How Flexible is that Functional Form?
Quantifying the Restrictiveness of Theories*

Drew Fudenberg†  Wayne Gao ‡  Annie Liang§

July 19, 2023

Abstract

We propose a restrictiveness measure for economic models based on how well they fit synthetic data from a pre-defined class. This measure, together with a measure for how well the model fits real data, outlines a Pareto frontier, where models that rule out more regularities, yet capture the regularities that are present in real data, are preferred. To illustrate our approach, we evaluate the restrictiveness of popular models in two laboratory settings—certainty equivalents and initial play—and in one field setting—takeup of microfinance in Indian villages. The restrictiveness measure reveals new insights about each of the models, including that some economic models with only a few parameters are very flexible.

*We thank Nikhil Agarwal, Victor Aguiar, Abhijit Banerjee, Tilman Börgers, Vincent Crawford, Glenn Ellison, Benjamin Enke, Ben Golub, Taisuke Imai, Shaowei Ke, David Laibson, Rosa Matzkin, John Quah, Kareen Rozen, Jesse Shapiro, Ludvig Sinander, Charles Sprenger, Dmitry Taubinsky, and Emanuel Vespa for helpful comments, and NSF grants SES 185162 and 1951056 for financial support. We thank Kyohei Okumura for excellent research assistance.
†Department of Economics, MIT; drewf@mit.edu
‡Department of Economics, University of Pennsylvania; waynegao@upenn.edu
§Department of Economics, Northwestern University; annie.liang@northwestern.edu.
1 Introduction

If a parametric model fits the data well, is it because the model captures structure specific to the observed data, or because the model is so flexible that it would fit almost all conceivable data? This paper provides a quantitative measure of restrictiveness that can distinguish between these two explanations, and is easy to compute in a variety of applications.

Our approach for evaluating the restrictiveness of a model is to generate synthetic data sets, and evaluate how well the model fits this synthetic data. Some models have known properties, for example Cumulative Prospect Theory requires that certainty equivalents for lotteries respect first-order stochastic dominance. For these models, the relevant question may not be whether the model is restrictive at all, but instead how much content it has beyond these known constraints. We define the eligible data to be those data sets that satisfy specified background constraints, and measure a model’s restrictiveness by its (normalized) average error across the eligible data.

We complement the evaluation of restrictiveness, which is based solely on synthetic data, with an evaluation of the model’s performance on actual data, using the completeness measure proposed in Fudenberg et al. (2022). Restrictiveness and completeness provide complementary perspectives, and define a Pareto frontier where models that rule out more regularities, yet capture the regularities that are present in real data, are preferred.1

Section 4 provides axioms for our restrictiveness measure to clarify its theoretical properties. The main axioms require that the measure is homogeneous in the unit scale used to quantify model error, and that the measure has a linearity property as the background constraints are varied. An additional “symmetry” axiom requires that the model’s ability to approximate different synthetic data sets has the same effect on the restrictiveness measure. Dropping this axiom returns a broader class of restrictiveness measures, where instead of

1These are not the only considerations that matter for evaluating models, and we do not speak to other important concerns such as parameter estimation and causal inference. Nevertheless, these two measures may be relevant to those problems as well: If a model can fit almost any data set, then its good fit to a specific real data set does not necessarily mean that the model is the “right” model.
averaging across synthetic data sets, the data sets are weighted by an analyst’s prior. We
develop estimators for both the restrictiveness and completeness measures in Section 5, and
establish their asymptotic properties so that users can compute confidence intervals.

A key feature of our restrictiveness measure is that is computable without the guidance
of theoretical results about the model’s implications or empirical content. This differentiates
restrictiveness from measures such as the model’s VC dimension, or its hit-rate and accuracy-
rate as defined in Selten (1991). This differentiates restrictiveness from measures such as the model’s VC dimension, or its hit-rate and accuracy-rate as defined in Selten (1991). (Section 3.4 reviews the related literature and relates it to our work.) The measure’s tractability makes it easy to apply to a variety of contexts, as we demonstrate by applying it to models from three economic domains: (1) predicting certainty equivalents for binary lotteries (where we evaluate Cumulative Prospect Theory and Disappointment Aversion); (2) predicting initial play in matrix games (where we evaluate the Poisson Cognitive Hierarchy Model (PCHM), Logit PCHM, and Logit Level-1); and (3) predicting takeup of microfinance in Indian villages (where we evaluate linear regression models based on economically-motivated regressors, and a structural model of diffusion).

The first two settings use data from the lab, our third application uses field data. In each of
these domains, these measures reveal new insights about the models we examine, which we
now summarize:

Application 1: Certainty Equivalents. We evaluate two models on a set of binary
lotteries from Bruhin et al. (2010): a popular three-parameter specification of Cumulative
Prospect Theory (Tversky and Kahneman, 1992), henceforth CPT, and a two-parameter
specification of Disappointment Aversion (Gul, 1991), henceforth DA. We find that CPT
performs strikingly well on the Bruhin et al. (2010) data, achieving a completeness of 95%,
while DA’s completeness is only 27%.

One explanation for this finding is that CPT is a much better model of risk preferences

---

2There are representation theorems for many non-parametric theories of individual choice, and some
analytic results for the sets of equilibria in games, but we are unaware of representation theorems for most
functional forms that are commonly used in applied work.

3In addition to these applications, Schwaninger (2022) uses our restrictiveness measure to evaluate models
of bargaining with inequity aversion, Ellis et al. (2022) uses it to evaluate models of consumer demand from
budget sets, and Ba et al. (2023) uses it to evaluate models of reaction to information.
than DA. Another possibility is that CPT is simply more flexible. We thus evaluate the restrictiveness of the two models, where our background constraints are that the synthetic average certainty equivalents must lie within the range of the lotteries’ possible payoffs, and must respect first-order stochastic dominance (FOSD). We find that CPT is indeed substantially less restrictive than DA: CPT performs better than DA not only on the real data set but also on the other eligible data sets. This tells us that FOSD constitutes a large part of the empirical content of CPT on the domain of binary lotteries, while DA imposes substantial additional restrictions.\footnote{DA’s low completeness suggests that these restrictions are not supported by the experimental data.}

Besides comparing distinct models such as CPT and DA, restrictiveness and completeness can be compared across nested models to reveal the role played by specific parameters. Adding a parameter always at least weakly increases completeness and decreases restrictiveness, but some parameters achieve greater improvements in completeness for the same decrease in restrictiveness. We find that several parameters lead to large drops in restrictiveness in return for only marginal improvements in completeness, suggesting that these parameters may add flexibility in the wrong directions. The CPT parameter that governs the curvature of the probability weighting function, however, achieves a large improvement in completeness compared to the flexibility it adds, so this parameter seems to capture an important part of risk preferences. Indeed, it is the curvature of the probability weighting function that has played a key role in many of the applications of CPT to financial data (e.g., Barberis and Huang (2008) and Green and Hwang (2012)).

**Application 2: Initial Play in Games.** Next, we evaluate three models on a set of 3 × 3 matrix games from Fudenberg and Liang (2019): the Poisson Cognitive Hierarchy Model, or \textit{PCHM} (Camerer et al., 2004); \textit{Logit PCHM} (Wright and Leyton-Brown, 2014), which allows for logistic best replies in the PCHM; and \textit{Logit Level-1}, which models the distribution of play as a logistic best reply to the uniform distribution. We impose the background constraint that strictly dominant actions are played at least as often as if by
chance (i.e. with probability at least 1/3) and that strictly dominated actions are played with probability no more than 1/3. We find that all three models are highly restrictive relative to these constraints, which shows that the constraints on the frequency of strictly dominated and strictly dominant strategies are a very small part of their empirical content. The restrictiveness of Logit PCHM and Logit Level-1 is nearly identical, although Logit PCHM has two parameters while Logit Level-1 has one.

**Application 3: Diffusion on a Social Network.** Finally, we consider the prediction of microfinance takeup rates in the set of Indian villages studied by Banerjee et al. (2013, 2019), and compare the performance of OLS regression on various economically-motivated regressors with that of an economically-motivated partially linear model built upon “network gossip centrality.” Here we find that the partially linear model is dominated by a simple OLS model based on the average eigenvector centrality of leaders: the latter has higher restrictiveness and higher completeness.

Besides these specific findings about each of these economic domains, our analyses make the high-level point that it is not sufficient to count parameters to understand a model’s restrictiveness. Even with just 3 parameters, CPT is not very restrictive on the domain of binary lotteries, and models with different numbers of parameters (such as Logit PCHM and Logit Level-1) turn out to be similarly restrictive. These comparisons are not obvious from the functional forms, but are easy to discover with our restrictiveness measure.

### 2 Example

Before formally defining our measure, we use a simple example to illustrate it. Suppose there is a binary covariate \( x \in \{x_0, x_1\} \) and an outcome variable \( y \in [0, 1] \). A **data set** is an observed outcome for each covariate value, i.e., a point in \( \mathbb{R}^2 \), and the **eligible data** \( \mathcal{F} \) is collection of possible data sets, i.e. a subset of \( \mathbb{R}^2 \). A **model** is also a subset of \( \mathbb{R}^2 \). The model explains a data set exactly if the data set is an element of the model.
Figure 1 considers eligible data $[0, 1]^2$ and depicts three models. Model A includes all of $[0, 1]^2$, and thus can exactly explain any (eligible) data set. Model B includes all data sets $(y_0, y_1)$ satisfying $y_1 > y_0$, and so can only explain data sets where the outcome is higher at covariate $x_1$ than at $x_0$. Model C discretizes the data into a grid and includes every other element of the grid.

One way of evaluating the restrictiveness of these models is the fraction of eligible data sets that they can fit exactly (Selten, 1991). But evaluating restrictiveness in this way obscures important differences between models such as B and C. Both exactly explain 50% of the data yet Model B appears to impose a more substantive restriction. Additionally, while it is easy to determine the fraction of possible data that are exactly consistent with each of these stylized models, this fraction can be difficult to determine when the data is high-dimensional and the models are complex.

Our restrictiveness measure takes as given a measure of how well a model approximates the data. We then uniformly sample over all eligible data and evaluate the model’s average approximation error to the realized datasets. To benchmark this error’s size, we compare it to the average approximation error of some constant model. In this example, we evaluate the minimal Euclidean distance between the model and each of the realized points (as depicted in Figure 1), and benchmark against the constant model $\{(1/2, 1/2)\}$. The restrictiveness
of Model B is numerically estimated to be approximately 0.30, while the restrictiveness of Model C is approximately 0.02. So Model B is substantially more restrictive by our measure. In general, because our approach relies on estimating approximation errors rather than determining exact fits, it is computationally straightforward to estimate in many applications even when we have very little analytical guidance about the model’s predictions.

3 Our Methodology

Section 3.2.1 formally defines our measure of restrictiveness. Section 3.2.2 reviews the measure of completeness from Fudenberg et al. (2022). Section 3.2.3 combines these concepts with the idea of a Pareto frontier of models that are undominated in completeness and restrictiveness. Section 3.3 further discusses the interpretation of our restrictiveness measure. Section 3.4 describes the relationship to the literature.

3.1 Setup

Our starting point is a data set of observations of \((X, Y)\), where \(X\) is a covariate vector and \(Y \in \mathcal{Y}\) is an outcome, with \(\mathcal{Y}\) being a subset of a finite-dimensional Euclidean space. We use \(\mathcal{X}\) to denote the set of covariate vectors, and \(P_X\) to denote the marginal distribution of \(X\). We assume that \(\mathcal{X}\) is finite, and \(P_X\) is chosen or known to the researcher.\(^5\) A prediction rule is a function \(f: \mathcal{X} \to \mathcal{Y}\). We denote the set of all such functions by \(\mathcal{F} \equiv \mathcal{Y}^{\left|\mathcal{X}\right|}\), and endow it with the norm topology.

**Example 1** (Predicting an Average Outcome). In our application to the prediction of certainty equivalents (Section 6), the covariate vectors are 25 binary lotteries, each described by two prizes and their probabilities, and the outcome space is the observed average (over subjects) certainty equivalent for each lottery in this data set. A prediction rule is any function from the 25 lotteries to average certainty equivalents.

\(^5\)In laboratory experiments the set of features and their relative frequencies are chosen by the experimenter while in field experiments these are chosen by Nature, but in either case we treat them as known.
Example 2 (Predicting a Distribution). In our application to initial play in 3x3 games (Section 7), the features are the 18 elements of the payoff matrix, and the outcomes are distributions over the row player’s actions. A prediction rule is a map from payoff matrices to probability distributions over row player actions.

3.2 Measures

3.2.1 Restrictiveness

We take as a primitive a *discrepancy* function \( d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+ \) where \( d(f, f') \) measures how different the two prediction rules \( f \) and \( f' \) are. For example, if \( Y \) is a vector in \( \mathbb{R}^n \), a natural choice for \( d \) is the expected mean-squared distance between the predictions (with respect to \( P_X \)), and if \( Y \) is a distribution a natural choice for \( d \) is the expected KL-divergence (again with respect to \( P_X \)). We allow for functions \( d \) that are not distances (such as KL-divergence), but require that \( d(f, f') = 0 \) if and only if \( f = f' \).

We will evaluate the restrictiveness of a parametric model \( \mathcal{F}_\Theta := \{ f_\theta \}_{\theta \in \Theta} \subseteq \mathcal{F} \), where the prediction rules \( f_\theta \) depend continuously on a parameter \( \theta \) from a compact set \( \Theta \).\(^6\) Restrictiveness is defined relative to a compact set of “eligible” rules \( \mathcal{F} \subseteq \mathcal{F} \) that reflect any constraints the model is known to have. For example, if a model is known to imply that choices respect first-order stochastic dominance, we can define \( \mathcal{F} \) to be all functions with this property, and measure the model’s additional restrictiveness beyond this. In general, the eligible set \( \mathcal{F} \) consists of all prediction rules that satisfy user-specified background constraints, where the special case of \( \mathcal{F} = \mathcal{F}_\Theta \) corresponds to the question of whether \( \mathcal{F}_\Theta \) imposes any restrictions at all.

We define the restrictiveness of a model to be its expected discrepancy to a prediction rule \( f \) drawn uniformly at random from the eligible set, normalized with respect to the expected discrepancy of a baseline prediction rule \( f_{\text{base}} \). The baseline prediction rule is chosen to

\(^6\)Because \( \mathcal{X} \) is assumed to be finite, \( \Theta \) can viewed as a subset of a finite-dimensional Euclidean space without loss of generality.
suit the setting, and we interpret its performance as a lower bound that any sensible model should outperform.\footnote{For example, in our application to predicting initial play in games, we define the baseline prediction rule to be a uniform distribution over actions. Note that while the choice of baseline affects the value of restrictiveness, it does not affect the comparative restrictiveness of two models on the same domain.}

**Definition 1.** The restrictiveness of model $\mathcal{F}_\Theta$ with respect to eligible set $\mathcal{F}$ is

\[
    r(\mathcal{F}_\Theta, \mathcal{F}) = \frac{\mathbb{E}_{\lambda_\mathcal{F}}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}_{\lambda_\mathcal{F}}[d(f_{\text{base}}, f)]}
\]

where $\lambda_\mathcal{F}$ denotes the uniform distribution on $\mathcal{F}$, and $d(\mathcal{F}_\Theta, f) := \inf_{f_\theta \in \mathcal{F}_\Theta} d(f_\theta, f)$.\footnote{Since $\mathcal{F}$ is a subset of bounded finite-dimensional Euclidean space, the uniform distribution on $\mathcal{F}$ is well-defined. Section 4 discusses a generalization to other distributions.}

Normalizing with respect to a baseline has several advantages: First, it makes our measure invariant to affine rescalings of the units of discrepancy. Second, whenever $f_{\text{base}}$ is chosen from $\mathcal{F}_\Theta$, restrictiveness ranges from 0 to 1. A model with $r = 0$ is completely unrestricted, while a model with $r = 1$ fits synthetic data no better than the baseline prediction rule does. If a model performs well on real data and is also highly restrictive, then its good performance occurs not simply because the model can fit any data, but because it precisely identifies regularities in real behavior.

The ratio in (1) is well-defined as long as the denominator exceed zero, so we will impose this an assumption going forward:

**Assumption 1.** $\mathbb{E}_{\lambda_\mathcal{F}}[d(f_{\text{base}}, f)] > 0$.

Section 4 provides axioms for the restrictiveness measure, which help to clarify the measure’s theoretical properties.
3.2.2 Completeness

While restrictive models are desirable holding all else equal, a restrictive model is not useful if it poorly fits real data. To evaluate model fit to real data, we use the completeness measure introduced in Fudenberg et al. (2022). This takes as a primitive a loss function \( l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \), which is assumed to be continuous. Let \( P_{Y|X} \) denote the distribution of \( Y \) given \( X \), and \( P := (P_X, P_{Y|X}) \) denote the joint distribution of \( X \) and \( Y \). The prediction rule that minimizes expected loss on the real data is given by

\[
f^* \in \arg \min_{f \in \mathcal{F}} e_P(f)
\]

where

\[
e_P(f) := \mathbb{E}_P[l(f(X), Y)] \quad \forall f \in \mathcal{F}.
\]

For example, if \( \mathcal{X} \) is a set of lotteries, \( \mathcal{Y} \) are subjects’ reported certainty equivalents for each lottery, and \( l \) is squared error, then \( f^* \) takes each lottery into its average certainty equivalent across subjects. If \( \mathcal{X} \) is a set of payoff matrices, \( \mathcal{Y} \) is the set of distributions over actions, and \( l(Y, Y') \) is Kullback-Leibler divergence from \( Y' \) to \( Y \), then \( f^* \) maps each game to the corresponding distribution over actions.

**Definition 2** (Fudenberg et al., 2022). The completeness of model \( \mathcal{F}_\Theta \) is defined by

\[
\kappa(\mathcal{F}_\Theta) := \frac{e_P(f_{\text{base}}) - \inf_{f \in \mathcal{F}_\Theta} e_P(f_\Theta)}{e_P(f_{\text{base}}) - e_P(f^*)}.
\]

By construction, \( \kappa \) lies within the unit interval. A model with \( \kappa = 1 \) matches the true \( f^* \) exactly, while a model with \( \kappa = 0 \) is no better at matching \( f^* \) than the baseline prediction rule \( f_{\text{base}} \).

We report both restrictiveness \( r \) and completeness \( \kappa \) for each application that we consider.

---

\( ^{10} \)In the special case where discrepancy is the expected mean-squared distance \( d(f, f') = \mathbb{E}_{P_X}[(f(X) - f'(X))^2] \) and the baseline prediction rule is constant at the expectation of \( Y \), \( f_{\text{base}} = \mathbb{E}_P[Y] \), completeness specializes to the familiar (population) definition of \( R^2 \), but completeness is applicable more generally.
Completeness is defined using the loss function $l$, while restrictiveness is defined using the discrepancy function $d$. When the discrepancy function $d$ and the loss function $l$ are “paired” in the sense of Online Appendix E,\textsuperscript{11} then $\kappa(\mathcal{F}_\Theta) = 1 - r(\mathcal{F}_\Theta, \{f^*\})$, so that completeness is the complement of the restrictiveness of model $\mathcal{F}_\Theta$ with respect to the true model $f^*$. Our first and third application use mean-squared error as the loss function and expected squared distance as the discrepancy function; our second application uses negative log-likelihood as the loss function and expected KL divergence as the discrepancy function. Both are examples of paired functions.

### 3.2.3 A “Pareto Frontier”

Our restrictiveness and completeness measures generate a “Pareto frontier” consisting of models that are undominated in the sense that none of the other models considered are simultaneously more restrictive and more complete. Although this is a very partial order, it has bite in our Application 6 (see Figure 2), as well as in the work of Ellis et al. (2022).

Unlike in typical economic problems, the Pareto frontier here need not be concave, so the preferred model may not maximize a weighted sum of the two scores. For example, the frontier might consist of 3 points with scores $(3/4,1/4)$, $(1/3,1/3)$, and $(1/4,3/4)$, and the analyst might prefer the model with scores $1/3$ each. Of course, given the estimated parameter values of two models on the actual and hypothetical data sets, one could make predictions by taking pointwise combinations of the two model’s predictions, which would mechanically lead to a weakly concave frontier of undominated models, but it seems hard to interpret this exercise.

While it is natural to prefer undominated models to dominated ones, it is less obvious how to aggregate the two measures to pick a preferred model, as the tradeoff between the measures is context-specific and also a matter of taste. Nevertheless, when two models have completeness-restrictiveness values that cannot be Pareto-ranked, one can consider the size

\textsuperscript{11}Loosely speaking, being paired means that $d(f, f^*)$ is the difference between the error of $f$ and the error of the best mapping $f^*$. 

10
of the improvement in completeness relative to the size of the reduction in restrictiveness. In Section 6.4 we show that adding an “elevation” parameter to a Cumulative Prospect Theory specification leads to a large drop in restrictiveness in return for only a small gain in completeness, while the parameter that governs the curvature of the probability weighting function leads to a sizeable improvement in completeness with only a small reduction in restrictiveness. We take this to mean that the curvature parameter plays a more important role in capturing risk preferences.\(^{12}\)

### 3.3 Discussion

**Context dependence.** Restrictiveness is context-specific, in the sense that it depends on the set of feature vectors \(X\) and the outcome to be predicted. For example, we show that the restrictiveness of Cumulative Prospect Theory depends on the support size of the lotteries that are considered. Evaluating the restrictiveness of a model across contexts can reveal that it is very restrictive for one kind of prediction problem but unrestrictive for others. An interesting direction for followup work would be to develop a measure of restrictiveness that takes into account how restrictive a model is across different contexts. For example, we might consider one model to be “generally more restrictive” than a second model if the distribution of restrictiveness values for the first model first-order stochastically dominates the distribution for the latter, as we find in Section 6.5.

**Choosing the eligible set.** The restrictiveness of a model is measured with respect to a specific eligible set, which is chosen based on what is known about the model. In Application 3, we investigate the restrictiveness of a structural model of network diffusion for predicting takeup of microfinance. Since there is relatively little known about the empirical content of this model, we define the eligible set to include all possible takeup rates, and study whether the model placed any restrictions at all. In contrast, the model of interest in Application 1, \(^{12}\)Ba, Bohren, and Imas (2023) conduct a similar exercise to compare two models which are not Pareto-ranked.
Cumulative Prospect Theory, implies that any lottery that first order stochastically dominates another must have a higher certainty equivalent. So we place this restriction on the eligible set, and see how much additional restrictiveness the model imposes.

In general, there is not a single correct choice of eligible set. While we focus on comparing the restrictiveness of models with respect to a given eligible set, an interesting complementary exercise is to fix a model and compare its restrictiveness relative to different eligible sets, as in Sections 6.5 and 7.3.

Why the uniform distribution? Section 4, which develops and axiomatizes a broader class of restrictiveness measures, provides an axiom that pins down the uniform distribution. Besides this axiom, there are many reasons to prefer the uniform distribution. First, once the eligible set is specified, the uniform distribution on this set is pinned down (under our assumptions that $\mathcal{X}$ is finite and $\mathcal{Y}$ is a subset of finite-dimensional Euclidean space). This reduces the number of primitives to be chosen, and helps prevent cherry-picking with respect to the distribution on $\mathcal{F}$. Second, the uniform distribution is computationally easy to implement, even for eligible sets $\mathcal{F}$ with potentially complicated structures.\textsuperscript{13} Finally, our use of the uniform distribution follows up on Becker (1962)’s proposal of the uniform distribution over budget-exhausting bundles as a model of irrational consumer behavior, and parallels Selten (1991)’s use of area (see Section 3.4).

Why are more restrictive models better? Our paper takes the perspective that restrictiveness is inherently desirable: if two models have the same level of predictive accuracy, we should prefer the one that imposes more restrictions to the more flexible alternative. A potential reason for this preference is that models are often meant to capture behavior in related but not-identical domains. Given enough data, models that are very unrestrictive

\textsuperscript{13}For example, in our application to prediction of certainty equivalents, we build monotonicity with respect to FOSD into our definition of $\mathcal{F}$, and it is straightforward to sample uniformly from $\mathcal{F}$ by first sampling from a larger space without the monotonicity constraints, and then only keeping the draws that satisfy the monotonicity constraints. In contrast, non-uniform weightings over $\mathcal{F}$ require additional specification of how exactly $\mathcal{F}$ is parametrized, making the dependence of restrictiveness on $\mathcal{F}$ less transparent.
will fit any specific data set well, but may do so by learning idiosyncratic details of those datasets that do not in fact transfer across settings. In contrast, if a highly specific and structured model happens to fit a data set well, this may generate more confidence that the model’s structure extends to other settings.\textsuperscript{14}

3.4 Relationship to the Literature

Our restrictiveness measure generalizes the notion of “observational restrictiveness" introduced in Koopmans and Reiersol (1950), where a model is observationally restrictive if the distributions permitted by the model are a proper subset of the distributions that would otherwise be possible.\textsuperscript{15} A model that is not observationally restrictive can perfectly match all data and so has $r = 0$. Our restrictiveness measure allows us to quantify just how restrictive a model is.

Section 2 already discussed Selten (1991)’s measure of flexibility, and showed how its use of exact instead of approximate fit can lead to very different conclusions than ours. The Selten measure has been applied by Beatty and Crawford (2011), Hey (1998), and Harless and Camerer (1994), and Blow et al. (2021) among others, to understand the restrictiveness of nonparametric economic models. It is typically difficult to determine whether a parametric model can exactly fit a given data set without the guidance of prior analytical results, while our measure is easy to compute in a variety of applications.\textsuperscript{16}

In considering approximate rather than exact fit, our approach is related to papers that measure the distribution of the Afriat index (Choi et al., 2007; Polisson et al., 2020).\textsuperscript{17} These

\textsuperscript{14}Andrews et al. (2022) compare the transfer performance of highly flexible black box models with less flexible economic models in a setting similar to our Application 1, and find that the black box models transfer more poorly.

\textsuperscript{15}As Koopmans and Reiersol (1950) points out, a special case of an observationally restrictive specification is an overidentifying restriction. See e.g. Sargan (1958), Hausman (1978), Hansen (1982), and Chen and Santos (2018) for econometric tests of overidentification.

\textsuperscript{16}Beatty and Crawford (2011) analytically derives the set of budget shares that are consistent with GARP, and Harless and Camerer (1994) uses results about generalized expected utility theories to determine whether choices between specially chosen pairs of lotteries (for example, lotteries sharing a common ratio of outcome probabilities) are consistent with those theories. But we do not know how to analytically determine the predictions that are consistent with PCHM or the structural model of microfinance takeup in Application 3.

\textsuperscript{17}Choi et al. (2007) and Polisson et al. (2020) relax the implications of expected utility maximization using
approaches are motivated by the testing of rationality of choices; our aim here is to show that similar techniques can be applied to a substantially broader class of models. Beatty and Crawford (2011) propose an alternative “smoothed out” version of Selten (1991)’s measure for the revealed preference setting that resembles restrictiveness, except that it does not allow for restrictions on the eligible data and normalizes by reference to a worst case.\footnote{Afriat’s “efficiency index” as an analog of our loss function. They compare the distribution of the efficiency indices of the actual subjects with its counterpart in randomly generated data.}

Our use of synthetic data to evaluate restrictiveness is similar to the use of simulated data to evaluate the power of a hypothesis test, as in Bronars (1987) and Andreoni et al. (2013). Their power measures are based on particular specifications of the alternative hypothesis, while we focus on an aggregate measure over a class of “alternative hypotheses.” Moreover, because our objective is to measure the content of a model’s restrictions and not hypothesis testing, we use approximate rather than exact fit.

There is a large literature in statistics and econometrics on model selection, which dates back to Cox (1961, 1962). Our restrictiveness measure adds to this literature but has several key differences. First, classic measures such as AIC and BIC are based on observed data, while our restrictiveness measure is not. This difference reflects a difference in objectives: A primary goal of model selection is to avoid overfitting a complex model to a finite (and small) quantity of data. Thus, these metrics typically trade off some notion of completeness against some notion of restrictiveness: For example, the AIC combines the log-likelihood, which is about fitness to real data (corresponding to “completeness”) and the number of parameters, which is about the flexibility of the model without reference to real data (corresponding to “restrictiveness”) in an additive way. In contrast, we view model restrictiveness (and relatedly, model parsimony) as intrinsically valuable even with an infinite data set. Our measure of restrictiveness thus does not depend on sample sizes or fit to real data.

Finally, our measure is also related to complexity measures from computer science and

\footnote{Another approach for model selection that does not require exact fit is de Clippel and Rozen (2022)’s suggestion to select models by comparing the ratio of the likelihood of observing the real data under the specified model to the likelihood under a uniform distribution over all possible models.}
statistics, including VC dimension, Rademacher complexity, and various notions of entropies. In most economic applications, it is difficult to determine the VC dimension of the relevant models.\textsuperscript{19} Rademacher complexity can be numerically computed, but differs from our measure in that it is indexed not only to the model but also to a sample size.\textsuperscript{20} Our restrictiveness measure is instead developed for a potentially infinite data set: For example, in the analysis of CPT, we do not generate synthetic finite data sets of lotteries and certainty equivalents, but rather directly generate maps from lotteries to conditional expectations.\textsuperscript{21} Various notions of metric entropy (based on the concept of covering numbers) and bracketing entropy (based on the concept of bracketing numbers) have also been used to measure the complexity of a function class in statistics, especially in the theory of empirical processes (Van Der Vaart and Wellner, 1996). Usually, analysts are only concerned with whether entropy integrals are finite and if so their growth rates, which are used to derive bounds on the convergence rates of various statistical quantities. Our measure is easier to compute, and can accommodate background constraints.

4 Axiomatic Foundation for Restrictiveness

This section provides an axiomatization for the un-normalized version of the restrictiveness measure (i.e., the numerator of (1)), which we call \textit{approximation error}. Readers primarily interested in applications of the measure can skip ahead to the next section.

We endow the set $\mathcal{F}$ with the Lebesgue $\sigma$-algebra and a $\sigma$-finite measure $\mu$, which can be interpreted as the analyst’s prior. An approximation error $e$ takes as input the model $\mathcal{F}_\Theta \subseteq \mathcal{F}$, a Lebesgue-measurable compact set of eligible prediction rules $\mathcal{F} \subseteq \mathcal{F}$, and a discrepancy function $d$. The quantity $e(\mathcal{F}_\Theta, \mathcal{F}, d)$ is interpreted as the approximation error

\textsuperscript{19}Basu and Echenique (2020) characterize the VC dimension of non-parametric models of decision-making under uncertainty, but it is not clear how to extend these results to parametric models.

\textsuperscript{20}This is again because of the difference in intended uses of the metrics: Rademacher complexity is typically used to bound generalization error, with a view towards avoiding overfitting.

\textsuperscript{21}We could loosely interpret our restrictiveness measure as analogous to a limiting case of Rademacher complexity for large samples, where we use the discrepancy function $d$, rather than correlation, to measure the model’s ability to fit the synthetic data.
of the model $\mathcal{F}_\theta$ to the eligible set $\mathcal{F}$, where the quality of the approximation is measured using $d$. We would like for this approximation error function to satisfy the following axioms. First, approximation error should always be nonnegative.

Axiom 1 (Nonnegativity). For every model $\mathcal{F}_\theta$, eligible set $\mathcal{F}$, and discrepancy $d$, $e(\mathcal{F}_\theta, \mathcal{F}, d) \geq 0$.

Second, if one model is better able to approximate every eligible prediction rule than another, the first model has lower approximation error.

Axiom 2 (Monotonicity). Fix any set of eligible mappings $\mathcal{F}$. If the sets $\mathcal{F}_\theta_1$ and $\mathcal{F}_\theta_2$ satisfy $d(\mathcal{F}_\theta_1, f) \geq d(\mathcal{F}_\theta_2, f)$ for all $f \in \mathcal{F}$, then $e(\mathcal{F}_\theta_1, \mathcal{F}, d) \geq e(\mathcal{F}_\theta_2, \mathcal{F}, d)$.

Third, any linear rescaling of the units of $d$ is inherited by the approximation error, and a linear rescaling of the discrepancy between a model $\mathcal{F}_\theta$ to each prediction rule $f$ leads to the same value of approximation error as rescaling the units of the discrepancy $d$.

Axiom 3 (Homogeneity). (a) Fix any model $\mathcal{F}_\theta$, set of eligible prediction rules $\mathcal{F}$, and discrepancy $d$. Then $e(\mathcal{F}_\theta, \mathcal{F}, \alpha \cdot d) = \alpha \cdot e(\mathcal{F}_\theta, \mathcal{F}, d)$ for every $\alpha \in \mathbb{R}_+$

(b) Fix any set of eligible prediction rules $\mathcal{F}$ and discrepancy $d$. If $\mathcal{F}_\theta_1$ and $\mathcal{F}_\theta_2$ satisfy $d(\mathcal{F}_1, f) = \alpha \cdot d(\mathcal{F}_2, f)$ for all $f \in \mathcal{F}$, then $e(\mathcal{F}_\theta_1, \mathcal{F}, d) = e(\mathcal{F}_\theta_2, \mathcal{F}, \alpha \cdot d)$.

Fourth, consider constraining the set of eligible prediction rules $\mathcal{F}$ to a subset $\mathcal{F}_1$ or its complement $\mathcal{F}_2$. The ex post approximation errors of a model $\mathcal{F}_\theta$ with respect to either of these new eligible sets is, respectively, $e(\mathcal{F}_\theta, \mathcal{F}_1, d)$ or $e(\mathcal{F}_\theta, \mathcal{F}_2, d)$. The subsequent axiom says that the ex ante approximation error $e(\mathcal{F}_\theta, \mathcal{F}, d)$ is a convex combination of the ex post approximation errors, where each ex post subset contributes to the ex ante approximation error in proportion to its measure.

Axiom 4 (Linearity). For any sequence of disjoint measurable sets $\mathcal{F}_\theta_1, \mathcal{F}_\theta_2, \ldots$ whose union $\mathcal{F}_\theta \equiv \cup_{i=1}^{\infty} \mathcal{F}_\theta_i$ has strictly positive measure,

$$e(\mathcal{F}_\theta, \mathcal{F}, d) = \sum_{i=1}^{\infty} \frac{\mu(\mathcal{F}_\theta_i)}{\mu(\mathcal{F})} \cdot e(\mathcal{F}_\theta_i, \mathcal{F}, d) \; \forall \mathcal{F}, d.$$
Finally, permuting the various discrepancies between the model and the eligible prediction rules $f$ does not affect the overall approximation error. This reflects a "principle of indifference" over the eligible prediction rules.

**Axiom 5** (Symmetry). Fix any eligible set $\mathcal{F}$ and any bijection $\tau$ from $\mathcal{F}$ to itself. Consider two sets $\mathcal{F}_{\Theta_1}$ and $\mathcal{F}_{\Theta_2}$ where $d(\mathcal{F}_{\Theta_1}, f) = d(\mathcal{F}_{\Theta_2}, \tau(f))$ for all $f \in \mathcal{F}$. Then $e(\mathcal{F}_{\Theta_1}, \mathcal{F}, d) = e(\mathcal{F}_{\Theta_2}, \mathcal{F}, d)$.

**Proposition 1.** An approximation error $e$ satisfies Axioms 1-4 if and only if there is a function $c: \overline{\mathcal{F}} \to \mathbb{R}$ such that

$$e(\mathcal{F}_{\Theta}, \mathcal{F}, d) = \mathbb{E}_{f \sim \mu_{\mathcal{F}}} \left[ c(f) \cdot \inf_{g \in \mathcal{F}} d(g, f) \right] \quad \forall \mathcal{F}_{\Theta}, \mathcal{F}, d \tag{2}$$

where $\mu_{\mathcal{F}}$ denotes the measure $\mu$ conditional on the event $\mathcal{F}$. If additionally $e$ satisfies Axiom 5, then

$$e(\mathcal{F}_{\Theta}, \mathcal{F}, d) = \mathbb{E}_{f \sim \lambda_{\mathcal{F}}} \left[ \inf_{g \in \mathcal{F}} c \cdot d(g, f) \right] \quad \forall \mathcal{F}_{\Theta}, \mathcal{F}, d \tag{3}$$

for a positive constant $c$, where $\lambda$ denotes the Lebesgue measure on $\overline{\mathcal{F}}$.

Our restrictiveness measure assumes (3), and normalizes the approximation error of model $\mathcal{F}$ relative to the approximation error of the baseline $f_{\text{base}}$.

## 5 Computation and Estimation

We now discuss how to implement our approach in practice. Recall that we restrict $\mathcal{X}$ to be finite, so $\mathcal{F}$ is finite-dimensional.

**Computing Restrictiveness** The following is an algorithm for computing $r$: Sample $M$ times independently from a uniform distribution on the eligible set $\mathcal{F}$. For each sampled
If \( f_m \in \mathcal{F} \), compute \( d(\mathcal{F}_\Theta, f_m) \) and \( d(f_{\text{base}}, f_m) \). Then

\[
\hat{r}_M := \frac{1}{M} \sum_{m=1}^{M} d(\mathcal{F}_\Theta, f_m) \quad \text{and} \quad \frac{1}{M} \sum_{m=1}^{M} d(f_{\text{base}}, f_m)
\]
is an estimator for restrictiveness \( r = r(\mathcal{F}_\Theta, \mathcal{F}) \). In principle, the number of simulations we run, \( M \), can be arbitrarily large, so \( \hat{r}_M \) can be made arbitrarily close to \( r \). Moreover, it is straightforward to obtain the formula for the asymptotic standard error of the simulated \( r \), based on which confidence intervals can be constructed.\(^{22}\)

**Estimating Completeness**

Suppose that the analyst has access to a finite sample of data \( \{Z_i := (X_i, Y_i)\}_{i=1}^{N} \) drawn from the unknown true distribution \( P^* \). To estimate completeness, which is defined based on the loss function \( l \) introduced in Section 3.2.2, we use \( K \)-fold cross-validation to estimate the out-of-sample prediction error of the model. (Our applications make the standard choice of \( K = 10 \).) Specifically, we randomly divide \( Z_N = (Z_1, \ldots, Z_N) \) into \( K \) (approximately) equal-sized groups. To simplify notation, assume that \( \frac{N}{K} \) is an integer. Let \( k(i) \) denote the group number of observation \( Z_i \), and fix an arbitrary set of maps \( \tilde{\mathcal{F}} \). In the \( k \)-th fold of cross-validation, we will use the observations in group \( k \) for testing and the remaining observations for training.

For each group \( k = 1, \ldots, K \), define \( \hat{f}^{-k} := \arg \min_{f \in \mathcal{F}} \frac{1}{N - J_N} \sum_{k(i) \neq k} l(f, Z_i) \) to be the minimizer in \( \tilde{\mathcal{F}} \) on the \( k \)-th training set (i.e., all observations outside of group \( k \)), and \( \hat{e}_k := \frac{1}{J_N} \sum_{k(i) = k} \left( \hat{f}^{-k}, Z_i \right) \) to be the out-of-sample error on the \( k \)-th test set. Then the average test error across the \( K \) folds, \( \hat{e}_{CV}(\tilde{\mathcal{F}}) := \frac{1}{K} \sum_{k=1}^{K} \hat{e}_k \), is an estimator for the unobservable expected error of the best prediction rule from class \( \tilde{\mathcal{F}} \). Setting \( \tilde{\mathcal{F}} \) to be \( \mathcal{F} \), \( \mathcal{G} \), or \( \{f_{\text{base}}\} \), we can compute \( \hat{e}_{CV}(\mathcal{F}) \), \( \hat{e}_{CV}(\mathcal{G}) \) and \( \hat{e}_{CV}(f_{\text{base}}) \) from the data, leading to the following

\(^{22}\)Under Assumption 1, \( \sqrt{M} \left( \hat{r}_M - r \right) / \hat{\sigma}_r \overset{d}{\to} \mathcal{N}(0, 1) \), where the asymptotic variance estimator \( \hat{\sigma}_r^2 \) is defined by \( \hat{\sigma}_r^2 := \left[ \hat{\sigma}_G^2 - 2 \hat{\sigma}_{G,f_{\text{base}}} + \hat{\sigma}_{f_{\text{base}}}^2 \right] / \left[ \frac{1}{M} \sum_{m=1}^{M} d(f_{\text{base}}, f_m) \right]^2 \), with \( \hat{\sigma}_G^2 \) being the sample variance of \( d(\mathcal{G}, f_m) \), \( \hat{\sigma}_{f_{\text{base}}}^2 \) the sample variance of \( d(f_{\text{base}}, f_m) \), and \( \hat{\sigma}_{G,f_{\text{base}}}^2 \) the sample covariance of \( d(\mathcal{G}, f_m) \) and \( d(f_{\text{base}}, f_m) \), across \( m = 1, \ldots, M \). We note that the standard error here simply measures the approximation error of \( r \) based on a finite number of simulations and do not reflect randomness in experimental data.
estimator for $\kappa$:

$$\hat{\kappa} = 1 - \frac{\hat{e}_{CV}(G) - \hat{e}_{CV}(F^*)}{\hat{e}_{CV}(f_{base}) - \hat{e}_{CV}(F^*)}.$$

It is crucial that the denominator in $\hat{\kappa}$ does not vanish asymptotically, so we impose the following assumption:

**Assumption 2** (Baseline is Imperfect). $e_P(f_{base}) - e_P(f^*) > 0$.

This assumption says that the baseline prediction rule performs strictly worse in expectation than the best prediction rule so there is some room for a model to do better. We show that $\hat{\kappa}$ is asymptotically normal by adapting Proposition 5 in Austern and Zhou (2020).

**Proposition 2.** Under Assumption 2 and some regularity conditions,\(^{23}\) $\sqrt{N} (\hat{\kappa} - \kappa) / \hat{\sigma}_{\hat{\kappa}} \xrightarrow{d} \mathcal{N}(0,1)$, where the variance estimator $\hat{\sigma}_{\hat{\kappa}}^2$ is as defined in Appendix C.2.

### 6 Application 1: Certainty Equivalents

**6.1 Setting**

Our first application is to the prediction of certainty equivalents for a set of 25 binary lotteries from Bruhin et al. (2010). Each lottery is described as a tuple $x = (\bar{z}, z, p)$, where $\bar{z} > z \geq 0$ are the possible prizes, and $p$ is the probability of the larger prize. Each observation consists of a lottery and a reported certainty equivalent by a given subject, so we can describe the feature space $\mathcal{X}$ by the 25 lottery tuples $(\bar{z}, z, p)$ in the Bruhin et al. (2010) data, and the outcome space by $\mathcal{Y} = \mathbb{R}$. Note that the residual uncertainty in $Y$ conditional on $X$ reflects heterogeneity in certainty equivalents reported across subjects for the same lottery.

We predict the average certainty equivalent (over subjects) for each lottery in this data set. A prediction rule for this problem is any function $f : \mathcal{X} \to \mathbb{R}$ from the 25 lotteries to their average certainty equivalents, and the discrepancy between two mappings is defined to be their average mean-squared distance $d(f, f') = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} (f(x) - f'(x))^2$.

\(^{23}\)See Appendix C for details of these assumptions.
We evaluate the restrictiveness and completeness of two economic models. First we consider a three-parameter version of *Cumulative Prospect Theory* indexed by $\theta = (\alpha, \gamma, \delta)$, which specifies a utility $w(p)v(\bar{z}) + (1 - w(p))v(\bar{z})$ for each lottery $(\bar{z}, \bar{z}, p)$, where

$$v(z) = z^\alpha, \quad w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma}.$$ \hfill (4)

The predicted certainty equivalent of a binary lottery is then given by $f_\theta(\bar{z}, \bar{z}, p) = v^{-1}(w(p)v(\bar{z}) + (1 - w(p))v(\bar{z}))$. Following the literature, we restrict $\alpha, \gamma \in [0, 1]$, and $\delta \geq 0$. We specify $\mathcal{F}$ as the set of all such functions $f_\theta$ with parameters $\theta$ in this range, and refer to this model simply as CPT. As a baseline, we consider the function $f_{\text{base}}$ that maps each lottery into its expected value, corresponding to $\alpha = \gamma = \delta = 1$.

Second, we consider the *Disappointment Aversion* model of Gul (1991), using a parametrization proposed in Routledge and Zin (2010) with the parameters $\lambda = (\alpha, \eta)$, where $\alpha \in [0, 1]$ and $\eta > -1$.\footnote{To facilitate comparison with CPT, we depart slightly from Routledge and Zin (2010) by imposing the functional form $v(z) = z^\alpha$ instead of $v(z) = z^\alpha/\alpha$.} The value function for money is the same as in (4), but the probability weighting function is given instead by $\bar{w}(p) = \frac{p}{1 + (1 - p)\eta}$. There are two parameters: $\alpha$ again reflects the curvature of the utility function, while $\eta > 0$ corresponds to “disappointment aversion,” i.e. aversion to realizations of the lottery that are worse than its certainty equivalent. Here the predicted certainty equivalent is $f_\lambda(\bar{z}, \bar{z}, p) = v^{-1}(\bar{w}(p)v(\bar{z}) + (1 - \bar{w}(p))v(\bar{z}))$. We specify $\mathcal{F}_\lambda$ as the set of all such functions and refer to this model as DA. Again, we use expected value as the baseline prediction, which corresponds to $\alpha = 1$ and $\eta = 0$ in DA.

### 6.2 Completeness

We evaluate completeness using mean-squared error as the loss function, i.e., if the reported certainty equivalent is $y$ when the model predicts $\hat{y}$, the loss in that observation is $(\hat{y} - y)^2$.

\footnote{This loss function is paired to the average mean-squared discrepancy function we used for measuring restrictiveness, see Appendix E for details.}

\footnote{This parametric form for $w(p)$ was used by Goldstein and Einhorn (1987) and Lattimore et al. (1992).}
CPT achieves a striking out-of-sample performance for predicting certainty equivalents in the Bruhin et al. (2010) data: it is 95% complete.\textsuperscript{27} Thus, the model achieves almost all of the possible improvement in prediction accuracy over the baseline.\textsuperscript{28} In contrast, DA is only 27% complete on the same data. One explanation is that CPT more precisely captures the observed risk preferences in the data than DA, but another possibility is that CPT is flexible enough to mimic most functions from binary lotteries to certainty equivalents, while DA imposes more substantial restrictions. These explanations have very different implications for how to interpret CPT’s empirical success compared to DA’s.

6.3 Restrictiveness

To distinguish between these explanations, we now compute the restrictiveness of the two models. We define the eligible set to be all prediction rules satisfying the following criteria:

(i) \( z \leq f(\tilde{z}, \bar{z}, p) \leq z \);

(ii) If \( \tilde{z} \geq \tilde{z}' \), \( z \geq z' \), and \( p \geq p' \) with at least one “\( \geq \)” strict, then \( f(\tilde{z}, \bar{z}, p) > f(\tilde{z}', \bar{z}', p') \).

Constraint (i) requires that the certainty equivalent is within the range of the possible payoffs, while (ii) is equivalent to monotonicity with respect to first-order stochastic dominance.\textsuperscript{29}

Table 1 reports the completeness and restrictiveness of both models.

\textsuperscript{27}Fudenberg et al. (2022) reports a similar finding for a sample of gain-domain and loss-domain lotteries.

\textsuperscript{28}This finding is consistent with Peysakhovich and Naecker (2017)’s result that CPT approximates the predictive performance of lasso regression trained on a high-dimensional set of features.

\textsuperscript{29}The CDF of a binary lottery with \( \tilde{z} > \bar{z} \) and \( 0 < p < 1 \) is \( F(z) = (1 - p)1\{\tilde{z} \leq z < \bar{z}\} + 1\{z \geq \bar{z}\} \), which is weakly decreasing in \((\tilde{z}, \bar{z}, p)\) for all \( z \), so \((\tilde{z}, \bar{z}, p)\) FOSD \((\tilde{z}', \bar{z}', p')\) if and only if \((\tilde{z}, \bar{z}, p) \gtrless (\tilde{z}', \bar{z}', p')\). There are many pairs of lotteries in the Bruhin et al. (2010) lottery data that can be compared via (ii), so these conditions are not vacuous.
<table>
<thead>
<tr>
<th># Param</th>
<th>Restrictiveness</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT</td>
<td>3</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>DA</td>
<td>2</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.06)</td>
</tr>
</tbody>
</table>

Table 1: Completeness for both models is estimated on the real data, which includes reported certainty equivalents by each of 179 subjects. Standard errors for the completeness estimates are computed using a block bootstrapping procedure that clusters together all observations from the same subjects, see Appendix D.1. Restrictiveness is estimated from 1000 simulations.

The restrictiveness of CPT is 0.28, so on average CPT’s approximation error is about one fourth of the error of the expected value. DA is more restrictive, with an average approximation error almost one half of the error of the baseline. Thus the two models are not directly comparable: CPT performs substantially better for predicting the real data, but would have performed well out-of-sample given sufficient data from almost any underlying data-generating process that respects first-order stochastic dominance. DA rules out more behaviors that satisfy first-order stochastic dominance, but in doing so is unable to well approximate the actual Bruhin et al. (2010) data.

6.4 The Role of a Parameter

In addition to comparing models such as CPT and DA, our approach can be used to learn more about the role played by specific parameters. Adding a parameter must at least weakly decrease restrictiveness and increase completeness, but we find that parameters can differ substantially in their effectiveness in trading off between these two goals. We also show that models with the same number of parameters can have very different levels of restrictiveness, and thus a simple parameter count is substantively less informative than our measure.

Specifically, we consider alternative specifications of CPT and DA with fewer parameters. Some of these specifications have been studied in the literature: CPT(α, γ), with δ = 1, is used in Karmarkar (1978); CPT(γ, δ), with α = 1, corresponds to a risk-neutral CPT agent.

\[ w(p) = \frac{p^\gamma}{p^\gamma + (1-p)^\gamma} \]

This specification with weighting function \( w(p) = \frac{p^\gamma}{p^\gamma + (1-p)^\gamma} \) is very similar to one used in Tversky and Kahneman (1992), where the weighting function was \( w(p) = \frac{p^\gamma}{p^\gamma + (1-p)^\gamma}^{1/\gamma} \).
whose utility over money is \( u(z) = z \) but exhibits nonlinear probability weighting; \( \text{CPT}(\alpha) \), with \( \delta = \gamma = 1 \), corresponds to an Expected Utility decision-maker whose utility function is as given in (4), and is also equivalent to \( \text{DA}(\alpha) \).\(^{31}\) The model \( \text{CPT}(\gamma) \), with \( \alpha = \delta = 1 \), and \( \text{CPT}(\delta) \), with \( \alpha = \gamma = 1 \) have not been studied in the literature, but we report them for comparison. We also consider \( \text{DA}(\eta) \) as in Gul (1991), with \( \alpha = 1 \), which corresponds to a disappointment-averse decision maker whose utility is linear in money.

![Figure 2: Comparison of models by their completeness and restrictiveness.](image)

Figure 2 plots restrictiveness and completeness for these alternative specifications, which reveals that some specifications fall in the interior of the restrictiveness-completeness Pareto frontier introduced in Section 3.2.3: Each of \( \text{CPT}(\delta) \), \( \text{CPT}(\alpha, \delta) \), and \( \text{DA}(\alpha, \eta) \) are dominated, in the sense that another model is simultaneously more complete and also more

\(^{31}\)See the survey Fehr-Duda and Epper (2012) for further discussion of these different parametric forms, and others which have been used in the literature.
restrictive.\textsuperscript{32} The figure also reveals substantial dispersion in the restrictiveness of these specifications (ranging from $r = 0.28$ to $r = 0.92$), even though all of the specifications use only a small number of parameters. This observation emphasizes the distinction between our method and a simple parameter count.

By looking more specifically at how restrictiveness and completeness vary across two nested specifications, we can better understand the role that any specific parameter plays. Figure 3 shows that the different parameters for probability weighting are not equally effective. Adding the parameter $\delta$, which governs the elevation of the probability weighting curve, to any specification of CPT leads to a large drop in restrictiveness in return for only a small gain in completeness. We find a similar result for the “disappointment aversion” parameter $\eta$ in DA, which barely improves upon the completeness of DA($\alpha$), but leads to a substantial drop in restrictiveness. In contrast, the parameter $\gamma$, which governs the curvature of the probability weighting function, appears to play an important role in capturing risk preferences: Adding $\gamma$ to any CPT specification leads to a sizeable improvement in completeness at the cost of a modest reduction in restrictiveness. This supports previous findings that probability distortions play an important role in fitting experimental and field data (Snowberg and Wolfers, 2010; Fehr-Duda and Epper, 2012; Barseghyan et al., 2013).

6.5 Robustness Checks

We show that the qualitative findings in this section are robust to changes in the eligible set and changes in the lotteries considered.

Different distribution over the eligible set. The uniform distribution is the same as beta(1, 1), so to test the sensitivity of the restrictiveness measure we consider nearby beta($a, b$) distributions with parameters ($a, b$) sampled from a uniform distribution over $[0, 1]$. See also Table 5 in Online Appendix D.1.

\textsuperscript{32}CPT($\delta$) is simultaneously less complete and less restrictive than DA($\eta$), while both CPT($\alpha, \delta$) and DA($\alpha, \eta$) are less complete and less restrictive than the single parameter model CPT($\gamma$).
Figure 3: Impact of the probability weighting parameters on completeness and restrictiveness.

[0.9, 1.1] × [0.9, 1.1]. For each (a, b) pair, we generate certainty equivalents from a beta(a, b) distribution over the prize range, again keeping only those functions f that satisfy FOSD. Over 100 such distributions beta(a, b), the average restrictiveness is 0.29, with a minimum value of 0.27 and a maximum value of 0.32.

**Different eligible set.** Next, we compute the restrictiveness of CPT(α, δ, γ) with respect to an eligible set that imposes the range restriction in (i) but drops the FOSD restrictions in (ii). The model’s errors are substantially higher when we drop FOSD (increasing from 63.75 to 102.41), but so are the errors of the Expected Value benchmark. The relative
performance of CPT(α, δ, γ) compared to the expected-value baseline is nearly identical regardless of whether or not we impose FOSD: the model’s restrictiveness relative to this larger eligible set is 0.29 (compared to 0.28 relative to the original eligible set).

**Other sets of binary lotteries.** In our main analysis, the feature space X consisted of 25 binary lotteries from Bruhin et al. (2010) data. Below we report the restrictiveness of CPT(α, γ, δ) and DA(α, η) with respect to alternative sets of binary lotteries, drawn from five additional papers (see Appendix D.3 for details). Figure 4 shows the CDF of restrictiveness values across these lotteries (including the Bruhin et al. (2010) lotteries) for both models. We find that CPT is not very restrictive on any of these sets of lotteries, and that the distribution of restrictiveness for DA first-order stochastically dominates that of CPT.

![Figure 4: CDF of restrictiveness values](image)

**Lotteries over the loss domain.** On 25 binary lotteries over the loss domain from Bruhin et al. (2010), the 3-parameter specification of CPT indexed to (β, γ, δ) predicts the certainty equivalent $v^{-1}((1 - w(1 - p)) \cdot v(z) + w(1 - p) \cdot v(\tilde{z}))$ for each lottery $(z, \tilde{z}, p)$, where $v(z) = -((-z)^\beta)$ and $w(p) = (\delta p_\gamma)/(\delta p_\gamma + (1 - p)_\gamma)$. The restrictiveness of CPT on these lotteries is 0.31, with a standard error of 0.02.
Lotteries with larger supports. Finally, we evaluate the restrictiveness of CPT(α, δ, γ) on gains-domain lotteries with more than two possible outcomes. For each lottery (z_1, ..., z_n; p_1, ..., p_n), where 0 ≤ z_1 < ... < z_n, the predicted certainty equivalent is

\[ v^{-1} \left( \sum_i u(x_i) \left[ w \left( \sum_k p_k \right) - w \left( \sum_{k=1}^{i-1} p_k \right) \right] \right), \]

where for i = 1 we define \( \sum_{k=1}^{0} p_k = 0 \), and v and w have the same functional forms as used above. On 18 three-outcome gain-domain lotteries from Bernheim and Sprenger (2020b), the restrictiveness of CPT is 0.57, with a standard error of 0.02. Thus CPT is about twice as restrictive for certainty equivalents on three-outcome lotteries as it is on binary lotteries. On a set of 10 six-outcome lotteries from Fudenberg and Puri (2021), the restrictiveness of CPT is 0.83, with a standard error of 0.01. These results suggest that CPT is more restrictive on lotteries with larger supports.

7 Application 2: The Distribution of Initial Play

7.1 Setting

Our second application is to predicting the distribution of initial play in games. Here the feature space \( \mathcal{X} \) consists of the 466 unique 3 × 3 payoff matrices from Fudenberg and Liang (2019). The outcome space is the set \( \mathcal{Y} = \Delta(\{a_1, a_2, a_3\}) \) of distributions of row player actions chosen by the participants in the experiments. The analyst seeks to predict this distribution for each game.

For any two prediction rules \( f \) and \( f' \), we define \( d(f, f') \) to be the average Kullback-Liebler divergence between the predicted distributions: \( d(f, f') = \frac{1}{466} \sum_{x \in \mathcal{X}} D(f(x)\|f'(x)), \)

These data are an aggregate of three data sets: the first is a meta data set of play in 86 games, collected from six experimental game theory papers in Wright and Leyton-Brown (2014); the second is a data set of play in 200 games with randomly generated payoffs, which were gathered on MTurk for Fudenberg and Liang (2019); the third is a data set of play in 200 games that were “algorithmically designed” for a certain model (level 1 with risk aversion) to perform poorly, again from Fudenberg and Liang (2019).
where $D$ denotes the Kullback-Leibler divergence.

We consider three economic models: The Poisson Cognitive Hierarchy Model (PCHM) of Camerer et al. (2004), the Level-1 model with logistic best replies (henceforth Logit Level-1), and the PCHM with logistic best replies (henceforth Logit PCHM). The PCHM supposes that there is a distribution over players of differing levels of sophistication: The level-0 player randomizes uniformly over his available actions, the level-1 player best responds to level-0 play (Stahl and Wilson, 1994, 1995; Nagel, 1995); and for $k \geq 2$, level-$k$ players best respond to a perceived distribution

$$p_k(h, \tau) = \frac{\pi_\tau(h)}{\sum_{l=0}^{k-1} \pi_\tau(l)} \quad \forall h \in \mathbb{N}_{<k}$$

over (lower) opponent levels, where $\pi_\tau$ is the Poisson distribution with rate parameter $\tau \geq 0$. The parameter $\tau$ is the single parameter of the model.

The Logit Level-1 prediction is defined as follows. For each row player action $a_i$, let $\bar{u}(a_i)$ be the expected payoff of $a_i$ when the column player uses a uniform distribution. The predicted frequency with which $a_i$ is played is $\exp(\lambda \cdot \bar{u}(a_i)) / \sum_{i=1}^{3} \exp(\lambda \cdot \bar{u}(a_i))$, where the logit parameter $\lambda \in \mathbb{R}_+$ is the single parameter of the model.

The Logit PCHM (see e.g. Wright and Leyton-Brown (2014)) replaces the assumption of exact maximization in the PCHM with a logit best response. That is, the level-0 player chooses $f_0 = (1/3, 1/3, 1/3)$ as in the PCHM, but we recursively construct the distribution of play for higher levels as follows. For each $k \geq 1$, define

$$v_k(a_i) = \sum_{h=0}^{k-1} p_k(h, \tau) \left( \sum_{j=1}^{3} f_h(a_j) u(a_i, a_j) \right)$$

to be the expected payoff of action $a_i$ against a player whose type is distributed according to $p_k(\cdot, \tau)$, where $p_k(h, \tau)$ is as given in (5). The distribution of play for a level-$k$ player is then $f_k(a_i) = \exp(\lambda \cdot v_k(a_i))/\sum_{j=1}^{3} \exp(\lambda \cdot v_k(a_j))$, where $\lambda \in \mathbb{R}_+$ is a logit parameter. We aggregate across levels using a Poisson distribution with rate parameter $\tau \in \mathbb{R}_+$ to yield the
predicted distribution of play.

Finally, we define the baseline prediction rule \( f_{\text{base}} \) to predict uniform play in every game \( x \). This prediction rule is nested in all three models.\(^{34}\)

7.2 Completeness

We evaluate completeness using negative log-loss as the loss function, i.e., if the chosen action is \( a_i \) when the model predicts distribution \( (p_1, p_2, p_3) \), the loss in that observation is \(-\log(p_i)\).\(^{35}\) The models PCHM, Logit Level-1, and Logit PCHM are 43.6%, 72.7%, and 72.9% complete. Thus, as observed in a related study by Wright and Leyton-Brown (2014), Logit PCHM provides much better predictions of the distribution of play than the baseline PCHM does. Perhaps surprisingly, almost all of Logit PCHM’s improved performance can be obtained by simply adding the logit parameter to the Level-1 model; the further improvement from allowing for multiple levels of sophistication is negligible.\(^{36}\)

7.3 Restrictiveness

We turn now to evaluating the restrictiveness of these models. We have relatively little understanding about their empirical content, but we do know that they all imply that if an action is strictly dominated, then the frequency with which it is chosen does not exceed 1/3, and that if an action is strictly dominant, then the frequency with which it is chosen is at least 1/3. We define the eligible set to be all prediction rules that satisfy these conditions.\(^{37}\)

All three models are very restrictive relative to this eligible set: Logit Level-1’s restric-

\(^{34}\)Let \( \tau = 0 \) in the PCHM or Logit PCHM, and let \( \lambda = 0 \) in Logit Level-1.

\(^{35}\)This loss function is paired to the Kullback-Leibler discrepancy function we used for measuring restrictiveness, see Appendix E for details.

\(^{36}\)Fudenberg and Liang (2019) found that the Level-1 model provides a good prediction of the modal action, but this does not imply that Logit Level-1 will perform well in predicting the full distribution of play. The fact that it does further suggests that initial play in many of these experiments is rather unstrategic.

\(^{37}\)In our data, the median frequency of a strictly dominated action is 0.03, and the highest frequency is 0.35; the median frequency for a strictly dominant action is 0.86, and the lowest frequency is 0.69. Payoff maximization implies that dominant strategies should have probability 1 and dominated strategies have probability 0, but this is inconsistent with observed play in most game theory experiments.
Table 2: Restrictiveness and Completeness for Initial Play

<table>
<thead>
<tr>
<th># Param</th>
<th>Restrictiveness</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCHM 1</td>
<td>0.992</td>
<td>0.436</td>
</tr>
<tr>
<td></td>
<td>(&lt;0.001)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>logit level-1 1</td>
<td>0.970</td>
<td>0.727</td>
</tr>
<tr>
<td></td>
<td>(&lt;0.001)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>logit PCHM 2</td>
<td>0.971</td>
<td>0.729</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.014)</td>
</tr>
</tbody>
</table>

Restrictiveness estimated from 1000 simulations.

...tiveness is 0.970, PCHM’s restrictiveness is 0.992, and Logit PCHM’s restrictiveness is 0.971. Since the models’ completeness ranges from 0.436 to 0.729, they are much better predictors of the real data than of the synthetic data. Table 7.3 reports completeness and restrictiveness measures for the models. We find that Logit Level-1 and Logit PCHM are substantially more complete than PCHM and only slightly less restrictive, but none of the models is dominated by another. Moreover, Logit Level-1 and Logit PCHM are almost identical in terms of completeness and restrictiveness, even though the parametric forms of the two models are not evidently related.\(^{38}\) Appendix F investigates the relationship between Logit Level-1 and Logit PCHM further by studying the correlation in their errors.

Finally, as a robustness check, we consider strengthening the background constraints imposed on the eligible set \(F\). For each \(t \in [0, 0.3]\), we define the eligible set \(F(t)\) to include all prediction rules \(f\) that satisfy the following conditions: (1) If an action is strictly dominated, then the frequency with which it is chosen does not exceed \(1/3 - t\); (2) If an action is strictly dominant, then the frequency with which it is chosen is at least \(1/3 + t\). The constraint imposed by these conditions increases in \(t\), and \(t = 0\) returns our original specification of \(F\). We find that across choices of \(t \in [0, 0.3]\), the restrictivenesses of PCHM, Logit PCHM, and Logit Level-1 do not fall below 0.89 (see Table 3 below). This tells us that constraints on the frequency of strictly dominated and strictly dominant strategies are a very small part of the empirical content of these models.

\(^{38}\)No value of \(\tau\) in the PCHM yields the Level-1 model, so Logit Level-1 is not nested within Logit PCHM.
<table>
<thead>
<tr>
<th>PCHM</th>
<th>Logit Level-1</th>
<th>Logit PCHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>max</td>
<td>0.993</td>
<td>0.969</td>
</tr>
<tr>
<td>min</td>
<td>0.974</td>
<td>0.890</td>
</tr>
</tbody>
</table>

Table 3: Highest and lowest smallest restrictiveness for \( t \in [0, 0.3) \).

8 Application 3: Diffusion in Social Networks

8.1 Setting

Our final application is to the prediction of microfinance takeup rates following diffusion of information in social networks. We use data from a study by Banerjee et al. (2013), in which certain “leaders” in 43 villages in Karnataka, India were given information about a microfinance program, and takeup of the program was then tracked.\(^{39}\)

For each village \( i \), let \( y_i \) be the average takeup rate among non-leader households.\(^{40}\) Our goal is to predict \( y_i \) given the observed characteristics \( X_i \) of village \( i \). Specifically, a village configuration \( X_i := (N_i, A_i, L_i) \) consists of a set \( N_i \) of villagers, an \( n_i \times n_i \) adjacency matrix \( A_i \) that represents the measured social network, and the set \( L_i \) of leaders in village \( i \). The feature space \( \mathcal{X} \) is the collection of 43 village configurations, and prediction rules are maps \( f : \mathcal{X} \to [0, 1] \) that from village configurations to the takeup rate among non-leaders. There are no obvious a priori restrictions on the takeup rates, so we set \( \mathcal{F} \) to be the set \([0, 1]^{43}\) of all possible prediction rules from \( \mathcal{X} \) to \([0, 1]\). We set the discrepancy function as \( d(f, g) := \frac{1}{43} \sum_{i=1}^{43} (f(x_i) - g(x_i))^2 \) and the loss function as \( l(f(x), y) := (f(x) - y)^2 \).

8.2 Models

The first parametric models we consider are OLS regressions with various subsets of the following eight network statistics as regressors: (1) average eigenvector centrality of leaders;\(^{39}\)

---

\(^{39}\)In 2007, the microfinance institution Bharatha Swamukti Samsthe invited leaders within each village to an information meeting, and asked the leaders to spread the information. The data set contains the resulting microfinance takeup rate for each village and some measures of social connections between households.

\(^{40}\)This is the outcome variable that Banerjee et al. (2013) focus on.
(2) average degree centrality of leaders; (3) average degree centrality of all villagers; (4) average betweenness centrality of leaders; (5) clustering coefficient of village network; (6) average path length in village network; (7) proportion of connected (non-isolated) villagers; (8) proportion of leaders.

We compute the restrictiveness and completeness of a sequence of OLS models by incrementally adding the regressors listed above. We set the baseline as OLS regression on a constant, which is a special case of all the linear models we consider. With the loss function 
\[ l(f(x), y) := (y - f(x))^2, \]
an estimator of completeness (computed based on in-sample errors without the use of cross validations) reduces to the \( R^2 \) of the OLS regression.\(^{41}\)

We also consider a partially linear model built upon the “network gossip centrality” described in Banerjee et al. (2019). To do this, we model each non-leader household’s takeup probability as a function of its position in the village. We define the “hearing matrix” of village \( i \) by 
\[ H_i(\theta_1) := \sum_{t=1}^{T} \theta_1 A_t^i, \]
where \( T \) is some given number of time periods for information diffusion.\(^{42}\) With \( \theta_1 = 1 \), the \( jk \)-th entry of \( H_i(1) \) can be interpreted as the expected number of times villager \( k \) hears a piece of information that originates from villager \( j \) within \( T \) periods of time. The parameter \( \theta_0 \in (0, 1) \) discounts longer paths of diffusion. For each non-leader \( k \) in village \( i \), we define \( x_{i,k}(\theta_1) := \sum_{j \in L_i} (H_i(\theta_1))_{jk} \) as the “network gossip centrality” of non-leader \( k \), which counts the (discounted) sum of number of paths from the leaders of village \( i \) to non-leader \( k \). Next, we model the takeup probability of non-leader \( k \) as function of \( k \)’s “network gossip centrality” based on a logistic model 
\[ p_{i,j}(\theta_0, \theta_1) := \frac{\exp(\theta_0 + x_{i,j}(\theta_1))}{1 + \exp(\theta_0 + x_{i,j}(\theta_1))}, \]
where \( \theta_0 \) is a location parameter.\(^{43}\) The expected village-level takeup rate among non-leaders can then be derived as the average \( p_{i,j}(\theta_0, \theta_1) \) among non-leaders. To allow additional flexibility, and to nest the naive constant model as a special case, we introduce

---

\(^{41}\) Recall that the \( R^2 \)-squared of an OLS regression is defined by 
\[ R^2 := 1 - \frac{SSR}{SST}, \]
where \( SSR := \sum_i (y_i - \hat{y}_i)^2 \) is the expected loss under an OLS regression model and \( SST := \sum_i (y_i - \bar{y})^2 \) is the expected loss under a constant model.

\(^{42}\) \( \sum_{t=1}^{T} A_t^i \) \( jk \) counts the number of paths from \( j \) to \( k \) of length up to \( T \). We set \( T = 5 \) following Banerjee et al. (2019).

\(^{43}\) Note that we do not include a scale parameter here, since if present, it will be absorbed into \( \theta_1 \).
two additional linear parameters \((\theta_2, \theta_3)\), and set: \(f_i(\theta) := \theta_2 + \theta_3 \cdot \frac{1}{|N_i \setminus L_i|} \sum_{j \notin L_i} p_{ij}(\theta_0, \theta_1)\). This model is very stylized; our purpose is to illustrate how our algorithmic approach can be used to evaluate the restrictiveness of a structural model whose flexibility is otherwise difficult to gauge.

### 8.3 Results

#### Table 4: Restrictiveness and Completeness for Microfinance Takeup Rates

<table>
<thead>
<tr>
<th></th>
<th># Param</th>
<th>Restrictiveness</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linear Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvector Centrality of Leaders</td>
<td>1</td>
<td>0.9762 (0.0003)</td>
<td>0.2577 (0.1101)</td>
</tr>
<tr>
<td>+ Degree Centrality of Leaders</td>
<td>2</td>
<td>0.9526 (0.0004)</td>
<td>0.3385 (0.1193)</td>
</tr>
<tr>
<td>+ Degree Centrality of All Villagers</td>
<td>3</td>
<td>0.9288 (0.0005)</td>
<td>0.3471 (0.1151)</td>
</tr>
<tr>
<td>+ Betweenness Centrality of Leaders</td>
<td>4</td>
<td>0.9053 (0.0006)</td>
<td>0.3475 (0.1158)</td>
</tr>
<tr>
<td>+ Clustering Coefficient</td>
<td>5</td>
<td>0.8816 (0.0007)</td>
<td>0.3516 (0.1191)</td>
</tr>
<tr>
<td>+ Average Path Length</td>
<td>6</td>
<td>0.8579 (0.0007)</td>
<td>0.3516 (0.1191)</td>
</tr>
<tr>
<td>+ Proportion of Connected Villagers</td>
<td>7</td>
<td>0.8342 (0.0008)</td>
<td>0.3575 (0.1229)</td>
</tr>
<tr>
<td>+ Proportion of Leaders</td>
<td>8</td>
<td>0.8101 (0.0008)</td>
<td>0.3604 (0.1237)</td>
</tr>
<tr>
<td><strong>Partially Linear Model</strong></td>
<td>4</td>
<td>0.9408 (0.0036)</td>
<td>0.0674 (0.0452)</td>
</tr>
</tbody>
</table>

Table 4 reports the restrictiveness and completeness of the models described above.\(^{44}\) The panel “Linear Models” contains results about a sequence of linear models, with a new regressor added to the OLS regression in each row.\(^{45}\) For example, the row “+ Degree Centrality” corresponds to an OLS regression of uptake rates on a constant, the leaders’

---

\(^{44}\)Table 4 displays the restrictiveness of the linear models based on \(M = 10000\) simulations, while restrictiveness for the partially linear models is computed using \(M = 100\) simulations. Completeness for all models is computed based the real data with \(N = 43\) villages.

\(^{45}\)We add the regressors sequentially according to the ordering above, and omit many other different orderings of the same set of regressors, since the regressions in Table 4 suffice to illustrate our main point.
average eigenvector, and the leaders’ average degree centrality.

The numerical results for linear models are as expected: as more regressors are added the model becomes more flexible, so restrictiveness decreases while completeness increases. While restrictiveness seems to be decreasing at an approximately linear rate starting from the second regression, the corresponding increases in completeness appear less uniform, and in particular, completeness barely changes when we add the regressor “average path length in the village.” Note that this does not mean that this additional regressor approximately lies in the linear span of all previously included regressors, since we do observe a nontrivial reduction in restrictiveness from the addition of this regressor: New regressors eventually barely improve fit to the data, but they continue to decrease restrictiveness.

A priori it is unclear how restrictive the partially linear model is. It turns out that its restrictiveness is very high, 0.94, suggesting that the individual-level modeling of takeup rates as a function of network gossip centrality imposes substantial restrictions across village configurations. However, this model’s completeness is only 0.07, so it does not capture much of the variation in village takeup rates.

This four-parameter partially linear model is dominated by the simple linear model with a constant and the average eigenvector centrality of leaders as the single regressor: the latter has both higher restrictiveness (0.9762 > 0.9408) and higher completeness (0.2577 > 0.0674). This shows that even a detailed, structured, and economically-motivated model may turn out to be more flexible than a simple linear model, and that the added flexibility need not help it fit real data.

9 Conclusion

When a theory fits the data well, it matters whether this is because the theory captures important regularities in the data, or whether the theory is so flexible that it can explain any behavior at all. We provide a practical, algorithmic approach for evaluating the re-
strictiveness of a theory, and demonstrate that it reveals new insights into models from two economic domains. The method is easily applied to models across diverse domains.

As highly flexible machine learning methods become more popular in economics, economic theory is distinguished in part by the structure it imposes on behaviors. We view these restrictions as an important part of the value added by economic theory, so it is natural to ask how restrictive economic models are compared to the highly flexible approaches used in machine learning. Our restrictiveness measure offers a way to quantify this.

References


Barseghyan, L., F. Molinari, T. O’Donoghue, and J. C. Teitelbaum (2013):


### A Proof of Proposition 1

Throughout this proof, we use $\Sigma$ to denote the Lebesgue $\sigma$-algebra on $\mathcal{F}$, and shorten $\Sigma$-measurable to simply “measurable.”

It is clear that A1-A4 are satisfied by the representation in (2), and A1-A5 are satisfied by the approximation error measure given in (3). For the other direction, we begin by demonstrating the following lemma:
Lemma A.1. Suppose $e$ satisfies A1 and A4. Then for every $\mathcal{F}_\Theta$ and $d$, there exists a function $h : \mathcal{F} \to \mathbb{R}$ such that $e(\mathcal{F}_\Theta, \mathcal{F}, d) = \mathbb{E}[h(f) : f \sim \mu_\mathcal{F}]$ for all measurable sets $\mathcal{F}$.

Proof. Fix an arbitrary $\mathcal{F}_\Theta$ and $d$, and define $e_\ast : \Sigma \to \mathbb{R}$ to satisfy $e_\ast(\mathcal{F}) = e(\mathcal{F}_\Theta, \mathcal{F}, d)$ for all measurable $\mathcal{F}$. The lemma follows if we can show that A4 implies the existence of a function $h : \mathcal{F} \to \mathbb{R}$ such that $e_\ast(\mathcal{F}) = \int_\mathcal{F} h(f) d\mu_\mathcal{F}$ for all measurable $\mathcal{F}$, where $\mu_\mathcal{F}$ denotes the measure $\mu$ conditional on the event $\mathcal{F}$.

Define $\nu : \Sigma \to \mathbb{R}$ to satisfy $\nu(\mathcal{F}) = \mu(\mathcal{F}) \cdot e_\ast(\mathcal{F})$ for all measurable $\mathcal{F}$. Then A4 implies that for any sequence $\mathcal{F}_{\Theta_1}, \mathcal{F}_{\Theta_2}, \ldots$, $\sum_{i=1}^{\infty} \nu(\mathcal{F}_{\Theta_i}) = \nu(\bigcup_{i=1}^{\infty} \mathcal{F}_{\Theta_i})$. Also, $\nu(\emptyset) = 0$ (since $\mu(\emptyset) = 0$) and $\nu$ is non-negative (by A1), so $\nu$ is a measure on $(\mathcal{F}, \Sigma)$. Moreover, $\nu$ is absolutely continuous with respect to $\mu$ by construction. So the Radon-Nikodym theorem implies existence of a function $h : \mathcal{F} \to \mathbb{R}$ such that $\nu(\mathcal{F}) = \int_\mathcal{F} h(f) d\mu_\mathcal{F}$ for all measurable $\mathcal{F}$.

Then $\mu(\mathcal{F}) e_\ast(\mathcal{F}) = \mu(\mathcal{F}) \int_\mathcal{F} h(f) \frac{d\mu}{\mu(\mathcal{F})} = \mu(\mathcal{F}) \int_\mathcal{F} h(f) d\mu_\mathcal{F}$, so $e_\ast(\mathcal{F}) = \int_\mathcal{F} h(f) d\mu_\mathcal{F}$. \hfill $\Box$

Now fix any $\mathcal{F}_\Theta$ and $d$, and let $h$ be the function given in Lemma A.1. We will show that A2 and A3 imply that for each $f \in \mathcal{F}$,

$$h(f) = c_f \cdot d(\mathcal{F}_\Theta, f) \quad \text{(A.1)}$$

for some constant $c_f \in \mathbb{R}_+$. 

Fix an arbitrary $f$. Lemma A.1 implies $e(\mathcal{F}_\Theta, \{f\}, d) = \int h(f') \cdot d\delta_f = h(f)$, where $\delta_f$ denotes the Dirac measure at $f$. So it is sufficient for (A.1) to show that there is a constant $c_f \in \mathbb{R}_+$ such that $e(\mathcal{F}_\Theta, \{f\}, d) = c_f \cdot d(\mathcal{F}_\Theta, f)$ for all $\mathcal{F}_\Theta, d$. By A2, models can be completely ordered for the eligible set $\{f\}$, where $e(\mathcal{F}_{\Theta_1}, \{f\}, d) \geq e(\mathcal{F}_{\Theta_2}, \{f\}, d)$ if and only if $d(\mathcal{F}_{\Theta_1}, f) \geq d(\mathcal{F}_{\Theta_2}, f)$. So there is a monotone increasing function $\Phi : \mathbb{R} \to \mathbb{R}$ such that

$$e(\mathcal{F}_\Theta, \{f\}, d) = \Phi(d(\mathcal{F}_\Theta, f)). \quad \text{(A.2)}$$

Now we will show that $\Phi$ must be linear. Choose an arbitrary $\alpha \in \mathbb{R}_+$. Define $d' = \alpha \cdot d$
and suppose some model $\mathcal{F}_{\Theta'}$ satisfies $d(\mathcal{F}_{\Theta'}, f) = \alpha \cdot d(\mathcal{F}_{\Theta}, f)$. Then $e(\mathcal{F}_{\Theta}, \{f\}, d') = \alpha \cdot e(\mathcal{F}_{\Theta}, \{f\}, d) = \alpha \cdot \Phi(d(\mathcal{F}_{\Theta}, f))$, where the first equality follows by (A3) and the second follows by (A.2). Also $e(\mathcal{F}_{\Theta'}, \{f\}, d) = \Phi(d(\mathcal{F}_{\Theta'}, f)) = \Phi(\alpha \cdot d(\mathcal{F}_{\Theta}, f))$, where the first equality follows by (A.2). A3 requires $e(\mathcal{F}_{\Theta'}, \{f\}, d) = e(\mathcal{F}_{\Theta}, \{f\}, d')$, so $\alpha \cdot \Phi(d(\mathcal{F}_{\Theta}, f)) = \Phi(\alpha \cdot d(\mathcal{F}_{\Theta}, f))$. Thus we can write $e(\mathcal{F}_{\Theta}, \{f\}, d) = c_f \cdot d(\mathcal{F}_{\Theta}, f)$ for some constant $c_f \in \mathbb{R}_+$. Repeating this argument for every $f$, there is a function $c : \mathcal{F} \to \mathbb{R}$ such that $e(\mathcal{F}, \mathcal{F}, d) = \mathbb{E}_{f \sim \mu_F} [c(f) \cdot d(G, f)]$ for all measurable $\mathcal{F}$, so we have the representation in (2).

Now suppose that A5 is satisfied in addition to the other axioms. The previous arguments imply that there is a function $c : \mathcal{F} \to \mathbb{R}$ such that

$$e(\mathcal{F}_{\Theta}, \mathcal{F}, d) = \mathbb{E}_{f \sim \mu_F} \left[ c(f) \cdot \inf_{f' \in \mathcal{F}_{\Theta}} d(f', f) \right] \quad \forall \mathcal{F}_{\Theta}, \mathcal{F}, d$$

Suppose towards contradiction that $e$ cannot be represented by (3). Then there must exist an eligible set $\mathcal{F}$ and $f, f' \in \mathcal{F}$ such that $c(f) \cdot \mu_F(f) > c(f') \cdot \mu_F(f')$. But then for any models $\mathcal{F}_{\Theta_1}$ and $\mathcal{F}_{\Theta_2}$ with the property that $d(\mathcal{F}_{\Theta_1}, f) = d(\mathcal{F}_{\Theta_2}, f') > d(\mathcal{F}_{\Theta_2}, f) = d(\mathcal{F}_{\Theta_1}, f')$, it follows that $e(\mathcal{F}_{\Theta_1}, \{f, f'\}, d) > e(\mathcal{F}_{\Theta_2}, \{f, f'\}, d)$, violating A5.
B A Guide for Practitioners

Below we provide detailed instructions for how to take the proposed measures to other applications.

B.1 Setup

The Prediction Problem and Model. We suppose that the researcher has a dataset that can be described as a set of observations \((x, y)\), where \(x\) is interpreted as an observable input, and \(y\) is interpreted as the outcome to be predicted. Define

- the set of features \(\mathcal{X}\) to consist of all unique instances of \(x\) in the analyst’s data (thus by construction finite).
- the set of outcomes \(\mathcal{Y} \subseteq \mathbb{R}^k\) to be the set in which \(y\) takes values.

Let \(\mathcal{F} = \mathcal{Y}^\mathcal{X}\) be the set of all mappings from \(\mathcal{X}\) to \(\mathcal{Y}\).

The researcher is interested in studying the properties of some parametric model \(\mathcal{F}_\Theta = \{f_\theta\}_{\theta \in \Theta}\), where each \(f_\theta\) belongs to \(\mathcal{F}\).
Baseline. Choose a “baseline mapping” $f_{\text{base}}$ from the model $\mathcal{F}_\Theta$. The purpose of the baseline is to provide a lower bound for error that any sensible model should outperform. Some possibilities for how to choose this baseline include:

- choosing a “degenerate” version of the model with the parameters fixed at some default values (for example, Expected Value as a degenerate case of Cumulative Prospect Theory, as in our Application 1)

- choosing a mapping that corresponds to “guessing at random” (e.g., predicting a uniform distribution over the possible outcomes, as in our Application 2)

- choosing a best constant prediction based on the data (e.g., regressing a linear model on a constant, as in our Application 3)

The choice of baseline mapping should be reported along with estimates of restrictiveness and completeness, and a natural robustness check is to verify that these estimates do not change significantly over different (reasonable) choices of baseline.

B.2 Evaluating Restrictiveness

The Admissible Set. The researcher first determines the admissible set $\mathcal{F}$, which is a subset of mappings from $\mathcal{X}$ to $\mathcal{Y}$ that satisfy some given properties. Which and how many properties to choose depends on what the researcher wants to understand. If the researcher wants to know whether the model imposes any restrictions at all, then the admissible set should include all mappings from $\mathcal{X}$ to $\mathcal{Y}$. If the researcher wants to know how restrictive the model is beyond imposing some Property A, then the admissible set should include only mappings that are consistent with Property A.

The Discrepancy Function $d$. Next the researcher chooses a discrepancy function $d : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}_+$ that tells us how different any two mappings $f$ and $f'$ are. Although we leave this specification open to the researcher, there are sometimes standard choices. When the
outcome space $\mathcal{Y}$ is real-valued, a natural choice is the expected squared distance between the predictions of $f$ and $f'$, namely

$$d(f, f') = \mathbb{E}_{P_X} [f(X) - f'(X)]^2$$

where $P^*_X$ is the empirical distribution on $X$ in the researcher’s dataset. When the outcome space $\mathcal{Y}$ consists of probability distributions, a natural choice is the expected Kullback-Liebler divergence between the predictions of $f$ and $f'$, namely

$$d(f, f') = \mathbb{E}_{P_X} [D(f(X) || f'(X))]$$

where $D$ denotes the Kullback-Liebler divergence. Nonstandard choices of $d$ should be explained and justified.

**Computing Restrictiveness.** By assumption that $\mathcal{Y}$ is a subset of finite-dimensional Euclidean space, the uniform distribution on any choice of admissible set $\mathcal{F}$ is well-defined. To compute restrictiveness, the researcher should:

1. Choose a sample size $M \in \mathbb{N}$ (for example, set $M = 1000$).

2. Sample $M$ mappings from the uniform distribution on the admissible set $\mathcal{F}$. Denote each generated mapping by $f_m$.

3. Compute the estimate of restrictiveness as follows:

$$\hat{r}_M = \frac{1}{M} \sum_{m=1}^{M} d(\mathcal{F}_\Theta, f_m) - \frac{1}{M} \sum_{m=1}^{M} d(f_{\text{base}}, f_m)$$

where $d(\mathcal{F}_\Theta, f) \equiv \min_{g \in \mathcal{F}_\Theta} d(g, f)$.

**Computing the Standard Error.** Let $\hat{\sigma}_{\mathcal{F}_\Theta}^2$ be the sample variance of $\{d(\mathcal{F}_\Theta, f_m)\}_{m=1}^{M}$, $\hat{\sigma}_{\{f_{\text{base}}\}}^2$ be the sample variance of $\{d(\{f_{\text{base}}\}, f_m)\}_{m=1}^{M}$, and $\hat{\sigma}_{\mathcal{F}_\Theta, \{f_{\text{base}}\}}$ be the sample covariance
of \{d(\mathcal{F}_\Theta, f_m)\}_{m=1}^M \text{ and } \{d(\{\text{base}\}, f_m)\}_{m=1}^M. \text{ Define}

\hat{\sigma}_r^2 \equiv \frac{\hat{\sigma}_{\mathcal{F}_\Theta}^2 - 2 \cdot \hat{\tau} \cdot \hat{\sigma}_{\mathcal{F}_\Theta\{\text{base}\}} + \hat{\tau}^2 \cdot \hat{\sigma}_{\{\text{base}\}}^2}{\left( \frac{1}{M} \sum_{m=1}^M d(\{\text{base}\}, f_m) \right)^2}

Then,

\frac{\sqrt{M}(\hat{r}_M - r)}{\hat{\sigma}_r} \to \mathcal{N}(0, 1)

so the standard error of \hat{r}_M is \hat{\sigma}_r/\sqrt{M}.

### B.3 Evaluating Completeness

**The Loss Function** \(\ell\). Choose a loss function \(\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+\) where \(\ell(f, (x, y))\) measures how wrong the prediction \(f(x)\) is when the true outcome is \(y\). We leave this specification open to the researcher, but there are natural choices of loss functions to use depending on the prediction problem and the choice of discrepancy \(d\). As we discuss in Appendix E, certain choices of discrepancy \(d\) and loss \(\ell\) are “paired,” and thus are natural to choose with one another. Specifically, when the outcome space \(\mathcal{Y}\) is real-valued and the discrepancy \(d\) is the expected squared distance, then consider choosing

\[\ell(f, (x, y)) = (y - f(x))^2\]

to be squared distance between the prediction and the outcome. When the outcome space \(\mathcal{Y}\) consists of a set of probability distributions and the discrepancy \(d\) is the expected KL divergence, then consider choosing

\[\ell(f, (x, y)) = -\log f(y | x)\]

to be the negative conditional likelihood of observing \(y\) given \(x\).
Computing Completeness. Let the researcher’s data be written as \( \{ Z_i := (X_i, Y_i) \}_{i=1}^N \).

We describe below a \( K \)-fold cross-validated estimator for completeness.

Choose any set of mappings \( \tilde{F} \). Estimate the out-of-sample prediction error of the model as follows:

1. Divide the data \( (Z_1, \ldots, Z_N) \) into \( K \) (approximately) equal-sized groups. To simplify notation, assume that \( J_N = \frac{N}{K} \) is an integer.

2. Let \( k(i) \) denote the group number of observation \( Z_i \). In each \( k \)-th iteration of cross-validation, the \( k \)-th test set consists of all observations belonging to group \( k \), and the \( k \)-th training set consists of all remaining observations.

3. For each group \( k = 1, \ldots, K \), define

\[
\hat{f}^{-k} := \arg \min_{f \in \tilde{F}} \frac{1}{N - J_N} \sum_{k(i) \neq k} l(f, Z_i)
\]

to be the element of \( \tilde{F} \) that minimizes error for prediction of the training data in iteration \( k \). This estimated mapping is used for prediction of the \( k \)-th test set, and

\[
\hat{e}_k := \frac{1}{J_N} \sum_{k(i) = k} l\left( \hat{f}^{-k}, Z_i \right)
\]

is its out-of-sample error.

4. Then,

\[
\hat{e}_{CV} (\tilde{F}) := \frac{1}{K} \sum_{k=1}^{K} \hat{e}_k
\]

is the average out-of-sample error across the \( K \) choices of test set.

Setting \( \tilde{F} \) to be \( \mathcal{F}, \mathcal{F}_\Theta \), or \( \{ f_{\text{base}} \} \), we can compute \( \hat{e}_{CV} (\mathcal{F}) \), \( \hat{e}_{CV} (\mathcal{F}_\Theta) \) and \( \hat{e}_{CV} (f_{\text{base}}) \).
from the data, leading to the following estimator for $\kappa$:

$$\hat{\kappa} = 1 - \frac{\hat{e}_{CV}(\mathcal{F}_\theta) - \hat{e}_{CV}(\mathcal{F})}{\hat{e}_{CV}(\hat{f}_{\text{base}}) - \hat{e}_{CV}(\mathcal{F})}.$$  

**Computing the Standard Error.** For the $k$-th test set, let $f_{\hat{\theta} - k}$ and $\hat{f}^{-k}$ be the estimated mappings from models $\mathcal{F}_\theta$ and $\mathcal{F}$, respectively. The difference in their test errors on observation $Z_i$ is

$$\Delta_{\theta,k}(Z_i) := l(f_{\hat{\theta} - k}, Z_i) - l(\hat{f}^{-k}, Z_i),$$

and the average difference across all observations in test fold $k$ is

$$\bar{\Delta}_{\theta,k} := \frac{1}{J_N} \sum_{k(i) = k} \Delta_{\theta,k}(Z_i).$$

The sample variance of the difference in test errors for the $k$-th fold is

$$\hat{\sigma}^2_{\Delta_{\theta,k}} := \frac{1}{J_N - 1} \sum_{k(i) = k} \left( \Delta_{\theta,k}(Z_i) - \bar{\Delta}_{\theta,k} \right)^2$$

which we then average over the $K$ folds and obtain

$$\hat{\sigma}^2_{\Delta_{\theta}} := \frac{1}{K} \sum_{k=1}^K \hat{\sigma}^2_{\Delta_{\theta,k}}.$$  

Similarly we define $\Delta_{\text{f}_{\text{base}},k}(Z_i) := l(f_{\text{base}}, Z_i) - l(\hat{f}^{-k}, Z_i)$, and correspondingly $\bar{\Delta}_{\text{f}_{\text{base}},k}$, $\hat{\sigma}^2_{\Delta_{\text{f}_{\text{base}},k}}$ and $\hat{\sigma}^2_{\Delta_{\text{f}_{\text{base}}}}$. Lastly, define the covariance estimator by

$$\hat{\sigma}_{\Delta_{\theta}\Delta_{\text{f}_{\text{base}}}} := \frac{1}{K} \sum_{k=1}^K \frac{1}{J_N - 1} \sum_{k(i) = k} \left( \Delta_{\theta,k}(Z_i) - \bar{\Delta}_{\theta,k} \right) \left( \Delta_{\text{f}_{\text{base}},k}(Z_i) - \bar{\Delta}_{\text{f}_{\text{base}},k}(Z_i) \right).$$
Based on $\hat{\sigma}^2_{\Delta \theta}, \hat{\sigma}^2_{\Delta f_{\text{base}}}$ and $\hat{\sigma}_{\Delta \theta \Delta f_{\text{base}}}$, we define the following variance estimator for $\hat{\kappa}$:

$$\hat{\sigma}^2_{\hat{\kappa}} := \frac{\hat{\sigma}^2_{\Delta \theta} - 2\hat{\kappa}\hat{\sigma}_{\Delta \theta \Delta f_{\text{base}}} + \hat{\kappa}^2\hat{\sigma}^2_{\Delta f_{\text{base}}}}{[\hat{e}_{CV}(f_{\text{base}}) - \hat{e}_{CV}(\bar{\mathcal{F}})]^2}$$  \hspace{1cm} (B.1)

so the standard error of $\hat{\kappa}$ is $\hat{\sigma}_{\hat{\kappa}}/\sqrt{N}$.

## C Proof of Proposition 2

### C.1 Preliminary Definitions

We now introduce some definitions and notation that will be useful in the derivation of the asymptotic distribution of the CV-based completeness estimator.

#### C.1.1 Finite-Sample Out-of-Sample Error

Let $\mathbf{Z}_N := (Z_i)_{i=1}^N$ be a random sample of observations in a given data set, and let $Z_{N+1} \sim P$ denote a random variable with the same distribution $P$ that is independent of $\mathbf{Z}_N$. For a given data set $\mathbf{Z}_N$ and a given model $\bar{\mathcal{F}}$, we define the conditional out-of-sample error (given data set $\mathbf{Z}_N$) as

$$e_{\bar{\mathcal{F}}}(\mathbf{Z}_N) := \mathbb{E} \left[ l \left( \hat{f}_{\mathbf{Z}_N}, Z_{N+1} \right) \mid \mathbf{Z}_N \right],$$

where $\hat{f}_{\mathbf{Z}_N} \in \bar{\mathcal{F}}$ is an estimator, or an algorithm, that selects a mapping $\hat{f}_{\mathbf{Z}_N}$ within the model $\bar{\mathcal{F}}$ based on data $\mathbf{Z}_N$. We also define the out-of-sample error, with expectation taken over different possible data sets $\mathbf{Z}_N$, as $e_{\bar{\mathcal{F}},N} := \mathbb{E} [e_{\bar{\mathcal{F}}}(\mathbf{Z}_N)]$.

From the definition of the K-fold cross-validation estimator, it can be shown that $\mathbb{E} \left[ \hat{e}_{CV}(\bar{\mathcal{F}}) \right] = e_{\bar{\mathcal{F}}, \frac{K-1}{K}N}$. The asymptotic distribution of $\hat{e}_{CV}(\bar{\mathcal{F}}) - e_{\bar{\mathcal{F}}, \frac{K-1}{K}N}$ has been studied in the statistics and machine learning literature. Our analysis below will be based on the results in Austern and Zhou (2020) on the asymptotic distribution of $\hat{e}_{CV}(\bar{\mathcal{F}}) - e_{\bar{\mathcal{F}}, \frac{K-1}{K}N}$.
C.1.2 Joint Parametrization of $F_\Theta$ and $\overline{F}$

Recall that the model $F_\Theta$ is parametrized by $\theta \in \Theta$, and $f_\theta$ denotes a generic function in $F_\Theta$. Motivated by the applications in this paper, we assume that the finite-dimensional parameter $\beta \in B \subseteq \mathbb{R}^d$ and use the notation $f[\beta] \in \overline{F}$ to denote a generic function in $\overline{F}$. Since by assumption $f^* \in \overline{F}$, we can define a parameter $\beta^*$ to represent it, i.e. $f[\beta^*] = f^*$.

For arbitrary $\theta$ and $\beta$, write

$$l(\theta, Z_i) := l(f_\theta, Z_i)$$

and

$$l_B(\beta, Z_i) := l(f[\beta], Z_i)$$

We define the estimation mappings by

$$\hat{\theta}(Z_N) := \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} l(\theta, Z_i)$$

and

$$\hat{\beta}(Z_N) := \arg\min_{\beta \in B} \frac{1}{N} \sum_{i=1}^{N} l_B(\beta, Z_i).$$

Let $\alpha := (\theta', \beta')'$ denote the concatenation of the parameters $\theta \in \Theta$ and $\beta \in B$, $\alpha^* := (\theta^*, \beta^*)'$ to be the parameters associated with the best mappings in $F_\Theta$ and $\overline{F}$, and also define

$$\hat{\alpha}(Z_N) := (\hat{\theta}(Z_N), \hat{\beta}(Z_N))' = \arg\min_{\theta \in \Theta, \beta \in B} \frac{1}{N} \sum_{i=1}^{N} [l(\theta, Z_i) + l_B(\beta, Z_i)]$$

to be an estimator for $\alpha^*$. Finally, define

$$\Delta l(\theta, \beta; Z_i) := l(f_\theta, Z_i) - l(f[\beta], Z_i) = l(\theta, Z_i) - l_B(\beta, Z_i).$$

C.2 Construction of Variance Estimator

To obtain the standard error of the estimate, we use a variance estimator adapted from Proposition 1 in Austern and Zhou (2020). Specifically, for the $k$-th test set, let $f_{\hat{\theta}^{-k}}$ and $\hat{f}^{-k}$ be the estimated mappings from models $F_\Theta$ and $\overline{F}$, respectively. The difference in their test errors on observation $Z_i$ is

$$\Delta_{\theta,k}(Z_i) := l(f_{\hat{\theta}^{-k}}, Z_i) - l(\hat{f}^{-k}, Z_i),$$

and the average difference across all observations in test fold $k$ is

$$\overline{\Delta_{\theta,k}} := \frac{1}{N} \sum_{k(i)=k} \Delta_{\theta,k}(Z_i).$$

The sample variance of the difference in test errors for the $k$-th fold is

$$\hat{\sigma}^2_{\Delta_{\theta,k}} := \frac{1}{J_N - 1} \sum_{k(i)=k} (\Delta_{\theta,k}(Z_i) - \overline{\Delta_{\theta,k}})^2.$$
which we average over the $K$ folds and obtain $\hat{\sigma}^2_{\Delta_\theta} := \frac{1}{K} \sum_{k=1}^{K} \hat{\sigma}^2_{\Delta_\theta,k}$.

Similarly, we define $\Delta_{f_{\text{base}},k}(Z_i) := l(f_{\text{base}}, Z_i) - l(f_k, Z_i)$, and correspondingly $\bar{\Delta}_{f_{\text{base}},k}$, $\hat{\sigma}^2_{\Delta_{f_{\text{base}},k}}$, and $\hat{\sigma}^2_{\Delta_{f_{\text{base}}}}$. Lastly, define the covariance estimator by

$$\hat{\sigma}_{\Delta_\theta \Delta_{f_{\text{base}}}} := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{J_N - 1} \sum_{(i)=k} \left( \Delta_{\theta,k}(Z_i) - \bar{\Delta}_{\theta,k} \right) \left( \Delta_{f_{\text{base}},k}(Z_i) - \bar{\Delta}_{f_{\text{base}},k}(Z_i) \right).$$

Based on $\hat{\sigma}^2_{\Delta_\theta}$, $\hat{\sigma}^2_{\Delta_{f_{\text{base}}}}$, and $\hat{\sigma}_{\Delta_\theta \Delta_{f_{\text{base}}}}$, we define the following variance estimator for $\hat{\kappa}$:

$$\hat{\sigma}^2_{\hat{\kappa}} := \frac{\hat{\sigma}^2_{\Delta_\theta} - 2\hat{\kappa}\hat{\sigma}_{\Delta_{f_{\text{base}}}} + \hat{\kappa}^2\hat{\sigma}^2_{\Delta_{f_{\text{base}}}}}{[\hat{e}_{\text{CV}}(f_{\text{base}}) - \hat{e}_{\text{CV}}(F)]^2}. \quad (C.1)$$

### C.3 Material Based on Austern and Zhou (2020)

**Assumption 3** (Conditions for Asymptotics of CV Estimator).

1. $l_\Theta(\theta, z)$ and $l_B(\beta, z)$ are twice differentiable and strictly convex in $\theta$ and $\beta$.

2. $\mathbb{E} \left[ \sup_{\theta \in \Theta} l_\Theta^4(\theta, Z_i) \right] < \infty$ and $\mathbb{E} \left[ \sup_{\beta \in B} l_B^4(\beta, Z_i) \right] < \infty$.

3. There exist open neighborhoods $O_{\theta^*}$ and $O_{\beta^*}$ of $\theta^*$ and $\beta^*$ in $\Theta$ and $B$ such that

   (a) $\mathbb{E} \left[ \sup_{\theta \in O_{\theta^*}} \|\nabla_\theta l_\Theta(\theta, Z_i)\|^{16} \right] < \infty$, $\mathbb{E} \left[ \sup_{\beta \in O_{\beta^*}} \|\nabla_\beta l_B(\beta, Z_i)\|^{16} \right] < \infty$.

   (b) $\mathbb{E} \left[ \sup_{\theta \in O_{\theta^*}} \|\nabla_\theta^2 l_\Theta(\theta, Z_i)\|^{16} \right] < \infty$, $\mathbb{E} \left[ \sup_{\beta \in O_{\beta^*}} \|\nabla_\beta^2 l_B(\beta, Z_i)\|^{16} \right] < \infty$.

   (c) there exists $c > 0$ such that $\lambda_{\text{min}}(\nabla_\theta^2 l_\Theta(\theta, Z_i)) \geq c$, $\lambda_{\text{min}}(\nabla_\beta^2 l_B(\beta, Z_i)) \geq c$ a.s. uniformly on $O_{\theta^*}$ and $O_{\beta^*}$.

**Lemma C.1.** Under Assumption 3:

$$\sqrt{N} \left[ \hat{e}_{\text{CV}}(F_\theta) - \hat{e}_{\text{CV}}(F) - \left( e_{F_\theta, \frac{k-1}{K}N} - e_{F, \frac{k-1}{K}N} \right) \right] \xrightarrow{d} \mathcal{N}(0, \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i))).$$

**Proof.** Proposition 5 of Austern and Zhou (2020) establishes the asymptotic normality of cross-validation risk estimator and its asymptotic variance under parametric settings where
the loss function used for training is the same as the loss function used for evaluation. Applying Proposition 5 of Austern and Zhou (2020) under Assumption 3 to $\theta, \beta$ and $\alpha = (\theta, \beta)$, we obtain:

$$
\sqrt{N} \left( \hat{e}_{CV}(F_{\theta}) - e_{F_{\theta}, \frac{k-1}{k} N} \right) \stackrel{d}{\rightarrow} \mathcal{N}(0, \text{Var} \left( l(f_{\theta^*}, Z_i) \right)),
$$

$$
\sqrt{N} \left( \hat{e}_{CV}(F) - e_{F, \frac{k-1}{k} N} \right) \stackrel{d}{\rightarrow} \mathcal{N}(0, \text{Var} \left( l(f^*, Z_i) \right)),
$$

$$
\sqrt{N} \left( \hat{e}_{CV}(F_{\theta}) + \hat{e}_{CV}(F) - e_{F_{\theta}, \frac{k-1}{k} N} - e_{F, \frac{k-1}{k} N} \right) \stackrel{d}{\rightarrow} \mathcal{N}(0, \text{Var} \left( l(f_{\theta^*}, Z_i) + l(f^*, Z_i) \right)).
$$

Using the equality $\text{Var} (X + Y) + \text{Var} (X - Y) = 2 \text{Var} (X) + 2 \text{Var} (Y)$, we then deduce that

$$
\sqrt{N} \left[ \hat{e}_{CV}(F_{\theta}) - \hat{e}_{CV}(F) - \left( e_{F_{\theta}, \frac{k-1}{k} N} - e_{F, \frac{k-1}{k} N} \right) \right] \stackrel{d}{\rightarrow} \mathcal{N}(0, \text{Var} \left( \Delta l(f_{\theta^*}, f^*; Z_i) \right)).
$$

\[\square\]

**Lemma C.2** (Application of Proposition 1 of Austern and Zhou, 2020). Under Assumption 3, $\hat{\sigma}_\Delta^2 \xrightarrow{p} \text{Var} (\Delta l(f_{\theta^*}, f^*; Z_i))$.

**Proof.** Applying Proposition 1 of Austern and Zhou (2020) under Assumption 3 to $\theta, \beta$ and $\alpha = (\theta, \beta)$:

$$
\hat{\sigma}_{F_{\theta}}^2 := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{J_{N} - 1} \sum_{k(i)=k} \left( l(f_{\hat{\theta}_k}, Z_i) - \frac{1}{J_{N}} \sum_{k(j)=k} l(f_{\hat{\theta}_k}, Z_j) \right)^2 \xrightarrow{p} \text{Var} (l(f_{\theta^*}, Z_i)),
$$

$$
\hat{\sigma}_{F}^2 := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{J_{N} - 1} \sum_{k(i)=k} \left( l(f_{\hat{\beta}_k}, Z_i) - \frac{1}{J_{N}} \sum_{k(j)=k} l(f_{\hat{\beta}_k}, Z_j) \right)^2 \xrightarrow{p} \text{Var} (l(f^*, Z_i)),
$$

and

$$
\hat{\sigma}_{F_{\theta}+F}^2
$$
\[ := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{J_N - 1} \sum_{k(i)=k} \left( l(f_{\theta - k}, Z_i) + l(f_{[\tilde{\beta} - k]}, Z_i) - \frac{1}{J_N} \sum_{k(j)=k} \left[ l\left( f_{[\tilde{\beta} - k]}, Z_j \right) + l(f_{\theta - k}, Z_i) \right] \right)^2 \]

\[ \rightarrow \text{Var} \left( l(f_{\theta^*}, Z_i) + l(f^*, Z_i) \right). \]

Hence, \[ \hat{\sigma}_\Delta^2 = 2\hat{\sigma}_F^2 + 2\hat{\sigma}_\bar{F}^2 \rightarrow \text{Var} \left( l(f_{\theta^*}, Z_i) \right) + 2\text{Var} \left( l(f^*, Z_i) \right) - 2\text{Var} \left( l(f_{\theta^*}, Z_i) + l(f^*, Z_i) \right) = \text{Var} \left( \Delta l(f_{\theta^*}, f^*; Z_i) \right). \]

\section*{C.4 Finishing the Proof}

Lemma C.1 characterizes the limit distribution of

\[ \sqrt{N} \left[ \hat{e}_{\text{CV}} (\mathcal{F}_\theta) - \hat{e}_{\text{CV}} (\bar{\mathcal{F}}) - \left( e_{\mathcal{F}_\theta, \frac{K-1}{K} N} - e_{\bar{\mathcal{F}}, \frac{K-1}{K} N} \right) \right] \]

which we show is also the limit distribution of \[ \sqrt{N} \left[ \hat{e}_{\text{CV}} (\mathcal{F}_\theta) - \hat{e}_{\text{CV}} (\bar{\mathcal{F}}) - (e_{\mathcal{F}_\theta} - e_{\bar{\mathcal{F}}}) \right]. \]

To see this, notice that

\[ e_{\mathcal{F}_\theta, \frac{K-1}{K} N} - e_{\mathcal{F}_\theta} = \mathbb{E} \left[ l_{\theta} \left( \hat{\theta}^{(i)}, Z_i \right) - l_{\theta} (\theta^*, Z_i) \right] \]

\[ = \mathbb{E} \left[ \nabla l_{\theta} (\theta^*, Z_i) \cdot \left( \hat{\theta}^{(i)} - \theta^* \right) + \left( \hat{\theta}^{(i)} - \theta^* \right) \cdot \nabla^2 l_{\theta} \left( \hat{\theta}, Z_i \right) \cdot \left( \hat{\theta}^{(i)} - \theta^* \right) \right] \]

\[ = 0 + \mathbb{E} \left[ \left( \hat{\theta}^{(i)} - \theta^* \right) \cdot \nabla^2 l_{\theta} \left( \hat{\theta}, Z_i \right) \cdot \left( \hat{\theta}^{(i)} - \theta^* \right) \right] \]

\[ = \frac{1}{N - J_N} \mathbb{E} \left[ \sqrt{N - J_N} \left( \hat{\theta}^{(i)} - \theta^* \right) \cdot \nabla^2 l_{\theta} \left( \hat{\theta}, Z_i \right) \cdot \sqrt{N - J_N} \left( \hat{\theta}^{(i)} - \theta^* \right) \right] \]

\[ = c \frac{1}{N - J_N} + o \left( \frac{1}{N - J_N} \right) = c \frac{K}{K - 1} \cdot \frac{1}{N} + o \left( \frac{1}{N} \right) \]

since \( J_N = N/K \). Therefore \( \sqrt{N} \left( e_{\mathcal{F}_\theta, \frac{K-1}{K} N} - e_{\mathcal{F}_\theta} \right) = o_p (1) \), and \( \sqrt{N} \left( e_{\bar{\mathcal{F}}, \frac{K-1}{K} N} - e_{\bar{\mathcal{F}}} \right) = o_p (1) \). Hence: \( \sqrt{N} \left[ \hat{e}_{\text{CV}} (\mathcal{F}_\theta) - \hat{e}_{\text{CV}} (\bar{\mathcal{F}}) - (e_{\mathcal{F}_\theta} - e_{\bar{\mathcal{F}}}) \right] \overset{d}{\rightarrow} \mathcal{N} \left( 0, \text{Var} \left( \Delta l(f_{\theta^*}, f^*; Z_i) \right) \right). \]

Now, we replicate the previous result with \( f_{\text{base}} \) in place of \( \mathcal{F}_\theta \) and obtain

\[ \sqrt{N} \left[ \hat{e}_{\text{CV}} (f_{\text{base}}) - \hat{e}_{\text{CV}} (\bar{\mathcal{F}}) - (e_{f_{\text{base}}} - e_{\bar{\mathcal{F}}}) \right] \overset{d}{\rightarrow} \mathcal{N} \left( 0, \text{Var} \left( \Delta l(f_{\text{base}}, f^*; Z_i) \right) \right). \]
and jointly
\[
\sqrt{N} \left( \begin{pmatrix} \hat{e}_{CV}(\mathcal{F}_\Theta) - \hat{e}_{CV}(\mathcal{F}) - (e_{\mathcal{F}_\Theta} - e_{\mathcal{F}}) \\ \hat{e}_{CV}(f_{base}) - \hat{e}_{CV}(\mathcal{F}) - (e_{f_{base}} - e_{\mathcal{F}}) \end{pmatrix} \right) \xrightarrow{d} \mathcal{N} \left( \begin{pmatrix} \sigma_{\Delta \theta}^2 & \sigma_{\Delta \theta \Delta f_{base}} \\ \sigma_{\Delta \theta \Delta f_{base}} & \sigma_{\Delta f_{base}}^2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right)
\]

with \( \sigma_{\Delta \theta}^2 := \text{Var}(\Delta l(f_{\theta^*}, f_*; Z_i)) \), \( \sigma_{\Delta f_{base}}^2 := \text{Var}(\Delta l(f_{base}, f_*; Z_i)) \), and \( \sigma_{\Delta \theta \Delta f_{base}} := \text{Cov}(\Delta l(f_{\theta^*}, f_*; Z_i), \Delta l(f_{base}, f_*; Z_i)) \).

By Lemma C.2, Assumption 2 and the Delta Method, we have
\[
\sqrt{N} (\hat{\kappa} - \kappa) \xrightarrow{d} \mathcal{N} \left( 0, \frac{\sigma_{\Delta \theta}^2 - 2\kappa \sigma_{\Delta \theta \Delta f_{base}} + \kappa^2 \sigma_{\Delta f_{base}}^2}{d^2(f_{base}, f^*)} \right).
\]

Since
\[
\hat{\sigma}_\kappa \xrightarrow{p} \frac{\sigma_{\Delta \theta}^2 - 2\kappa \sigma_{\Delta \theta \Delta f_{base}} + \kappa^2 \sigma_{\Delta f_{base}}^2}{d^2(f_{base}, f^*)},
\]
we have
\[
\frac{\sqrt{N} (\hat{\kappa} - \kappa)}{\hat{\sigma}_\kappa} \xrightarrow{d} \mathcal{N}(0, 1).
\]
D Supplementary Material to Application 1

D.1 Estimates for Application 1

Table 5: Restrictiveness and Completeness for Certainty Equivalents

<table>
<thead>
<tr>
<th></th>
<th># Param</th>
<th>Restrictiveness</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPT Specifications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha, \delta, \gamma$</td>
<td>3</td>
<td>0.28 (0.003)</td>
<td>0.95 (0.02)</td>
</tr>
<tr>
<td>$\delta, \gamma$</td>
<td>2</td>
<td>0.37 (0.004)</td>
<td>0.95 (0.02)</td>
</tr>
<tr>
<td>$\alpha, \gamma$</td>
<td>2</td>
<td>0.51 (0.006)</td>
<td>0.95 (0.02)</td>
</tr>
<tr>
<td>$\alpha, \delta$</td>
<td>2</td>
<td>0.49 (0.005)</td>
<td>0.27 (0.05)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1</td>
<td>0.91 (0.005)</td>
<td>0.25 (0.05)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>1</td>
<td>0.68 (0.009)</td>
<td>0.26 (0.06)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>1</td>
<td>0.59 (0.006)</td>
<td>0.71 (0.06)</td>
</tr>
<tr>
<td><strong>DA Specifications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha, \eta$</td>
<td>2</td>
<td>0.47 (0.006)</td>
<td>0.27 (0.06)</td>
</tr>
<tr>
<td>$\eta$</td>
<td>1</td>
<td>0.69 (0.009)</td>
<td>0.27 (0.05)</td>
</tr>
</tbody>
</table>

Restrictiveness is estimated from 1000 simulations and we report the analytic standard errors. Because of potential dependence among the reported certainty equivalents of subjects, we compute the standard errors for completeness using a block bootstrapping procedure that clusters together all observations from the same subject.\(^{46}\) We then carry out our (cross-validated) estimation of completeness on each bootstrap sample, and compute the standard errors based on 1000 bootstrap samples. These bootstrapped standard errors are similar to the analytic standard errors we get under a revision of the formulas in Section 5 to accommodate clustering on subjects (see the following section).

\(^{46}\)When generating a bootstrap sample, we randomly sample the 179 subjects with replacement, and include all the reported certainty equivalents of the drawn subjects with replacement.
D.2 Analytical SE with Clustering

We discuss here an alternative method for calculating clustered standard errors for completeness.

We consider each subject’s reported certainty equivalents for the 25 lotteries as a 25-dimensional vector. We assume that this 25-dimensional vector is i.i.d. across subjects, but leave the dependence within this subject-specific vector unrestricted. Specifically, define the feature space $\mathcal{X}$ to be a singleton consisting of the $25 \times 3$ matrix whose rows are the different lottery tuples $(z, \bar{z}, p)$ in the Bruhin et al. (2010) data. The outcome space is $\mathcal{Y} = \mathbb{R}^{25}$, where a typical element is a vector of 25 certainty equivalents for the 25 lotteries. The expected certainty equivalent vector over subjects is represented by a mapping $f : \mathcal{X} \rightarrow \mathbb{R}^{25}$, which is simply a vector in $\mathbb{R}^{25}$.

Finally, let the loss function $l$ be

$$l(f, Y_i, X) := \frac{1}{25} \|Y_i - f_\theta(X)\|^2 = \frac{1}{25} \sum_{h=1}^{25} (Y_{i,h} - f_h)^2.$$ 

This loss function groups together the squared losses of each individual subject across the 25 lotteries. Under this setup, the analytical formula for standard errors provided in Section 5 and Appendix C.2 can be directly applied, with sample size $N = 179$. Table D.2 reports the standard errors for completeness computed in this way.
### CPT Specifications

<table>
<thead>
<tr>
<th>Param Set</th>
<th># Parameters</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha, \delta, \gamma$</td>
<td>3</td>
<td>0.95 (0.09)</td>
</tr>
<tr>
<td>$\delta, \gamma$</td>
<td>2</td>
<td>0.95 (0.08)</td>
</tr>
<tr>
<td>$\alpha, \gamma$</td>
<td>2</td>
<td>0.95 (0.09)</td>
</tr>
<tr>
<td>$\alpha, \delta$</td>
<td>2</td>
<td>0.27 (0.09)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1</td>
<td>0.25 (0.05)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>1</td>
<td>0.26 (0.06)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>1</td>
<td>0.71 (0.06)</td>
</tr>
</tbody>
</table>

### DA Specifications

<table>
<thead>
<tr>
<th>Param Set</th>
<th># Parameters</th>
<th>Completeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha, \eta$</td>
<td>2</td>
<td>0.27 (0.06)</td>
</tr>
<tr>
<td>$\eta$</td>
<td>1</td>
<td>0.27 (0.05)</td>
</tr>
</tbody>
</table>

### D.3 Restrictiveness on Alternative Sets of Lotteries

We report here the restrictiveness values used to construct the CDFs in Figure 4 as well as the papers the corresponding sets of lotteries were derived from, and the number of lotteries from each paper.

#### Table 6: Restrictiveness

<table>
<thead>
<tr>
<th>Source Paper</th>
<th># Lotteries</th>
<th>CPT($\alpha, \delta, \gamma$)</th>
<th>DA($\alpha, \eta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdellaoui et al. (2015)</td>
<td>3</td>
<td>0.04 (0.00)</td>
<td>0.31 (0.01)</td>
</tr>
<tr>
<td>Murad et al. (2016)</td>
<td>25</td>
<td>0.25 (0.00)</td>
<td>0.38 (0.00)</td>
</tr>
<tr>
<td>Sutter et al. (2013)</td>
<td>4</td>
<td>0.46 (0.01)</td>
<td>0.46 (0.01)</td>
</tr>
<tr>
<td>Fan et al. (2019)</td>
<td>19</td>
<td>0.23 (0.00)</td>
<td>0.25 (0.00)</td>
</tr>
<tr>
<td>Bernheim and Sprenger (2020a)</td>
<td>7</td>
<td>0.13 (0.00)</td>
<td>0.45 (0.01)</td>
</tr>
</tbody>
</table>
E “Pairing” Completeness and Restrictiveness

In this section, we show that completeness and restrictiveness are related via the equation

\[ \kappa(\mathcal{F}_\Theta) = 1 - r(\mathcal{F}_\Theta, \{f^*\}), \]  

(E.1)

when the loss function \( l \) used to define \( e_P \), and the discrepancy function \( d \) used to define \( r \), are “paired” in a coherent way, which we now explain.

We first provide more details about the formulation of completeness. Suppose that besides \( X \), there is a random outcome \( Z \). We will consider hypothetical joint distributions \( \tilde{P} \) with different conditional distribution \( \tilde{P}_{Z|X} \), where the marginal distribution \( \tilde{P}_X \) is held fixed. The analyst wants to learn a statistic of the conditional distribution of \( Z \) given \( X \), which we denote by \( Y \in \mathcal{Y} \). Two leading cases of this problem are: (a) prediction of the conditional expectation \( \mathbb{E}_{\tilde{P}}[Z|X] \), and (b) prediction of the conditional distribution \( \tilde{P}_{Z|X} \) itself. As in the main text, a prediction is any function \( f : \mathcal{X} \to \mathcal{Y} \), and we define \( \mathcal{F} \) to be the set of all such mappings.

Let \( l : \mathcal{F} \times \mathcal{X} \times \mathcal{Z} \to \mathbb{R} \) be a loss function, where \( l(f, (x, z)) \) is the loss assigned to predicting \( f(x) \) when the realized outcome is \( z \). We define the expected error of a prediction rule \( f \) with respect to the distribution \( \tilde{P} \) by

\[ e_{\tilde{P}}(f) := \mathbb{E}_{\tilde{P}}[l(f, (X, Z))], \]  

(E.2)

and let \( f^*_{\tilde{P}} \) denote the prediction rule that minimizes the expected error under \( \tilde{P} \):

\[ f^*_{\tilde{P}} := \min_{f \in \mathcal{F}} e_{\tilde{P}}(f). \]

As in the main text, \( P \) denotes the distribution from which real data is generated. Then the
completeness of a model $F_\Theta$ as defined in Fudenberg et al. (2022) can be written as

$$\kappa(F_\Theta) = \frac{e_P(f_{\text{base}}) - e_P(F_\Theta)}{e_P(f_{\text{base}}) - e_P(F)} \equiv 1 - \frac{e_P(F_\Theta) - e_P(F)}{e_P(f_{\text{base}}) - e_P(F)}.$$ 

We now formally define the meaning of “pairing” between the discrepancy function $d$ and the loss function $l$.

**Definition E.1.** The loss function $l$ and discrepancy $d : F \times F \to \mathbb{R}$ are paired if

$$d(f, f^*_{\bar{P}}) = e_{\bar{P}}(f) - e_{\bar{P}}(f^*_{\bar{P}})$$

(E.3)

for every distribution $\bar{P} \in \Delta(\mathcal{X} \times \mathcal{Z})$ whose marginal distribution on $\mathcal{X}$ is $P_X$. That is, $d(f, f^*_{\bar{P}})$ is the difference between the error of prediction rule $f$ and the error of the best prediction rule $f^*_{\bar{P}}$.\textsuperscript{47}

As noted in the main text, if $l$ and $d$ are paired, then (E.1) holds, where $f^* = f^*_{\bar{P}}$. Moreover, as also noted in the main text, the following functions are paired:

- Let $\mathcal{Y} = \mathbb{R}$. Then squared loss $l(f, (x, z)) := (z - f(x))^2$ and the squared distance discrepancy $d_{\text{MSE}}(f, g) := \mathbb{E}_{P_X}[(f(X) - g(X))^2]$ are paired.

- Let $\mathcal{Y}$ be the set of distributions over a finite set $\mathcal{Z}$. Then negative (conditional) log-likelihood $l(f, (x, z)) := -\log f(z|x)$ and the KL-divergence discrepancy

$$d_{KL}(f, g) := \mathbb{E}_{P_X} \left[ \sum_{z \in \mathcal{Z}} g(z|x) \left( \log g(z|x) - \log f(z|x) \right) \right]$$

are paired.

\textsuperscript{47}This relation resembles but differs from the coupling of the “cost of uncertainty” and the “value of information” in Frankel and Kamenica (2019), which concerns comparisons of different signal structures, as opposed to comparing model classes.
A Loss Function That Cannot be Paired with any Discrepancy

When $Y$ is the set of distributions on $Z$, then every loss function $l$ has a paired discrepancy function, since we can define $d(f, f_{\tilde{P}}) := e_{f_{\tilde{P}}}(f) - e_{f_{\tilde{P}}}(f_{\tilde{P}})$\footnote{This is because $\tilde{P}$ is completely pinned down by $f_{\tilde{P}}$ given $P_X$, so $e_{\tilde{P}} = e_{f_{\tilde{P}}}$.}. But in general, for some prediction problems and loss functions $l$, there may not exist a discrepancy $d$ such that $l$ and $d$ are paired, as the next example shows. In these cases, we can still evaluate restrictiveness and completeness, but they will not have an evident relationship.

Consider a setting where $X$ is degenerate, i.e., $X$ is a singleton, so that the joint distribution $\tilde{P}$ is completely characterized by the distribution of $Y$. Furthermore, let $Y := [0, 1]$. If $f^* := \text{med}(Y) \in Y = [0, 1]$, then a mapping $f : X \to S$ is just a number in $[0, 1]$. When the loss function is the absolute deviation $l(f, y) := |y - f|$, and the error function is mean absolute deviation $e_{\tilde{P}}(f) := E_{\tilde{P}}[|Y - f|]$, the true median $f^*$ minimizes the error, i.e. $f^* \in \arg \min_{f \in [0, 1]} e_{\tilde{P}}(f)$. However, it is not true that $|f - f^*| = e_{\tilde{P}}(f) - e_{\tilde{P}}(f^*)$ for any $f \in [0, 1]$. To see this, suppose that $Y \sim U[0, 1]$ under $\tilde{P}$. Then $f^* = 0.5$ and $e_{\tilde{P}}(f^*) = 0.25$. However, for $f = 0.4$, we have $e_{\tilde{P}}(f) = 0.26$. but $|f - f^*| = 0.1 \neq 0.01 = e_{\tilde{P}}(f) - e_{\tilde{P}}(f^*)$.

Moreover, there is no function $d : [0, 1]^2 \to [0, 1]$ such that decomposability (E.3) holds, which would require that $d(f, \tilde{f}_{\tilde{P}}) = e_{\tilde{P}}(f) - e_{\tilde{P}}(\tilde{f}_{\tilde{P}})$ for any distribution $P$ of $Y$ supported on $[0, 1]$. To see this, suppose that $Y \sim U[0, 1]$ under $\tilde{P}_1$, we have

$$e_{\tilde{P}_1}(f) - e_{\tilde{P}_1}(\tilde{f}_{\tilde{P}_1}) = (f - 0.5)^2 = (f - \tilde{f}_{\tilde{P}_1})^2, \quad \forall f \in [0, 1].$$

However, supposing that, under $\tilde{P}_2$, the probability density function of $Y$ is given by $2y$ for $y \in [0, 1]$, we have $f_{\tilde{P}_2} = \sqrt{2}/2$ and $e_{\tilde{P}_2}(f_{\tilde{P}_2}) = (2 - \sqrt{2})/3$ but

$$e_{\tilde{P}_2}(f) - e_{\tilde{P}_2}(f_{\tilde{P}_2}) = \frac{1}{3} \left(2f^3 - 3f^2 + \sqrt{2}\right) \neq (f - f_{\tilde{P}_2})^2.$$
How Different are Two Models?

The restrictiveness of parametric models does not tell us whether the two models rule out similar kinds of behaviors. For example, consider DA(ν)—which achieves a restrictiveness of 0.69—and CPT(γ)—which achieves a restrictiveness of 0.59. Despite imposing different functional forms, these models may essentially capture the same risk behaviors, leading to their similar absolute levels of restrictiveness. Another possibility is that the two models embody rather different restrictions, so that mappings which are well approximated by DA(ν) are poorly approximated by CPT(γ), and vice versa. We next provide a measure for determining whether two models are restrictive “in the same way.”

Consider two parametric models $\mathcal{F}_{\Theta_1}$ and $\mathcal{F}_{\Theta_2}$, where we assume that $f_{\text{base}} \in \mathcal{F}_{\Theta_1}, \mathcal{F}_{\Theta_2}$.

For an arbitrary mapping $f$, define

$$
\delta_1^f := \frac{d(\mathcal{F}_{\Theta_1}, f)}{d(f_{\text{base}}, f)} \quad \delta_2^f := \frac{d(\mathcal{F}_{\Theta_2}, f)}{d(f_{\text{base}}, f)}
$$

to be the normalized discrepancy between the model and $f$.

**Definition F.1.** The $\delta$-correlation between models $\mathcal{F}_{\Theta_1}$ and $\mathcal{F}_{\Theta_2}$ is the correlation coefficient for the pair $(\delta_1^f, \delta_2^f)$ where $f$ follows a uniform distribution on the admissible set $\mathcal{F}$.

Two models with a high $\delta$-correlation do relatively well on the same mappings, while the $\delta$-correlation is negative if the mappings that one model approximates well are relatively harder for the other to approximate.

The size of $\delta$-correlation between two models does not directly imply anything about their absolute levels of restrictiveness.\(^{49}\) The size of $\delta$-correlation also does not tell us which model is more restrictive. If two models perform better on the same mappings, but one model fits all mappings better than the other, the $\delta$-correlation measure will not reveal which of the two models is more flexible. The measure of $\delta$-correlation, however, can be usefully paired

\(^{49}\)For example, the models $\mathcal{F}_{\Theta_1} \equiv \mathcal{F}$ and $\mathcal{F}_{\Theta_2} \equiv \mathcal{F}$ have a $\delta$-correlation of 1, as do the models $\mathcal{F}_{\Theta_1} \equiv \{f_{\text{base}}\}$ and $\mathcal{F}_{\Theta_2} \equiv \{f_{\text{base}}\}$. But the first pair of models is maximally unrestrictive while the second is maximally restrictive.
with the restrictiveness of the two models to provide further insight into their comparison, as we now demonstrate.

F.1 Certainty Equivalents

Table 7 reports the $\delta$-correlation between the models CPT-$\gamma$, CPT-$(\delta,\gamma)$, CPT-$(\alpha,\delta,\gamma)$, DA($\eta$), and DA($\alpha,\eta$).

<table>
<thead>
<tr>
<th></th>
<th>CPT($\gamma$)</th>
<th>CPT($\delta,\gamma$)</th>
<th>CPT($\alpha,\delta,\gamma$)</th>
<th>DA($\eta$)</th>
<th>DA($\alpha,\eta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPT($\gamma$)</td>
<td>1</td>
<td>0.38</td>
<td>-0.76</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>CPT($\delta,\gamma$)</td>
<td>-</td>
<td>1</td>
<td>0.37</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>CPT-$(\alpha,\delta,\gamma)$</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.48</td>
<td>0.40</td>
</tr>
<tr>
<td>DA($\eta$)</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>DA($\alpha,\eta$)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7: $\delta$-correlation between various pairs of models

CPT($\delta,\gamma$) and CPT($\alpha,\delta,\gamma$) are highly correlated. Since the two models have similar absolute levels of restrictiveness ($r = 0.28$ for CPT($\alpha,\delta,\gamma$) and $r = 0.37$ for CPT($\delta,\gamma$)), this suggests that the two models rule out very similar behaviors.

The two models DA($\eta$) and CPT($\gamma$) also have similar absolute levels of restrictiveness ($r = 0.69$ for DA($\eta$) and $r = 0.59$ for CPT($\gamma$)). But their $\delta$-correlation turns out to be quite negative, suggesting that the two models perform relatively well on different mappings. The models are thus different in empirical content and not simply in the statement of their functional forms. Interestingly, the gap in restrictiveness between CPT($\gamma$) and DA($\alpha,\eta$) is not substantially larger, but the $\delta$-correlation between these models rises to 0.40, suggesting that introduction of the $\alpha$ parameter in addition to the $\eta$ parameter in DA re-directs the model’s predictions in the direction of CPT($\gamma$).

The remaining $\delta$-correlations are all positive but not large, suggesting that there are substantial differences between the models. The imperfect correlation is not surprising, since these model pairs are differentiated in both restrictiveness and completeness.
F.2 Initial Play

Table 8 compares the δ-correlation between models PCHM, Logit Level-1, and Logit PCHM.

<table>
<thead>
<tr>
<th></th>
<th>PCHM</th>
<th>Logit Level-1</th>
<th>Logit PCHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCHM</td>
<td>1</td>
<td>0.67</td>
<td>0.77</td>
</tr>
<tr>
<td>Logit Level-1</td>
<td>-</td>
<td>1</td>
<td>0.94</td>
</tr>
<tr>
<td>Logit PCHM</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8: Correlation between errors of the two models

The δ-correlation between Logit PCHM and Logit Level-1 is close to 1, so the distributions that these models fit relatively better and relatively worse are very similar. Since the absolute levels of restrictiveness for the two models are not statistically different, the near-perfect correlation in errors suggests that the two models have approximately the same empirical content. In contrast, PCHM—which is less complete and more restrictive than both Logit Level-1 and Logit PCHM—has a lower δ-correlation with each of these models, although PCHM’s δ-correlation with Logit PCHM is slightly higher. This reflects the fact that the predictions of PCHM are more similar to the predictions of Logit PCHM than to the predictions of Logit Level-1.