) | |
| Irreducible Error | 61.64 | 100% | 0.32 | 100% |
| | (3.00) | | (0.03) | |

Table I. We report the completeness of the CPT and PCHM models in their respective prediction tasks.

We find that CPT is nearly complete, achieving 91% of the feasible reduction in prediction error, while its absolute level of prediction error is 64.92. The PCHM achieves 76% of the achievable reduction, which is good, but leaves room for improvements that capture additional regularities. We note additionally that the best PCHM model on this data set is the simpler 0-parameter *Level-1* model, which predicts the action that is a best response to uniform play.

In Fudenberg and Liang [2019], we trained a bagged decision tree algorithm to predict play in the games considered in Table I. This algorithm led to a further improvement in predictive accuracy. We then examined the 14 (out of 86) games where play was predicted correctly by our algorithm, but not by level-1/PCHM. Each of these games had an action whose average payoffs closely approximated the level-1 action, but which led to lower variation in possible payoffs. Players were more likely in the data to choose this "almost" level-1 action than the actual level-1 action.

One explanation for this behavior is that players maximize a concave function over game payoffs, as if they are risk averse. This led us to add a single parameter $\alpha$

to the level-1 model, so that the prediction is the level-1 action when dollar payoffs $u$ are transformed under $f(u) = u^\alpha$. The performance of this model, called level-1($\alpha$), weakly improved upon the decision tree ensemble, which shows that atheoretical prediction rules fit by machine learning algorithms can help researchers discover interpretable and portable extensions of existing models.

## 3. RESTRICTIVENESS

The high completeness of CPT and level-1($\alpha$) suggest that these models capture many of the regularities in the data. But because each of these models has free parameters that are chosen to maximize fit, one explanation for the high completeness measures is simply that these models are flexible enough to accommodate any pattern of behavior.[1] We would thus like to distinguish high completeness because a model includes most of the functions from $X$ to $Y$ from high completeness because the model includes the "right" regularities, namely those that are observed in actual data. In Fudenberg et al. [2020], we propose an algorithmic method for quantifying the restrictiveness of a model, which allows us to separate these cases.

Our strategy is to generate random mappings $f : X \to Y$ from a set $\mathcal{F}_\mathcal{M}$ of "permissible mappings"—for example, all mappings of certainty equivalents that are consistent with the property that people prefer more money to less—and evaluate how well these mappings can be approximated using the model $\mathcal{F}_\Theta$. The more mappings from $\mathcal{F}_\mathcal{M}$ that can be approximated by a model, the less restrictive that class is. To operationalize our measure, we define restrictiveness relative to a distribution $\mu$ on $\mathcal{F}_\mathcal{M}$ chosen by the analyst, where we interpret $\mu$ as the analyst's prior over the space of mappings. (One natural option would be a uniform prior.)

Formally, for any two mappings $f$ and $f'$, define $d(f, f') = \mathbb{E}_{P_X}(l(f(x), f'(x))$ to be the (average) distance between their outcomes, where $P_X$ is the marginal distribution over the feature space. If $f'$ describes the actual relationship between the features $x$ and the outcome $y$, and the distance between $f$ and $f'$ is large, then predictions using the mapping $f$ will (in expectation) lead to large errors. Further define $d(\mathcal{F}_\Theta, f) = \inf_{f' \in \mathcal{F}_\Theta} d(f', f)$ to be the distance between $f$ and the closest mapping in $\mathcal{F}_\Theta$, so that $d(\mathcal{F}_\Theta, f)/d(f_n, f)$ is a **normalized distance** between $\mathcal{F}_\Theta$ and $f$, relative to the naive prediction rule introduced in Section 2. The models that we study nest the associated naive rule, so $d(\mathcal{F}_\Theta, f) \leq d(f_n, f)$. Thus the normalized distance lies between 0 and 1 on any prediction problem.

The restrictiveness of model $\mathcal{F}_\Theta$ is then defined to be the average normalized distance between random mappings $f$ (drawn according to distribution $\mu$ on $\mathcal{F}_\mathcal{M}$) and the model $\mathcal{F}_\Theta$.

*Definition* 3.1. The *restrictiveness* of model $\mathcal{F}_\Theta$ is $r := \mathbb{E}_\mu \left[ \frac{d(\mathcal{F}_\Theta, f)}{d(f_n, f)} \right]$.

Larger $r$ corresponds to a more restrictive model: If $r = 1$, then the model fails to improve upon the naive mapping for most maps $f$, which implies that $\mathcal{F}_\Theta$ is very restrictive. If $r = 0$, then $\mathcal{F}_\Theta$ includes all mappings from the permissible set $\mathcal{F}_\mathcal{M}$, so it is completely unrestrictive. In Figure 1, we report a histogram of normalized

[1] Although there are "representation theorems" that characterize which data are consistent with a general CPT specification, the empirical content for the 3-parameter functional form is not known, and the same is true for the PCHM.

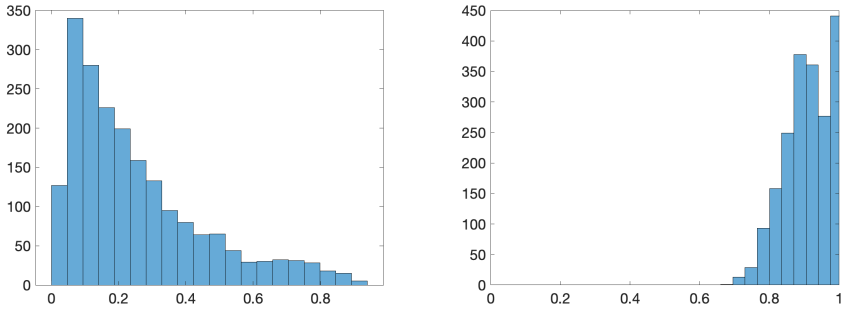


Fig. 1. *Left:* Distribution of normalized distances between CPT and random mappings; *Right:* Distribution of normalized distances between level-1($\alpha$) and random mappings.

distances between CPT and level-1($\alpha$) and 2000 random mappings of lotteries or games to certainty equivalents or modal actions respectively.

The estimated restrictiveness of CPT is 0.25, so in expectation CPT approximates a randomly selected mapping four times as well as the naive mapping does. In contrast, the restrictiveness of level-1($\alpha$) is 0.91, meaning that the level-1($\alpha$) model barely improves upon a naive mapping for approximating random mappings between games and initial play.

Since the level-1($\alpha$) model is a substantially more restrictive theory than CPT, its high completeness is suggestive that it more precisely captures the observed regularities.

## 4. ALGORITHMIC EXPERIMENTAL DESIGN

Our completeness and restrictiveness measures both depend on the underlying marginal distribution $P_X$ over the feature space. Although we expect the conditional distribution $P(y \mid s)$ to be a fixed distribution describing the dependence of the outcome on the specified set of features, the marginal distribution on $X$ is a choice variable for the experimenter. For example, we used a data set of certainty equivalents for the set of binary lotteries selected by Bruhin et al. [2010], and we used observations of initial play in $3 \times 3$ matrix games that had been chosen by different teams of authors with different purposes in mind. It isn't feasible, however, to run experiments on all lotteries or $3 \times 3$ games. The idea of *algorithmic experimental design* is to use machine learning to determine which test cases in $X$ would be most informative.

In Fudenberg and Liang [2019], we used this approach to select which $3 \times 3$ games to include in a new experiment. Our goal was to identify games where behavior was likely to depart from the level-1($\alpha$) model, as this data could then allow us to discover further regularities in play. We trained a machine learning algorithm to predict the frequency of the level-1($\alpha$) action, and then selected games that achieved low predicted frequencies according to this algorithm. This approach is related in spirit to adversarial machine learning [Huang et al. 2011] and generative adversarial networks [Goodfellow et al. 2014] in that we are generating instances to trick the level-1($\alpha$) model, although our goal is to design new instances for *data*

*collection* instead of refining predictions for a given data set.

We experimentally elicited play on these "algorithmically-generated" games on the platform Mechanical Turk, and found that the frequency of level-1($\alpha$) play is indeed low in these games. In keeping with our desire for interpretable conclusions, we did not simply look for the best black-box algorithm on our new data set. Instead, we developed a hybrid approach: We identified two models, each of which fit some of the data reasonably well, and trained a decision tree to predict which model would perform better on which games. This hybrid model outperformed its two constituent models, and studying the optimal assignment of games to models shed light on when the level-1($\alpha$) model is outperformed by an equally simple alternative model.

## 5. CONCLUSION

As we have shown, machine learning and associated algorithmic techniques can aid in the improvement of economic theories. When theories are incomplete, machine learning can help researchers identify regularities that are not captured by existing models and then develop new theories that predict better. Conversely, when a theory is highly complete, algorithmic techniques can show whether this is simply due to the theory's ability to fit any possible data, or whether the good fit results from the theory describing behaviors in the real world. Finally, machine learning can be used to guide researchers in choosing which experiments to run.

## REFERENCES


Bruhin, A., Fehr-Duda, H., and Epper, T. 2010. Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica 78,* 4, 1375–1412.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics 119,* 3, 861–898.

Fudenberg, D., Gao, W., and Liang, A. 2020. Quantifying the restrictiveness of theories. Working Paper; available at `http://economics.mit.edu/faculty/drewf/working`.

Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. 2019. Measuring the completeness of theories. Working Paper; available at `http://economics.mit.edu/faculty/drewf/working`.

Fudenberg, D. and Liang, A. 2019. Predicting and understanding initial play. *American Economic Review 109,* 12, 4112–41.

Goldstein, W. M. and Einhorn, H. J. 1987. Expression theory and the preference reversal phenomena. *Psychological review 94,* 2, 236.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. Generative adversarial networks. *Advances in neural information processing systems*, 2672–2680.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B., and Tygar, J. 2011. Adversarial machine learning. *Proceedings of 4th ACM Workshop on Artificial Intelligence and Security*, 43–58.

Lattimore, P. K., Baker, J. R., and Witte, A. D. 1992. The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization 17,* 3, 315–436.

Nagel, R. 1995. Unraveling in guessing games: An experimental study. *American Economic Review 85,* 5, 1313–1326.

Stahl, D. O. and Wilson, P. W. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior and Organization 25,* 3, 309–327.

Stahl, D. O. and Wilson, P. W. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior 10,* 1, 218–254.

Wright, J. R. and Leyton-Brown, K. 2014. Level-0 meta-models for predicting human behavior in games. *Proceedings of the fifteenth ACM conference on Economics and computation*, 857–874.