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

Drew Fudenberg[†] Wayne Gao [‡] Annie Liang[§]

November 7, 2020

Abstract

We propose a new way to quantify the restrictiveness of an economic model, based on how well the model fits simulated, hypothetical data sets. The data sets are drawn at random from a distribution that satisfies some application-dependent content restrictions (such as that people prefer more money to less). Models that can fit almost all hypothetical data well are not restrictive. To illustrate our approach, we evaluate the restrictiveness of popular behavioral models in two experimental settings—certainty equivalents and initial play—and explain how restrictiveness reveals new insights about each of the models.

[*] We thank Nikhil Agarwal, Victor Aguiar, Abhijit Banerjee, Tilman Börgers, Vince Crawford, Glenn Ellison, Shaowei Ke, Rosa Matzkin, John Quah, Charles Sprenger, and Emanuel Vespa for helpful comments, and NSF grant SES 1643517 for financial support.

[†] Department of Economics, MIT

[‡] Department of Economics, U. Pennsylvania

[§] Department of Economics, U. Pennsylvania

1 Introduction

If a parametric model fits the available data well, is it because the model captures structure that is 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 model restrictiveness that can distinguish between these two explanations and is easy to compute across a variety of applications. We test the restrictiveness of a model by simulating hypothetical data sets and seeing how well the model can fit this data. A restrictive model performs poorly on most of the hypothetical data, while an unrestrictive model approximates almost all conceivable data.

The specification of the "conceivable" data plays an important role in our measure, and what is viewed as conceivable in a given setting reflects ex-ante knowledge or intuition. For example, an analyst might think that everyone prefers more money to less, or that players are less likely to choose strictly dominated actions. To measure restrictiveness, we propose that the analyst first selects a conceivable set by imposing some constraints on the data, and then generates random data sets that satisfy the specified constraints. Our measure is then determined from the model's performance on this hypothetical data: The restrictiveness of a model is how much it restricts behavior beyond these background restrictions.

We complement the evaluation of restrictiveness, which is based solely on hypothetical data, with an evaluation of the model's performance on actual data, using the measure of completeness proposed in Fudenberg et al. (2019). If a model is very unrestrictive, then its completeness on the real data does not directly speak to its relevance. In contrast, a model that is simultaneously restrictive and complete encodes important structure.

Our restrictiveness measure can be computed from data without guidance from analytical results about the model's implications or empirical content, so it can be used in settings where there no such results are available.[1] We provide estimators for restrictiveness and completeness, and characterize their asymptotic distributions and

[1] There 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 the functional forms that are commonly used in applied work.

standard errors. These results tell us, among other things, how many hypothetical data sets need to be generated in order to achieve a given level of approximation to our measure of restrictiveness. We then apply our method and estimators to evaluate parametric models from two classic settings in experimental economics: predicting certainty equivalents for binary lotteries and predicting initial play in matrix games. In each of these domains, these measures reveal new insights about the models we examine.

Our first application considers 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 evaluate the completeness of these models using a set of binary lotteries from Bruhin et al. (2010), and find that CPT's completeness is 95%, which is almost as high as possible, while DA is only 27% complete.

One possible explanation for this finding is that CPT is substantially more flexible than DA. We evaluate the restrictiveness of the two models by evaluating the fit of these models on a large number of hypothetical sets of certainty equivalents, and find that CPT is indeed less restrictive than DA: The restrictiveness of CPT is 0.32, while the restrictiveness of DA is 0.46. Thus, while CPT performs substantially better for predicting the real data, DA rules out more behaviors. Some ways of trading off completeness and restrictiveness imply that CPT's higher completeness more than compensates for its greater flexibility, but we do not take a stand on this here.

Next, we evaluate completeness and restrictiveness for lower-parameter specifications of the two models. Of the nine specifications we consider, three turn out to be dominated, in the sense that there is an alternative specification which is simultaneously more complete and also more restrictive. By comparing nested models that differ by only one parameter, we can better understand that parameter's effect. Adding a parameter always at least weakly increases completeness and decreases restrictiveness, but some parameters handle this tradeoff better than others. We find that the nonlinear probability weighting parameters in CPT and DA are more effective than their utility curvature parameters, and one of the nonlinear probability weighting parameters in CPT is especially effective.

Our second application is to the prediction of initial play in 3×3 matrix games, using a data set from Fudenberg and Liang (2019). We evaluate the restrictiveness of the Poisson Cognitive Hierarchy Model (PCHM) (Camerer et al., 2004) by generating hypothetical distributions of play and evaluating how well the PCHM fits the hypothetical data. We find that in contrast to CPT, the PCHM is very restrictive: Most hypothetical distributions are poorly fit by the PCHM for any parameter values, suggesting that the PCHM isolates a systematic regularity.

We next compare the PCHM with two alternative models: *Logit Level-1*, which models the distribution of play as a logistic best reply to the uniform distribution, and *Logit PCHM*, which allows for logistic best replies in the PCHM (Wright and Leyton-Brown, 2014). Logit Level-1 and Logit PCHM turn out to be almost identical from the perspective of completeness and restrictiveness, suggesting that their empirical content is quite similar. This observation is somewhat surprising to us, since the functional forms of these two models do not bear an obvious relationship. Both models are more complete and less restrictive than PCHM.

Our measure of restrictiveness provides a new perspective on the problem of how richly to parameterize a model. In contrast to cross-validation or other methods (such as AIC and BIC) for guarding against overfitting in finite samples, our approach supposes an intrinsic preference for more parsimonious models even with an infinite data set as in e.g. Gabaix and Laibson (2008).

Our analysis also demonstrates that it is not sufficient to count parameters to understand a model's restrictiveness. As we show, even with just 3 parameters, CPT is not very restrictive on the domain of binary lotteries. CPT can become more restrictive when some of these free parameters are removed, but it matters which ones: Some two-parameter versions of CPT are much more restrictive than others. Moreover, models with different numbers of parameters (such as Logit PCHM and Logit level-1) can be quite similar in their level of restrictiveness. These comparisons are not easy to see from the functional forms associated with a model, but they are revealed by our restrictiveness measure.

2 Related Work

Koopmans and Reiersol (1950) defined a model to be *observationally restrictive* if the distributions it allows are a proper subset of the distributions that would otherwise be possible. Their definition is with respect to an ambient family of outcome distributions; when this ambient family consists of every distribution, a non-restrictive theory cannot be refuted from data.[2]

Selten (1991) subsequently proposed measuring the restrictiveness of a model by the fraction of possible data sets that it can exactly explain. To compute this measure, the analyst needs to know which data sets are consistent with the model. This is a demanding criterion that is only satisfied in some special cases.[3] In contrast, our measure of restrictiveness is based on approximate rather than exact fit to a model, and we compute the model's fit numerically. In this respect, our approach is closer to papers that measure the distribution of the Afriat index (Choi et al., 2007; Polisson et al., 2020).[4]

The use of simulated data to evaluate restrictiveness is similar in spirit to the use of simulated data to evaluate the power of a hypothesis test, as in Bronars (1987)'s numerical evaluation of a test of GARP proposed by Varian (1982), but it is not linked to hypothesis testing. We also provide statistical estimators for our proposed measures and standard errors for these estimates.

Our work complements the representation theorems of decision theory, which describe the empirical content of different models. For example, although there are theorems that characterize which data are consistent with a general Cumulative Prospect

[2] 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.

[3] These cases include whether individual choices from budget sets are consistent with maximization of a utility function (Beatty and Crawford, 2011) and whether individual choices between certain pairs of lotteries are consistent with expected utility, or one of its generalizations (Hey, 1998; Harless and Camerer, 1994).

[4] Choi et al. (2007) and Polisson et al. (2020) relax the implications of expected utility maximization using Afriat's "efficiency index" as an analog of our loss function. They then compare the distribution of the efficiency indices of the actual subjects with the distribution of efficiency indices in randomly generated data. Our approach is designed to evaluate parametric models, while GARP is nonparametric, but can be seen as a way of extending a similar idea to other problem domains and "loss functions."

Theory specification (Quiggin, 1982; Yaari, 1987), we know of no representation theorems for the popular functional forms we study here. We are also unaware of representation theorems for the Poisson Cognitive Hierarchy Model. Moreover, even when a representation theorem is available, it can be computationally challenging to determine whether a given data set is consistent with the characterization.[5]

Our paper is related to the vast literature in statistics and econometrics on model selection, which dates back to Cox (1961, 1962). Unlike classic measures, including AIC and BIC, restrictiveness is not based on observed data.[6] Various complexity measures, such as VC dimension[7], metric entropy (with or without bracketing) and Rademacher complexity, are related to our restrictiveness measure at a high level, but it is generally nontrivial to analytically derive these measures on any given model. Moreover, in statistics and econometrics, often analysts are only concerned with bounds on the "order of magnitude" of such measures: see, e.g. Van Der Vaart and Wellner (1996). In contrast, our restrictiveness measure is (by design) easy to compute.

Finally, our paper joins a recent literature on experimental design and the interpretation of experimental results. This includes, for example, DellaVigna and Pope (2019), de Quidt et al. (2018), and Shmaya and Yariv (2016) on the external validity of experimental results; Andrews and Kasy (2019), Frankel and Kasy (2019), and DellaVigna and Linos (2020) on p-hacking and publication bias; Fudenberg and Levine (2020) on the interpretation of natural experiments; and Chassang et al. (2012), Chemla and Hennessy (2019), and Banerjee et al. (2020) on the intepretation of randomized controlled trials.

[5] For example, the Harless and Camerer (1994) exercise would be much harder to implement on larger menus of binary lotteries, or on 3-outcome lotteries, or if subjects had been asked to report real-valued certainty equivalents.

[6] Note that our notion of restrictiveness is different than the question of how the "restrictiveness" of an econometric model can affect the identifiability of parameters and the efficiency of estimators.

[7] The VC dimension is known for very few economic models. A recent exception is the work of Basu and Echenique (2020) for various models of decision-making under uncertainty.

3 Approach

3.1 Preliminaries

Let X be an observable (random) *feature vector* taking values in a finite set $\mathcal{X}$, and Y be a random *outcome variable* taking values in a finite-dimensional set $\mathcal{Y}$. We use P^* to denote the joint distribution of (X, Y), P^*_X to denote the marginal distribution of X and $P^*_{Y|X}$ to denote the conditional distribution of Y given X. We assume that the marginal P^*_X is known to the analyst, while the conditional distribution is not.[8]

The analyst wants to learn a function of the conditional distribution, $s(P^*_{Y|X=x}) \in \mathcal{S}$, where $\mathcal{S}$ is finite-dimensional. We call any function $f : \mathcal{X} \to \mathcal{S}$ a *predictive mapping*, or simply *mapping*, and denote the *true mapping* f^* by $f^*(x) := s(P^*_{Y|X=x})$. The set of all possible mappings is denoted by $\mathcal{F}$.

We focus on two leading cases of this problem whose structure makes our methods easier to explain; Section 8 explains how to extend our approach to more general problems.

Prediction of a Conditional Expectation. When the statistic of interest is $\mathbb{E}_{P^*}[Y|\,X]$, the analyst's objective is to learn the average outcome for each realization of X. To evaluate the error of predicting $f(x)$ when the realized outcome is y, we use squared loss $l\left(f,(x,y)\right) := \left(y - f(x)\right)^2$. The *expected error* of a mapping f is then $e_{P^*}(f) = \mathbb{E}_{P^*}\left[(Y - f(X))^2\right]$, which is minimized by the true mapping $f^*(x) = \mathbb{E}_{P^*}\left[Y|\,X = x\right]$. It is a standard result that the difference between the error $e_{P^*}(f)$ of an arbitrary mapping $f \in \mathcal{F}$ and the best possible error $e_{P^*}(f^*)$ is

$$d_{MSE}(f, f^*) := e_{P^*}(f) - e_{P^*}(f^*) := \mathbb{E}_{P^*_X}\left[\left(f^*(X) - f(X)\right)^2\right], \tag{1}$$

i.e. the expected mean-squared difference between the predicted outcomes. (See Appendix C for details.)

Our first application, predicting the average reported certainty equivalent for binary lotteries, falls into this setting. Each lottery is described as a tuple $x = (\overline{z}, \underline{z}, p)$,

[8] For example, in a decision theory experiment the experimenter knows the distribution over menus that the subjects will face.

and the feature space $\mathcal{X}$ consists of the 25 tuples associated with lotteries in a data set from Bruhin et al. (2010). The outcome space of certainty equivalents is $\mathcal{Y} = \mathbb{R}$, and we seek to predict the population average of certainty equivalents for each lottery $x \in \mathcal{X}$. A predictive mapping for this problem specifies an average certainty equivalent for each of the 25 binary lotteries.

Prediction of a Conditional Distribution. Here the statistic of interest is $P^*_{Y|X}$, so the analyst's objective is to learn the conditional distribution itself. To evaluate the error of predicting the distribution $f(x)$ when the realized outcome is y, we use the negative (conditional) log-likelihood $l\left(f,(x,y)\right) := -\log f\left(y|\, x\right)$. The expected error of mapping f is $e_{P^*}\left(f\right) = \mathbb{E}_{P^*}\left[-\log f\left(Y|\, X\right)\right]$, which is minimized by the true conditional distribution $f^*(x) = P^*_{Y|X}(x)$. As we show in Appendix C, the difference between the error of an arbitrary mapping $f \in \mathcal{F}$, $e_{P^*}(f)$ and the best possible error, $e_{P^*}(f^*)$, is

$$d_{KL}(f, f^*) := e_{P^*}(f) - e_{P^*}(f^*) := \mathbb{E}_{P^*_X}\left[\sum_y f^*\left(y|\, x\right)\left[\log f^*\left(y|\, x\right) - \log f\left(y|\, x\right)\right]\right], \tag{2}$$

i.e. the expected Kullback-Liebler divergence from f to the true distribution f^*.

Our second application, predicting initial play in in matrix games, is an example of this case. Here the feature space $\mathcal{X}$ consists of the 466 unique 3×3 matrix games from Fudenberg and Liang (2019), each described as a vector in $\mathbb{R}^{18}$. The outcome space is $\mathcal{Y} = \{a_1, a_2, a_3\}$ (the set of row player actions) and the analyst seeks to predict the conditional distribution over $\mathcal{Y}$ for each game, interpreted as choices made by a population of subjects for the same game. Thus, $\mathcal{S} = \Delta(\mathcal{Y})$, the set of all distributions over row player actions. A predictive mapping is any function $f : \mathcal{X} \to \mathcal{S}$ taking the 466 games into predicted distributions of play.

3.2 Restrictiveness

Our goal is to evaluate the restrictiveness of parametric models $\mathcal{F}_\Theta = \{f_\theta\}_{\theta\in\Theta} \subseteq \mathcal{F}$, where the permitted mappings f_θ are indexed by a finite dimensional parameter θ and Θ is a compact set. If the model $\mathcal{F}_\Theta$ contains a mapping that can approximate the predictions of the true mapping f^*, then $\inf_{f \in \mathcal{F}_\Theta} e_{P^*}(f)$ also approximates the true mapping's error, $e_{P^*}(f^*)$. Given enough data to train on, such a model will thus predict about as well as possible, but a good fit to the data could be because the model isolates the "right" regularities, or because it is simply flexible enough to accommodate any pattern of behavior.

Our strategy to determine the restrictiveness of a model is to generate mappings f from a set $\mathcal{F}_\mathcal{M} \subseteq \mathcal{F}$ of "conceivable" mappings. This set encodes prior knowledge or intuition about the setting.[9]

We generate random mappings from $\mathcal{F}_\mathcal{M}$ according to a distribution $\mu \in \Delta(\mathcal{F}_\mathcal{M})$, which we interpret as the analyst's prior over the space of conceivable models. When the analyst does not have a priori reasons to consider one mapping from the conceivable set more likely than another, it seems natural to specify that μ is the uniform distribution. The use of a uniform prior is standard also in many computer science literatures: For example, in computational complexity, the average-case time complexity of an algorithm measures the amount of time used by the algorithm, averaged over all possible inputs (Goldreich and Vadhan, 2007).

An alternative perspective is that μ and $\mathcal{F}_\mathcal{M}$ are user inputs. Just as it can be interesting how a model's parameter estimates and fit vary across subject populations and settings, it can be instructive to understand how a model's restrictiveness varies with respect to different choices of μ and $\mathcal{F}_\mathcal{M}$. We conduct such an exercise for each of our leading applications.

Given a fixed μ with support on $\mathcal{F}_\mathcal{M}$, we generate random mappings according to μ and evaluate restrictiveness of a model $\mathcal{F}_\Theta$ based on the model's ability to fit these generated mappings. We evaluate "fit" using a function $d(f, f')$ that tells us how close the generated mapping f is to candidate mappings f' from the model $\mathcal{F}_\Theta$. When predicting conditional expectations, we define $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ to extend d_{MSE} (as given in (1)) to

$$d_{MSE}(f', f) := \mathbb{E}_{P_X^*}\left[\left(f'(X) - f(X)\right)^2\right].$$

[9] The choice of $\mathcal{F}_\mathcal{M}$ is somewhat analogous to the choice of what alternatives to consider when computing the power of a statistical test. In both cases, the right choice is guided by intuition and prior knowledge, and not derived from formal considerations.

When predicting a conditional distribution, we define d to extend d_{KL} (as given in (2)) to

$$d_{KL}(f', f) := \mathbb{E}_{P_X^*}\left[\sum_y f\left(y|\, x\right)\left[\log f\left(y|\, x\right) - \log f'\left(y|\, x\right)\right]\right].$$

Since our subsequent definitions and results hold for both of these functions, we simply use the notation d, understanding that it means d_{MSE} for predicting the conditional expectation, and d_{KL} for predicting the conditional distribution.

The model's approximation error to a generated mapping f is $d(\mathcal{F}_\Theta, f) := \inf_{\theta\in\Theta} d(f_\theta, f)$. We normalize this raw error relative to a benchmark *naive mapping* $f_{\text{naive}} \in \mathcal{F}_\Theta$ chosen to suit the problem, which we interpret as a lower bound that any sensible model should outperform.[10]

Definition 1. The *f-discrepancy* of model $\mathcal{F}_\Theta$ is

$$\delta_f := \frac{d(\mathcal{F}_\Theta, f)}{d(f_{naive}, f)}$$

with $\delta_f := 0$ if $d(\mathcal{F}_\Theta, f) = d(f_{naive}, f) = 0$.

Since f_{naive} is assumed to be an element of $\mathcal{F}_\Theta$, the f-discrepancy of $\mathcal{F}_\Theta$ is bounded above by 1, and since d is nonnegative, the f-discrepancy is also bounded below by 0.[11] Thus, the f-discrepancy in any problem must fall between 0 and 1. Large values of δ_f imply that the model does not approximate f much better than the naive mapping does. Since the naive mapping itself has no free parameters and therefore does not have the flexibility to accommodate most mappings, concentration of the distribution of δ_f around values close to 1 implies that the model rules out many kinds of regularities.

The *restrictiveness* of model $\mathcal{F}_\Theta$ is its average f-discrepancy.

Definition 2. The *restrictiveness* of model $\mathcal{F}_\Theta$ is $r := \mathbb{E}_\mu\left[\delta_f\right].$

[10] For example, in our application to predicting initial play in games, we define the naive mapping to predict a uniform distribution of play in every game.

[11] If $d(f_{naive}, f) = 0$, then from our assumption that $f_{naive} \in \mathcal{F}_\Theta$, it follows that $d(\mathcal{F}_\Theta, f) = 0$. In this case we set f-discrepancy δ_f to be 0. As long as μ has no atoms, the definition of δ_f at $d(f_{naive}, f) = 0$ does not matter substantially.

If $\mathcal{F}_\Theta = \mathcal{F}_M$ (so that the model is completely unrestrictive), then $r = 0$ for every choice of μ with support on $\mathcal{F}_M$.

3.3 Completeness

While restrictive models are desirable, a restrictive model is not particularly useful if it fails to predict real data. We would like models to embody regularities that are present in actual behavior, and rule out conceivable regularities that are not. We thus evaluate models from the dual perspectives of how restrictive they are and how well they predict actual data. The latter can be measured using the f^*-discrepancy of the model, where f^* is the true mapping. This measure is tightly linked to the notion of *completeness* introduced in Fudenberg et al. (2019).

Definition 3 (Fudenberg et al., 2019)*.* The *completeness* of model $\mathcal{F}_\Theta$ is

$$\kappa^* := \frac{e_{P^*}(f_{\text{naive}}) - e_{P^*}(\mathcal{F}_\Theta)}{e_{P^*}(f_{\text{naive}}) - e_{P^*}(f^*)},$$

where $e_{P^*}(\mathcal{F}_\Theta) := \inf_{\theta \in \Theta} e_{P^*}(f_\theta)$.

Completeness is the complement of the f^*-discrepancy, since

$$\kappa^* = 1 - \frac{e_{P^*}(\mathcal{F}_\Theta) - e_{P^*}(f^*)}{e_{P^*}(f_{\text{naive}}) - e_{P^*}(f^*)} = 1 - \frac{d(\mathcal{F}_\Theta, f^*)}{d(f_{\text{naive}}, f^*)} = 1 - \delta_{f^*}. \tag{3}$$

A model's completeness can be interpreted as the ratio of the reduction in error achieved by the model (relative to the naive baseline), compared to the largest achievable reduction. By construction, the measure κ^* is scale-free and lies within the unit interval. A large κ^* suggests that the model is able to approximate the real data well: at the extremes, a model with $\kappa^* = 1$ matches the true mapping f^* exactly, while a model with $\kappa^* = 0$ is no better at matching f^* than the naive model. We will report both restrictiveness r and completeness κ^* for each of the models that we consider.

3.4 Discussion of Measures

An alternative "area" measure. Selten's *area measure* of model flexibility is $a := \lambda(\mathcal{F}_\Theta)$, where λ is the Lebesgue measure, i.e. the fraction of possible mappings that are exactly consistent with the model. Our measure of restrictiveness differs both by normalization with respect to the performance of a naive model, and by measuring how well the model $\mathcal{F}_\Theta$ *approximates* a randomly drawn mapping f in $\mathcal{F}_\mathcal{M}$, which allows us to quantify the degree of error. A model that does not include most mappings from $\mathcal{F}_\mathcal{M}$ would be considered highly restrictive under the Selten measure, but would have low restrictiveness by our measure if it approximated most mappings very well.

Role of the normalization. We define restrictiveness to be the average value of $d(\mathcal{F}_\Theta, f)/d(f_{\text{naive}}, f)$, rather than its un-normalized counterpart $d(\mathcal{F}_\Theta, f)$. Normalizing relative to a naive mapping has several advantages compared to the unit-dependent raw error $d(\mathcal{F}_\Theta, f)$: If we were to scale up the payoffs in the binary lotteries in our first application, then $d(\mathcal{F}_\Theta, f)$ would mechanically scale up as well, even though the flexibility of the model has not changed, which makes it hard to say what constitutes a "large" value of $d(\mathcal{F}_\Theta, f)$. Normalizing relative to the naive error returns a unitless quantity that is easier to interpret, and can more easily be compared across problems that use different error metrics.

Relationship to Generalized and Pseudo R^2. With the squared loss function $e_{P^*}(f) := \mathbb{E}[(Y - f(X))^2]$, and $f_{naive}(x) \equiv \mathbb{E}[Y]$, the population analog of the more familiar concept of R *squared* in econometrics would be given by

$$R^2 = \frac{e_{P^*}(f_{\text{naive}}) - e_{P^*}(\mathcal{F}_\Theta)}{e_{P^*}(f_{\text{naive}})}$$

which is similar to but still different from the completeness measure in Definition 3. Specifically, the denominator of completeness is given by $e_{P^*}(f_{\text{naive}}) - e_{P^*}(f^*)$ rather than $e_{P^*}(f_{\text{naive}})$, since in our context we explicitly acknowledge the *irreducible error* $e_{P^*}(f^*)$, and normalize the error reduction achieved by $\mathcal{F}_\Theta$ from f_{naive} with the "largest possible" reduction $e_{P^*}(f_{\text{naive}}) - e_{P^*}(f^*)$. With the log-likelihood loss function and application-specific choices of f_{naive}, completeness would look similar to "generalized pseudo R squared,"[12] subject to the same difference in the denominators and the corresponding differences in the interpretation.

Sensitivity to μ. For any two measures $\mu, \mu' \in \Delta(\mathcal{F})$,

$$\mathbb{E}_{\mu}\left[\delta_f\right] - \mathbb{E}_{\mu'}\left[\delta_f\right] \leq \int \delta_f \cdot |d\mu - d\mu'| \leq 2 \cdot d_{TV}(\mu, \mu'), \tag{4}$$

where d_{TV} is the total variation distance. Thus for any two measures that are close in total variation distance, the corresponding restrictiveness measures must also be close. We complement this theoretical bound with a numerical sensitivity check in Section 5.3, where we evaluate restrictiveness with respect to beta distributions that are close to our specification that μ is uniform. The resulting variation in restrictiveness is quite small.

Combining κ^* and r. Ideal models have high κ^*, so they approximate the real data well, but also high restrictiveness r, so they rule out regularities that could have been present but are not. These two criteria generate a partial order on models, and there are many ways to complete it. One possibility is to use a lexicographic ordering, where models are ordered first by κ^* and then by r. Another is to impose a functional form that combines κ^* and restrictiveness r, such as $r - (1 - \kappa^*) = \mathbb{E}_{\mu}[\delta_f] - \delta_{f^*}$.[13] Yet another possibility is to use the probability that the model fits the actual data better than it fits a randomly generated data set, namely the quantile of δ_{f^*} under the distribution of f-discrepancies. In the present paper, we report κ^* and r separately, and leave it to the analyst's discretion whether or how to combine these two metrics.

Point-Identified and Set-Identified Models. Note that f-discrepancy, restrictiveness, and completeness are well-defined regardless of whether the parametric model $\mathcal{F}_\Theta$ is point-identified or set-identified. This is because the definitions of

[12] By generalized R^2 we mean the generalization of R^2 by replacing a constant with a general chosen naive model f_{naive}, as in Anderson-Sprecher (1994). This is also often referred to as partial R^2. By pseudo R^2 we mean generalizing R^2 by replacing squared loss with other loss functions, such as the negative log-likelihood in McFadden (1974).

[13] Selten (1991) provided an axiomatic characterization of the similar aggregator $m = r - a$, where r is the pass rate of the model on the actual data and a is the area measure we discussed above.

$d(\mathcal{F}_\Theta, f)$, restrictiveness, and $e_{P^*}(\mathcal{F}_\Theta)$ do not rely on the uniqueness of the minimizers. In other words, we evaluate the parametric model $\mathcal{F}_\Theta$ with d and e_{P^*}, so our measures do not differentiate point-identified models from set-identified models that yield the same $d(\mathcal{F}_\Theta, f)$ and $e_{P^*}(\mathcal{F}_\Theta)$.

4 Estimates and Test Statistics

We now discuss how to implement our approach in practice. Recall that we restrict $\mathcal{X}$ to be finite so that $\mathcal{F}_\mathcal{M}$ is finite-dimensional. In Appendix F, we provide a discussion on how to compute restrictiveness and estimate completeness when $\mathcal{X}$ is a continuum and $\mathcal{F}_\mathcal{M}$ is infinite-dimensional.

4.1 Computing Restrictiveness r

We provide an algorithm for computing r: Sample M times from the distribution μ on $\mathcal{F}_\mathcal{M}$, and for each sampled $f_m \in \mathcal{F}_\mathcal{M}$, compute $\delta_m := \frac{d(\mathcal{F}_\Theta, f_m)}{d(f_{\text{naive}}, f_m)}$. The sample mean $\overline{\delta}_M := \frac{1}{M}\sum_{m=1}^{M} \delta_m$ is an estimator for restrictiveness. In principle, the number of simulations we run, M, can be taken as large as we want, so $\overline{\delta}$ can be made arbitrarily close to r by the Law of Large Numbers. Moreover, the approximation error under a given finite M can be quantified using standard statistical inference methods. We focus on the case where the distribution of δ_m is nondegenerate.

Assumption 1. *The distribution of δ_m is non-degenerate.*

Assumption 1 is a very mild condition that can be easily verified, as it is sufficient for any two δ_m and δ'_m to be distinct.

The sample variance is

$$\hat{\sigma}_\delta^2 := \frac{1}{M}\sum_{m=1}^{M} \left(\delta_m - \overline{\delta}_M\right)^2 \tag{5}$$

and the standard Central Limit Theorem gives the following result.

Proposition 1. *Under Assumption 1,*

$$\frac{\sqrt{M}\left(\overline{\delta}_M - r\right)}{\hat{\sigma}_\delta} \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

The $(1-\alpha)$*-th confidence interval for* r *is given by*

$$\left[\overline{\delta}_M - q_{1-\alpha/2}\cdot\frac{1}{\sqrt{M}}\hat{\sigma}_\delta,\ \overline{\delta}_M - q_{\alpha/2}\cdot\frac{1}{\sqrt{M}}\hat{\sigma}_\delta\right],$$

where $\hat{\sigma}_\delta$ *is given in (5) and* q_α *denotes the* α*-th quantile of the standard normal distribution.*

One-sided confidence intervals for r can also be constructed in standard ways.[14] We again note that the confidence intervals here simply measure the approximation error of r based on a finite number of simulations and do not reflect randomness in experimental data.

4.2 Estimating Completeness κ^*

In this section, we show how to estimate completeness, κ^*.

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, we use K-fold cross-validation to estimate the out-of-sample prediction error of the model.[15] (In our applications, we take the standard choice of $K = 10$.) Specifically, we randomly divide $\mathbf{Z}_N$ into K (approximately) equal-sized groups. To simplify notation, assume that $J_N = \frac{N}{K}$ is an integer. Let $k\left(i\right)$ denote the group number of observation Z_i, and for each group $k = 1, ..., K$, define

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

to be the mapping from $\widetilde{\mathcal{F}}$ that minimizes error for prediction of observations outside

[14] For example, $\left[0,\ \overline{\delta}_M - q_\alpha\cdot\frac{1}{\sqrt{M}}\hat{\sigma}_\delta\right]$ is the $1-\alpha$ confidence interval that has the tightest upper confidence bound on restrictiveness.

[15] Alternatively, we can use the in-sample error estimator without cross validation, if we are not concerned with out-of-sample errors in finite samples. See, e.g., the estimator in Appendix F.

of group 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 on the k-th test set. Then,

$$CV\left(\widetilde{\mathcal{F}}\right) := \frac{1}{K} \sum_{k=1}^{K} \hat{e}_k$$

is the average test error across the K folds. This is an estimator for the unobservable expected error of the best mapping from class $\widetilde{\mathcal{F}}$.

Setting $\widetilde{F}$ to be $\mathcal{F}_\Theta$, $\mathcal{F}$, or $\mathcal{F}_{\text{naive}} = \{f_{\text{naive}}\}$, we can compute $CV\left(\mathcal{F}_\Theta\right)$, $CV\left(\mathcal{F}\right)$ and $CV\left(\mathcal{F}_{\text{naive}}\right)$ from the data, leading to the following estimator for κ^*:

$$\hat{\kappa}^* = 1 - \frac{CV\left(\mathcal{F}_\Theta\right) - CV\left(\mathcal{F}\right)}{CV\left(\mathcal{F}_{naive}\right) - CV\left(\mathcal{F}\right)}.$$

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

Assumption 2 (Naive Rule is Imperfect)**.** $e\left(f_{naive}\right) - e\left(f^*\right) > 0$.

This assumption is quite weak, as it simply says that the naive mapping performs strictly worse in expectation than the best mapping.

Under additional technical conditions, we show, by applying and adapting Proposition 5 in Austern and Zhou (2020), that $\hat{\kappa}^*$ is asymptotically normal.

Theorem 1. *Under Assumption 2 and some regularity conditions,*[16]

$$\frac{\sqrt{N}\left(\hat{\kappa}^* - \kappa^*\right)}{\hat{\sigma}_{\hat{\kappa}^*}} \xrightarrow{d} \mathcal{N}\left(0, 1\right),$$

where the estimate $\hat{\sigma}_{\hat{\kappa}^*}$ of the standard deviation is as defined in Appendix D.2. Consequently, the $(1-\alpha)$-th two-sided confidence interval for κ^* is given by

$$\left[\hat{\kappa}^* - q_{1-\alpha/2} \cdot \frac{1}{\sqrt{N}} \hat{\sigma}_{\hat{\kappa}^*},\ \hat{\kappa}^* - q_{\alpha/2} \cdot \frac{1}{\sqrt{N}} \hat{\sigma}_{\hat{\kappa}^*}\right],$$

[16] See Appendix D for details of these assumptions.

where q_α denotes the α-th quantile of the standard normal distribution.

5 Application 1: Certainty Equivalents

5.1 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 $x = (\overline{z}, \underline{z}, p)$, where $\overline{z} > \underline{z} \geq 0$ are the two possible prizes, and p is the probability of the larger prize $\overline{z}$. The feature space $\mathcal{X}$ consists of the 25 tuples associated with lotteries in the Bruhin et al. (2010) data, and the outcome space is $\mathcal{Y} = \mathbb{R}$. Each observation in the data is a pair consisting of a lottery and a reported certainty equivalent by a given subject. Note that the variation in Y for fixed X reflects heterogeneity in certainty equivalents reported across subjects for the same lottery. (In Appendix A.5, we discuss how to extend our approach to allow for subject-level heterogeneity.)

We predict the average (over subjects) certainty equivalent for each lottery in this data set. A mapping for this problem is any function $f : \mathcal{X} \to \mathbb{R}$ from the 25 lotteries to average certainty equivalents, and the distance between $d(f, f')$ between two mappings f and f' is defined to be the expected mean-squared distance between the two mappings' predictions, as in (1).

We evaluate two economic models. First we consider a three-parameter version of *Cumulative Prospect Theory* indexed by $\theta = (\alpha, \gamma, \zeta)$, which specifies a "utility"

$$w(p)v(\overline{z}) + (1 - w(p))v(\underline{z}) \tag{6}$$

for each lottery $(\overline{z}, \underline{z}, p)$, where

$$v(z) = z^\alpha \tag{7}$$

is a value function for money, and

$$w(p) = \frac{\zeta p^\gamma}{\zeta p^\gamma + (1-p)^\gamma} \tag{8}$$

is a probability weighting function.[17] The predicted certainty equivalent of a binary lottery is

$$f_\theta(\overline{z}, \underline{z}, p) = v^{-1}\left(w(p)v(\overline{z}) + (1 - w(p))v(\underline{z})\right).$$

Following the literature, we restrict $\alpha, \gamma \in [0,1]$, and $\delta \geq 0$. We specify $\mathcal{F}_\Theta$ as the set of all such functions f_θ with parameters θ in this range, and refer to this model simply as CPT. As a naive benchmark, we consider the function f_{naive} that maps each lottery into its expected value, corresponding to $\alpha = \gamma = \zeta = 1$.

Second, we consider the *Disappointment Aversion* model of Gul (1991), using a parametric form proposed in Routledge and Zin (2010) with the parameters $\lambda = (\alpha, \eta)$, where $\alpha \in [0,1]$ and $\eta > -1$.[18] The value function for money is the same as in (7), but the probability weighting function is given instead by

$$\widetilde{w}(p) = \frac{p}{1 + (1-p)\eta}$$

There are two parameters: α again reflects the curvature of the utility function, while $\eta > 0$ corresponds to "disappointment aversion", meaning aversion to realizations of the lottery that are worse than its certainty equivalent.

The predicted certainty equivalent is

$$g_\lambda(\overline{z}, \underline{z}, p) = v^{-1}(\widetilde{w}(p)v(\overline{z}) + (1 - \widetilde{w}(p))v(\underline{z})).$$

We specify $\mathcal{G}_\Lambda$ as the set of all such functions g_λ, and refer to this model simply as DA. The naive benchmark is again the function that maps each lottery into its expected value, corresponding to $\alpha = 1$ and $\eta = 0$ in DA.

[17] This parametric form for $w(p)$ was first suggested by Goldstein and Einhorn (1987) and Lattimore et al. (1992). Following Bruhin et al. (2010) and much of the literature, we will estimate separate values of these parameters for losses (see Section 5.6), so in a sense the "overall CPT model" has 6 parameters.

[18] To facilitate comparison with CPT, we depart slightly from Routledge and Zin (2010) by imposing the functional form $v(z) = z^\alpha$ instead of $v(z) = z^\alpha/\alpha$.

5.2 Completeness

CPT achieves a striking out-of-sample performance for predicting the average certain equivalents in Bruhin et al. (2010) data: it is 95% complete.[19] Thus, the model achieves almost all of the possible improvement in prediction accuracy over the naive baseline.[20] 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 substantial restrictions. These explanations have very different implications for how to interpret CPT's empirical success compared to DA's.

5.3 Restrictiveness

To distinguish between these explanations, we now compute the restrictiveness of the two models. Our primitive distribution μ is a uniform distribution over the set of all mappings satisfying the following criteria:[21]

1. $\underline{z} \leq f(\overline{z}, \underline{z}, p) \leq \overline{z}$

2. if $\overline{z} > \overline{z}'$, $\underline{z} > \underline{z}'$, and $p \geq p'$ then $f(\overline{z}, \underline{z}, p) > f(\overline{z}', \underline{z}', p')$

3. if $\overline{z} \geq \underline{z}$, $p > p'$, then $f(\overline{z}, \underline{z}, p) > f(\overline{z}, \underline{z}, p')$

Constraint (1) requires that the certainty equivalent is within the range of the possible payoffs, while constraints (2) and (3) are equivalent to first-order stochastic dominance. There are many pairs of lotteries in the Bruhin et al. (2010) lottery data that can be compared via (2) and (3), so these conditions are not vacuous.

The restrictiveness of the two models is given in the table below.

[19] $\frac{CV(\mathcal{F}_{naive})-CV(\mathcal{F}_{\Theta})}{CV(\mathcal{F}_{naive})-CV(\mathcal{F})} = \frac{98.32-63.75}{98.32-61.87} = 0.95$. A similar result was reported in Fudenberg et al. (2019) for the pooled sample of gain-domain and loss-domain lotteries.

[20] 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.

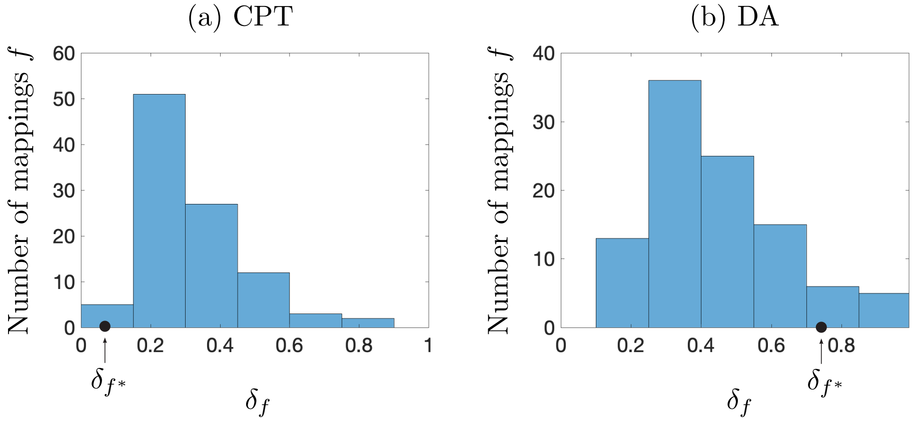
[21] This uniform distribution is well-defined since $\mathcal{F}_M$ is a bounded subset of $\mathbb{R}^{50}$.

	Completeness κ^*	Restrictiveness r
CPT	0.95	0.32
	(0.02)	(0.02)
DA	0.27	0.46
	(0.06)	(0.02)

Table 1: Completeness and restrictiveness measures for CPT and DA. Completeness is estimated from 8906 observations; restrictiveness is estimated from 100 simulations.

The restrictiveness of CPT is 0.32, so on average, CPT's approximation error is about a third of the error of the naive (expected-value) mapping. This implies that CPT—while not completely unrestrictive—is quite flexible. DA's approximation error is roughly a half of the error of the naive mapping on average, so it is more restrictive. 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.

To gain further insight into the restrictiveness of the two models, we plot below the distribution of f-discrepancies for 100 random mappings f for each model. (Recall that restrictiveness is equal to the average f-discrepancy.) For comparison, we also plot the f^*-discrepancy estimated on the actual data (i.e. $\delta_{f^*} = 1 - \kappa^*$).


(a) CPT

(b) DA

Figure 1: Distribution of f-discrepancies for 100 randomly generated mappings f

For 91 of the 100 simulated mappings, DA's completeness exceeds its completeness on the real data,[22] so DA better approximates most of our simulated data sets compared to the actual Bruhin et al. (2010) data. In contrast, CPT's completeness improves over its completeness on the real data for only 1 of the 100 hypothetical mappings. Taking both completeness and restrictiveness into account via a composite such as the difference $r - (1 - \kappa^*)$, which is 0.27 for CPT but -0.27 for DA, CPT's high completeness on real data more than compensates for its moderately low restrictiveness.[23] We do not take a stand on which composite measure is the right one.

5.4 The Value of a Parameter

In addition to comparing distinct models such as CPT and DA, our approach can also be used to compare nested models. To illustrate this, we now consider alternative specifications of CPT and DA with fewer free parameters. Some of these specifications have been studied in the literature: CPT(α, γ), with ζ set to 1, is the specification used in Karmarkar (1978)[24]; CPT(γ, ζ), with $\alpha = 1$, corresponds to a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but exhibits nonlinear probability weighting; CPT(α), with $\zeta = \gamma = 1$, corresponds to an Expected Utility decision-maker whose utility function is as given in (7), and is also equivalent to DA(α).[25] The model CPT(γ), with $\alpha = \zeta = 1$, and CPT(ζ), with $\alpha = \gamma = 1$ have not been studied in the prior literature, but we report them for comparison. We also consider DA(η) as in Gul (1991), with $\alpha = 1$, which corresponds to a disappointment-averse decision maker whose utility is linear in money. Figure 2 plots restrictiveness and completeness for these alternative specifications (see also Table A.1 in the appendix).

We find that three models—CPT(ζ), CPT(α, ζ), and DA(α, η)—are dominated, in the sense that another model is simultaneously more complete and also more re-

[22] Specifically, there are 91 mappings f for which $\delta_f < \delta_{f^*}$.

[23] See Section 3.4 for further discussion of this composite measure and others.

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

[25] See the survey Fehr-Duda and Epper (2012) for further discussion of these different parametric forms, and others which have been proposed in the literature.

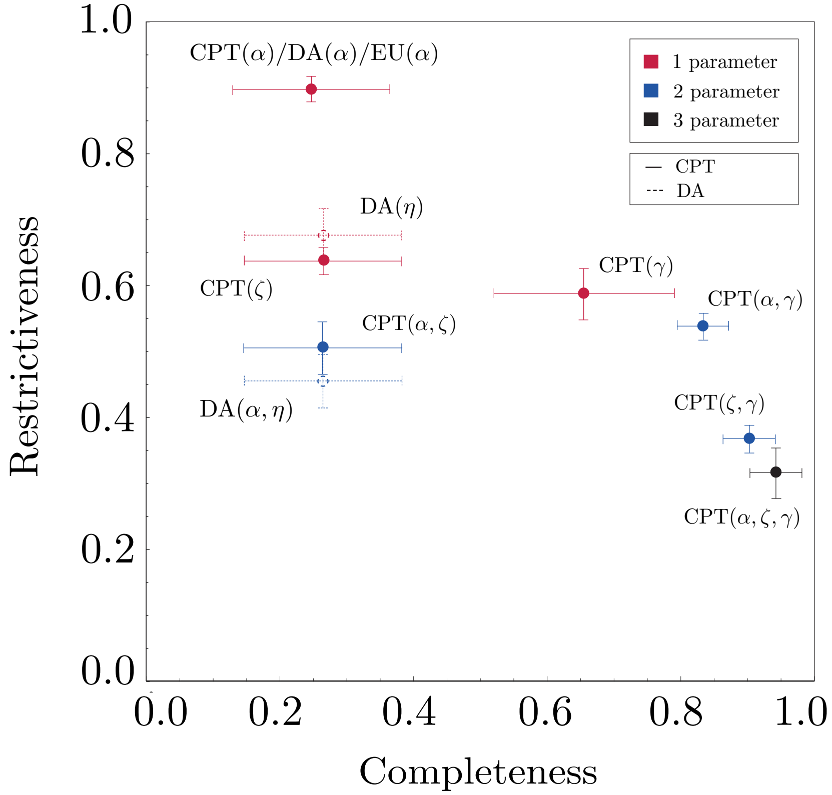


Figure 2: Comparison of models by their completeness and restrictiveness.

strictive. CPT(ζ) is simultaneously less complete and less restrictive than DA(η), while both CPT(α,ζ) and DA(α,η) are less complete and less restrictive than the single parameter model CPT(γ).

On this data, the risk-aversion parameter α does not play important role: The Expected Utility model with utility function $u(x) = x^\alpha$ and $\alpha \in [0,1]$ is the most restrictive of the models we consider ($r = 0.90$), but the least complete ($\kappa^* = 0.25$). Adding the risk-aversion parameter α to the nonlinear probability weighting parameters ζ and γ in CPT leads to only a slight improvement in completeness (κ^* increases from 0.91 to 0.95), and a comparable drop in restrictiveness (r falls from 0.37 to 0.32), while adding the curvature parameter α to the probability weighting parameter η in DA does not improve completeness on the actual data (κ^* remains at 0.27), but leads to a large drop in restrictiveness (r falls from 0.70 to 0.54). These findings suggest that nonlinear probability weighting is an essential part of modeling risk preferences.[26]

The different parameters used for nonlinear probability weighting are not equally effective. Two of the dominated CPT specifications involve the nonlinear probability weighting parameter ζ: CPT(ζ) and CPT(α, ζ). (Perhaps correspondingly, neither of these specifications has been suggested in the literature.) The parameter ζ is more effective when combined with the probability weighting parameter γ: Adding ζ to CPT(γ) improves completeness by 0.25 at a cost of a decrease in restrictiveness of 0.22, and adding ζ to CPT(α, γ) improves completeness by 0.11 at a cost of a decrease in restrictiveness of 0.22.

The nonlinear probability weighting parameter η in DA is also not particularly effective: Adding η over the risk aversion parameter α improves completeness from $\kappa^* = 0.25$ to $\kappa^* = 0.27$ but reduces restrictiveness from $r = 0.90$ to $r = 0.46$. Moreover, the single-parameter DA(η) model is almost dominated by the Expected Utility model with risk aversion: DA(η) is substantially less restrictive than the Expected Utility model (compare $r = 0.90$ with $r = 0.68$) but only marginally more complete (compare $\kappa^* = 0.25$ with $\kappa^* = 0.27$).

In contrast, the nonlinear probability weighting parameter γ seems to model risk attitudes well. The model CPT(γ) is by far the most complete single-parameter specification. Furthermore, adding γ over α improves completeness by 0.59 while decreasing restrictiveness by 0.34; adding γ over ζ improves completeness by 0.64 while decreasing restrictiveness by 0.27; and adding γ over both ζ and α improves completeness by 0.68 while decreasing restrictiveness by 0.19. While these are real reductions in restrictiveness, the sizeable improvements in completeness may compensate for this additional flexibility.

5.5 Dependence on μ

The restrictiveness measure depends on the choice of distribution μ, which we chose above to be uniform over the set $\mathcal{F}_{\mathcal{M}}$ of mappings that respect first-order stochastic

[26] Our finding is consistent with previous studies which find that probability distortions play an important role in explaining experimental and field data (Snowberg and Wolfers, 2010; Fehr-Duda and Epper, 2012; Barseghyan et al., 2013).

dominance. We next consider the robustness of these results to several alternative choices for μ (see Appendix A.2 for further detail).

First, we consider different distributions over the same set of conceivable mappings $\mathcal{F}_{\mathcal{M}}$. 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 $[0.9,1.1]\times[0.9,1.1]$. 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 0.32, with a min value of 0.25 and a max value of 0.40. Thus our finding that CPT is quite flexible is robust to these perturbations in μ.[27]

Next, we compute the restrictiveness of the model with respect to a different background constraint, dropping the FOSD restrictions in (2) and (3) while keeping the range restriction in (1). We would expect the restrictiveness of CPT to increase in this case, since (for all parameter values) CPT obeys first-order stochastic dominance. However, the restrictiveness of CPT relative to this larger conceivable set, 0.33, is not significantly higher than 0.32, the restrictiveness of CPT under the main specification for $\mathcal{F}_{\mathcal{M}}$.[28] This reinforces our finding that CPT is not very restrictive on the domain of binary lotteries.

5.6 Cross-Context Variation

Just as a model's parameter estimates and fit are understood to potentially vary across subject populations and settings, our measures of completeness and restrictiveness depend on the underlying context. Here we consider the performance of CPT(α,ζ,γ) on lotteries from two alternative contexts: binary lotteries over the loss domain, and lotteries over three outcomes.

[27] The variation in restrictiveness is bounded by the total variation distance between the primitive choices of μ (see (4)), but it can be difficult to compute the total variation distance between complex choices of μ.

[28] Normalization plays an important role here: CPT'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. CPT's *relative* performance compared to the naive benchmark is very similar regardless of whether we impose FOSD or not.

First we consider 25 binary lotteries over the loss domain from Bruhin et al. (2010). On this data, the 3-parameter specification of CPT indexed to (β, γ, ζ) predicts the certainty equivalent

$$v^{-1}\left((1-w(1-p)) \cdot v(\overline{z}) + w(1-p) \cdot v(\underline{z})\right)$$

for each lottery $(\overline{z}, \underline{z}, p)$, where $v(z) = -((-z)^{\beta})$ and $w(p) = (\zeta p^{\gamma})/(\zeta p^{\gamma} + (1-p)^{\gamma})$.

Table 2 reports completeness and restrictiveness measures for the four specifications of CPT that have appeared in the prior literature. The results are very similar to what we found for the gain domain.

	Completeness κ^*	Restrictiveness r
CPT(β, γ, ζ)	0.92	0.34
	(0.03)	(0.02)
CPT(β, γ)	0.67	0.47
	(0.08)	(0.02)
CPT(ζ, γ)	0.92	0.36
	(0.02)	(0.01)
CPT(β)	<0.01	0.83
	(0.09)	(0.03)

Table 2: Completeness and restrictiveness are reported for the new data set. Completeness is estimated from 8906 observations; restrictiveness is estimated from 100 simulations.

Next we evaluate the restrictiveness of CPT(α, ζ, γ) on a set of 18 three-outcome gain-domain lotteries from Bernheim and Sprenger (2020) (see Appendix A.4). For each lottery $(z_1, z_2, z_3; p_1, p_2, p_3)$, where $z_1 \geq z_2 \geq z_3 \geq 0$, the predicted certainty equivalent is

$$v^{-1}\big(v(z_1) + w(p_2+p_3)(v(z_2) - v(z_1)) + w(p_3)(v(z_3) - v(z_2))\big)$$

where v and w have the same functional forms as used above..

A predictive mapping f is a map from these 18 lotteries into average certainty equivalents. The set of conceivable mappings $\mathcal{F}_M$ is again defined to satisfy: (1) each certainty equivalent has to be in the range of the lottery outcomes, and (2) if a lottery first-order stochastically dominates another, then its certainty equivalent must be higher. We generate 100 random mappings from a uniform distribution over mappings satisfying these properties.

Below, we compare the distribution of f-discrepancies from Figure 3 with the distribution of f-discrepancies that we find for these three-outcome lotteries.

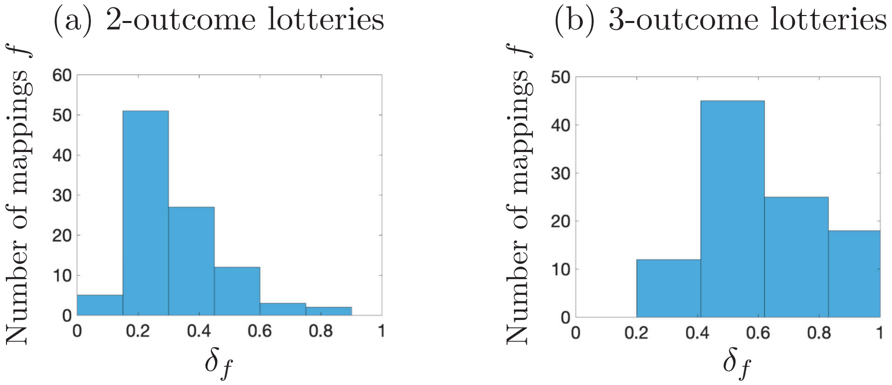


Figure 3: *Left:* Comparison of distribution of f-discrepancies.

The restrictiveness of CPT on this set of three-outcome lotteries is 0.63, with a standard error of 0.02. Thus CPT is about twice as restrictive on three-outcome lotteries than on binary lotteries. Besides imposing FOSD, CPT imposes the property of "rank dependence" for lotteries with more than two outcomes, which means roughly that the probability weighting function is applied to increments in utility and not to utility itself.[29] We view the increase in restrictiveness as a quantification of the additional constraints implied by this property.

6 Application 2: The Distribution of Initial Play

6.1 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 matrix games from Fuden-

[29] See Strzalecki (2020) for a discussion of the rank-dependence property of CPT and of the related RDEU model (Quiggin, 1982).

berg and Liang (2019),[30] each described as a vector in $\mathbb{R}^{18}$. The outcome space is $\mathcal{Y} = \{a_1, a_2, a_3\}$ (the set of row player actions) and the analyst seeks to predict the conditional distribution over $\mathcal{Y}$ for each game, interpreted as choices made by a population of subjects for the same game. Thus, $\mathcal{S} = \Delta(\mathcal{Y})$, the set of all distributions over row player actions. A mapping for this problem is any function $f : \mathcal{X} \to \mathcal{S}$ taking the 466 games into predicted distributions of play. For any two mappings f and f', let $d(f, f')$ be the expected Kullback-Liebler divergence between the predicted distributions, as in (2).

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

$$p_k(h, \tau) = \frac{\pi_\tau(h)}{\sum_{l=0}^{k-1} \pi_\tau(l)} \qquad \forall\ h \in \mathbb{N}_{<k} \tag{9}$$

over (lower) opponent levels, where π_τ is the Poisson distribution with rate parameter $\tau \geq 0$. The parameter τ is the only free parameter of the model, and the naive mapping is nested as $\tau = 0$.

The *Logit Level-1* model has a single free parameter $\lambda \geq 0$. Let $\bar{u}(a_i)$ be the expected payoff of a_i when the column player uses a uniform distribution. Then for each action a_i, the predicted frequency with which a_i is played is

$$\frac{\exp\left(\lambda \cdot \bar{u}(a_i)\right)}{\sum_{i=1}^{3} \exp\left(\lambda \cdot \bar{u}(a_i)\right)}.$$

[30] This data includes a meta data-set of experimental data aggregated in Wright and Leyton-Brown (2014) from 86 games from six experimental game theory papers, in addition to Mechanical Turk data from 200 games with randomly chosen payoff matrices and 180 games whose payoff matrices were algorithmically designed. The games from past papers on average had more pure strategy Nash equilibria and more actions that are pure-strategy rationalizable than the games from previous studies; in the algorithmic games it was even more common for all 3 actions to be rationalizable.

This model nests prediction of uniform play (our naive rule) as $\lambda = 0$, and predicts a degenerate distribution on the level-1 action when λ is sufficiently large.

Finally, *Logit PCHM* (see e.g. Wright and Leyton-Brown (2014)) replaces the assumption of exact maximization in the PCHM with a logit best response. This model has two free parameters: $\lambda, \tau \in \mathbb{R}_+$. The level-0 player chooses $g_0 = (1/3, 1/3, 1/3)$, as in the PCHM. Recursively define for each $k \geq 1$

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

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 defined in (9), and define

$$g_k(a_i) = \frac{\exp(\lambda \cdot v_k(a_i))}{\sum_{j=1}^{3} \exp(\lambda \cdot v_k(a_j))}$$

to be the distribution of level-k play. We aggregate across levels using a Poisson distribution with rate parameter τ.

Finally, we use the uniform distribution $f_{\text{naive}}(x) = (1/3, 1/3, 1/3)$ as the naive prediction for every game x.

6.2 Completeness

The models PCHM, Logit Level-1, and Logit PCHM are 43.6%, 72.7%, and 72.9% complete on the actual data. (Equivalently, their f^*-discrepancies are 0.564, 0.273, and 0.271.) 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. Fudenberg and Liang (2019) found that the Level-1 model provides a good prediction of the modal action, but it is harder to predict the full distribution of play, so it is not obvious from the previous result that Logit Level-1 would perform so well here. The fact that it did, combined with its strong performance for predicting modal play, suggests that initial play in many of these experiments is rather unstrategic.[31]

6.3 Restrictiveness

We turn now to evaluating the restrictiveness for these models. Compared to the case of preferences over binary lotteries, economic theory provides very little in the way of a priori restrictions on initial play.[32] We thus define the conceivable set $\mathcal{F}_M$ to include all mappings satisfying the following very weak conditions:[33]

1. If an action is strictly dominated, then the frequency with which it is chosen does not exceed 1/3.
2. If an action is strictly dominant, then the frequency with which it is chosen is at least 1/3.

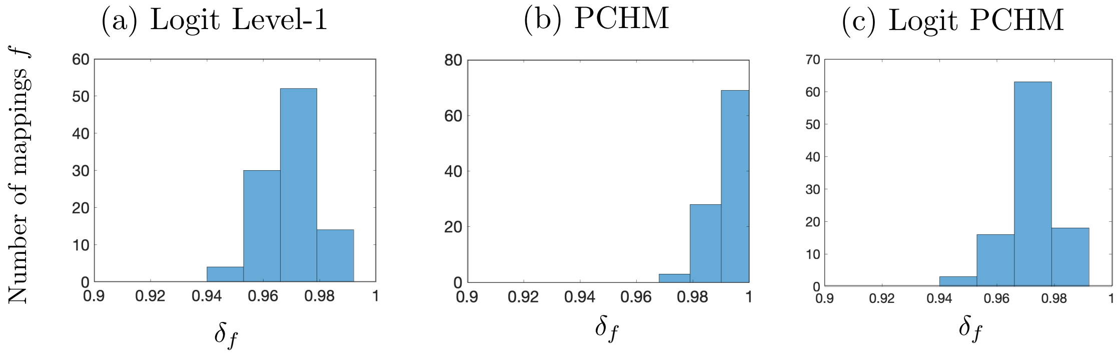
For each of the PCHM, Logit Level-1, and Logit PCHM models, we generate 100 mappings f from a uniform distribution over the set of conceivable mappings $\mathcal{F}_{\mathcal{M}}$, and evaluate the f-discrepancies with respect to these mappings.[34] The distributions of f-discrepancies are shown in the figure below.

[31] Fudenberg and Liang (2019) found that modal play in some sorts of games is better described by equilibrium notions than level-1. Since such regularities cannot be accommodated by the logit level-1 model, these may explain the gap between the completeness of logit level-1 and full completeness. Costa-Gomes et al. (2001) find a sizable fraction of level-2 players in their experimental data, which may further help to explain this gap.

[32] Classic game theory suggests that dominant strategies should have probability 1 and dominated strategies have probability 0, but this is inconsistent with our data (and most experimental data of play in games).

[33] In the actual data, the median frequency for 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.

[34] To define the uniform distribution over $\mathcal{F}_M$ we embed it in $[0,1]^{466\times 3}$.



(a) Logit Level-1

(b) PCHM

(c) Logit PCHM

Figure 4: Distribution of f-discrepancies for the three models.

These models are very restrictive: Logit Level-1's restrictiveness is 0.969, PCHM's restrictiveness is 0.993, and Logit PCHM's restrictiveness is 0.972. Indeed, across all of these mappings and models, the f-discrepancy is always at least 0.943, so the completenesses of these models across the simulated mappings is bounded above by 0.057. Since the completeness of these models on the actual data ranged from 0.436 to 0.729, these models are much better predictors of the real data than of the hypothetical data sets.

Figure 5 plots completeness and restrictiveness measures for all three models (see also Table B.1 in the appendix). Compared to PCHM, Logit Level-1 and Logit PCHM are substantially more complete (compare $\kappa^* = 0.436$ to $\kappa^* = 0.727$ and $\kappa^* = 0.729$) and only slightly less restrictive (compare $r = 0.993$ to $r = 0.969$ and $r = 0.972$). Logit Level-1 and Logit PCHM are almost identical in terms of completeness and restrictiveness, even though the parametric forms of the two models do not appear related.[35] In the subsequent Section 7, we investigate the relationship between Logit Level-1 and Logit PCHM further by studying the correlation in their errors.

One takeaway from these results is that each of the models very precisely captures the observed regularities in actual play. A second takeaway is that, in contrast to the certainty equivalent setting, the functional forms of PCHM, Logit Level-1, and Logit PCHM imply substantial restrictions beyond those of the background restrictions. (In Appendix B.1, we show that increasing the strength of the constraints on the proba-

[35] No value of τ in the PCHM yields the Level-1 model, so Logit Level-1 is not nested within Logit PCHM.

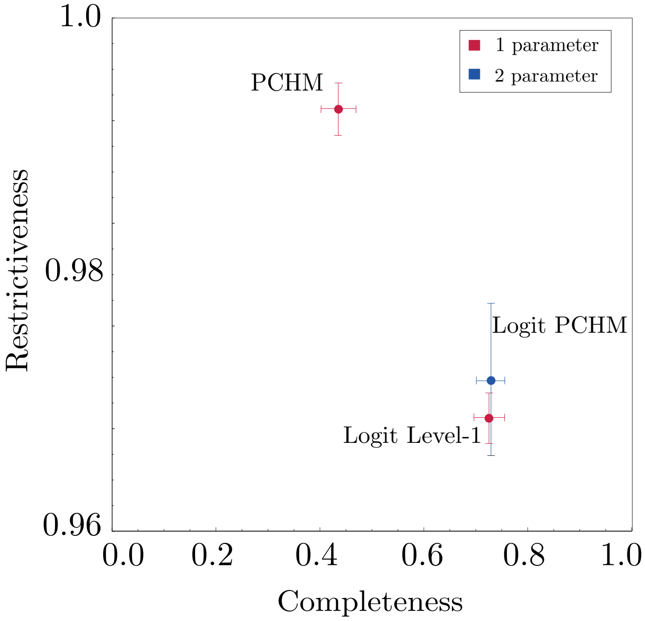


Figure 5: Comparison of models by their completeness and restrictiveness.

bilities of the strictly dominated and strictly dominant actions does not appreciably change estimated restrictiveness.) This is not so surprising, because the background constraints we use here are quite weak, but it shows that we do not yet understand yet which basic properties are implied by and captured in the parametric models PCHM, Logit Level-1, and Logit PCHM. In particular, we do not understand why Logit Level-1 and Logit PCHM entail the same empirical content, or what empirical content that is. This suggests that there is room still for a better understanding of what governs initial play.

7 How Different are Two Models?

The restrictivenesses of two parametric models allows us to compare how many behaviors are ruled out by each model, but does not tell us whether the two models rule out similar kinds of behaviors. For example, consider DA(η)—which achieves a restrictiveness of 0.68—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(η) 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 $f_{\text{naive}} \in \mathcal{F}_{\Theta_1}, \mathcal{F}_{\Theta_2}$. For an arbitrary mapping f, define

$$\delta_f^{\Theta_1} := \frac{d(\mathcal{F}_{\Theta_1}, f)}{d(f_{naive}, f)} \qquad \delta_f^{\Theta_2} := \frac{d(\mathcal{F}_{\Theta_2}, f)}{d(f_{naive}, f)}$$

to be the f-discrepancies of the respective models. As in the definition of restrictiveness, let μ be a primitive distribution over mappings in $\mathcal{F}_{\mathcal{M}}$.

Definition 4. The *δ-correlation* between models $\mathcal{F}_{\Theta_1}$ and $\mathcal{F}_{\Theta_2}$ is the correlation coefficient for the pair $(\delta_f^{\Theta_1}, \delta_f^{\Theta_2})$ where $f \sim \mu$.

Two models with a high δ-correlation do relatively well on the same mappings, while the δ-correlation is negative if the mappings that one model approximates well are relatively harder for the other to approximate. The size of δ-correlation between two models does not directly imply anything about their absolute levels of restrictiveness. [36] The size of δ-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 δ-correlation measure will not reveal which of the two models is more flexible. The measure of δ-correlation, however, can be usefully paired with the restrictiveness of the two models to provide further insight into their comparison, as we now demonstrate.

7.1 Certainty Equivalents

Table 3 reports the δ-correlation between the models CPT-γ, CPT-(ζ, γ), CPT-(α, ζ, γ), DA(η), and DA(α, η).

[36] For example, the models $\mathcal{F}_{\Theta_1} \equiv \mathcal{F}_{\mathcal{M}}$ and $\mathcal{F}_{\Theta_2} \equiv \mathcal{F}_{\mathcal{M}}$ have a δ-correlation of 1, as do the models $\mathcal{F}'_{\Theta_1} \equiv \{f_{naive}\}$ and $\mathcal{F}'_{\Theta_2} \equiv \{f_{naive}\}$. But the first pair of models is maximally unrestrictive while the second is maximally restrictive.

	CPT(γ)	CPT(ζ, γ)	CPT(α, ζ, γ)	DA(η)	DA(α, η)
CPT(γ)	1	0.38	0.26	-0.76	0.40
CPT(ζ, γ)	-	1	0.99	0.37	0.43
CPT-(α, ζ, γ)	-	-	1	0.48	0.40
DA(η)	-	-	-	1	0.47
DA(α, η)	-	-	-	-	1

Table 3: δ-correlation between various pairs of models

CPT(ζ, γ) and CPT(α, ζ, γ) are nearly perfectly correlated. Since the two models have similar absolute levels of restrictiveness ($r = 0.32$ for CPT(α, ζ, γ) and $r = 0.37$ for CPT(ζ, γ)), this suggests that the two models rule out essentially the same behavior.

The two models DA(η) and CPT(γ) also have similar absolute levels of restrictiveness ($r = 0.68$ for DA(η) and $r = 0.59$ for CPT(γ)). But their δ-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(γ) and DA(α, η) is not substantially larger, but the δ-correlation between these models rises to 0.40, suggesting that introduction of the α parameter in addition to the η parameter in DA re-directs the model's predictions in the direction of CPT(γ).

The remaining δ-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.

7.2 Initial Play

Table 4 compares the δ-correlation between models PCHM, Logit Level-1, and Logit PCHM.

	PCHM	Logit Level-1	Logit PCHM
PCHM	1	0.67	0.77
Logit Level-1	-	1	0.94
Logit PCHM	-	-	1

Table 4: 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.

8 Application to General Prediction Problems

In the two leading cases we have analyzed in the main text (Section 3.1), the function d is derived from a primitive loss function l. We call the general property that permits this *decomposability.*[37]

Definition 5 (Decomposablity). Consider an arbitrary loss function $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y}$ and define $e_{P^*}(f) := \mathbb{E}_{P^*}[l(f, (X, Y))]$ to be the expected loss of mapping f. For any distribution P, let $f_P = \min_{f \in \mathcal{F}} e_P(f)$ denote the error-minimizing mapping under that distribution. Say that the problem is *decomposable* if there exists a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ such that

$$d(f, f_P) = e_P(f) - e_P(f_P) \tag{10}$$

[37] Decomposability looks similar to the coupling of the "cost of uncertainty" and the "value of information" in Frankel and Kamenica (2019), which is also satisfied in the cases of predicting conditional expectations with loss mean squared error, and predicting conditional distributions with loss equal to KL divergence. But they are primarily concerned with comparisons of different signal structures, while we compare model classes. Also, we do not consider a Bayesian problem, so it is not clear how to relate their axiom of order invariance to our setting.

for every distribution P (with fixed marginal distribution P^*_X). That is, $d(f, f_P)$ is the difference between the error of mapping f and the error of the best mapping f_P.

Note that when predicting the whole conditional distribution, i.e., $\mathcal{S} = \Delta(Y)$, decomposability holds for any loss function, since we can define $d(f, f_P) := e_{f_P}(f) - e_{f_P}(f_P)$.[38] But in general, prediction problems need not be decomposable. For example, suppose the objective is to predict the conditional median, and the loss function is $l(f, (x, y)) = |y - f(x)|$ instead of squared loss. The expected error is then $e_{P^*}(f) = \mathbb{E}_{P^*}|Y - f(X)|$, and the error-minimizing function f^* takes each x into the median value of Y at x. We might want to use

$$d(f, f') = \mathbb{E}_{P^*}(|f(X) - f'(X)|) \tag{11}$$

as a measure of how different the predictions are under f and f', but this function does not satisfy (10). For the absolute value loss function, there is in fact no function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that satisfies (10), because the difference in errors cannot be determined from f and f^* alone, but depends on further properties of the conditional distribution P^*. (See Appendix E.2 for more details.)

When the problem is decomposable, as in the cases analyzed in the main text, then our approach is applicable without change by setting d to be the function satisfying (10). If the problem is not decomposable, we take d as a primitive, rather than deriving it from the loss function l. The key concepts of f-discrepancies and restrictiveness are defined as above using this primitive d. What we lose is the equivalence between the $1 - \delta_{f^*}$ and completeness κ^*, as described in (3). One can report restrictiveness r (based on the primitive d) and completeness κ^* (based on the primitive l), understanding that there is no inherent relationship between these concepts. Larger values of r and κ^* can still be interpreted as more restrictive and more complete models. A second alternative is to report $1 - \delta_{f^*}$ instead of completeness. Since δ_{f^*} is derived from d, this second approach does not require specification of a loss function at all. A new estimation procedure for δ_{f^*} is needed, however, as our approach in Section 4.2 makes use of the relationship $\delta_{f^*} = 1 - \kappa^*$. We provide an alternative estimator

[38] This is because P is completely pinned down by f_P given P^*_X, so $e_P = e_{f_P}$.

for δ_{f^*} in Appendix E.1 for this purpose.

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 restrictiveness of a theory, and demonstrate that it reveals new insights into models from two economic domains. The method is easily applied to models from different domains.[39]

References

Anderson-Sprecher, R. (1994): "Model comparisons and R 2," *The American Statistician*, 48, 113–117.

Andrews, I. and M. Kasy (2019): "Identification of and Correction for Publication Bias," *American Economic Review*, 109, 2766–2794.

Austern, M. and W. Zhou (2020): "Asymptotics of Cross-Validation," *arXiv preprint arXiv:2001.11111.*

Banerjee, A., S. Chassang, S. Montero, and E. Snowberg (2020): "A Theory of Experimenters: Robustness,Randomization, and Balance," *American Economic Review*, 110, 1206–30.

Barseghyan, L., F. Molinari, T. O'Donoghue, and J. C. Teitelbaum (2013): "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103, 2499–2529.

Basu, P. and F. Echenique (2020): "On the falsifiability and learnability of decision theories," *Theoretical Economics*, forthcoming.

Beatty, T. and I. Crawford (2011): "How Demanding Is the Revealed Preference Approach to Demand?" *American Economic Review*, 101, 2782–95.

Bernheim, D. and C. Sprenger (2020): "Direct Tests of Cumulative Prospect Theory," Working Paper.

Bronars, S. (1987): "The Power of Nonparametric Tests of Preference Maximization," *Econometrica*, 55, 693–698.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica*, 78, 1375–1412.

[39] For example, to measure the restrictiveness of rational aggregate demand, one could generate random demand functions on a finite collection of budget sets, and compute the "distance" between these functions and one that satisfies GARP. (We thank Tilman Börgers for this suggestion.)

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

Chassang, S., G. P. i Miquel, and E. Snowberg (2012): "Selective Trials: A Principal-Agent Approach toRandomized Controlled Experiments," *American Economic Review*, 102, 1279–1309.

Chemla, G. and C. Hennessy (2019): "Controls, belief updating, and bias in medical RCTs," *Journal of Economic Theory*, 184.

Chen, X. and A. Santos (2018): "Overidentification in regular models," *Econometrica*, 86, 1771–1817.

Choi, S., R. Fisman, D. Gale, and S. Kariv (2007): "Consistency and Heterogeneity of Individual Behavior under Uncertainty," *American Economic Review*, 97, 1–15.

Costa-Gomes, M., V. P. Crawford, and B. Broseta (2001): "Cognition and behavior in normal-form games: An experimental study," *Econometrica*, 69, 1193–1235.

Cox, D. R. (1961): "Tests of separate families of hypotheses," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 105–123.

——— (1962): "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 406–424.

de Quidt, J., J. Haushofer, and C. Roth (2018): "Measuring and Bounding Experimenter Demand," *American Economic Review*, 108, 3266–3302.

DellaVigna, S. and E. Linos (2020): "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," .

DellaVigna, S. and D. Pope (2019): "Stability of Experimental Results: Forecasts and Evidence," .

Fehr-Duda, H. and T. Epper (2012): "Probability and Risk: Foundations and Economic Implication of Probability-Dependent Risk Preferences," *Annual Review of Economics*, 4, 567–593.

Frankel, A. and E. Kamenica (2019): "Quantifying information and uncertainty," *American Economic Review*, 109, 3650–80.

Frankel, A. and M. Kasy (2019): "Which findings should be published?" Working Paper.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2019): "Measuring the Completeness of Theories," Working Paper.

Fudenberg, D. and D. Levine (2020): "Learning in Games and the Interpretation of NaturalExperiments," *American Economic Journal: Microeconomics*.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–4141.

Gabaix, X. and D. Laibson (2008): "The Seven Properties of Good Models," in *The Methodologies of Modern Economics: Foundations of Positive and Normative Economics*.

Goldreich, O. and S. Vadhan (2007): "Special issue on worst-case versus average-case complexity editors' foreword," *Computational Complexity*, 16, 325–330.

Goldstein, W. M. and H. J. Einhorn (1987): "Expression theory and the preference reversal phenomena," *Psychological review*, 94, 236–254.

Gul, F. (1991): "A Theory of Disappointment Aversion," *Econometrica*, 59, 667–686.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Harless, D. and C. Camerer (1994): "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica*, 62, 1251–1289.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251–1271.

Hey, J. D. (1998): "An application of Selten's measure of predictive success," *Mathematical Social Sciences*, 35, 1–15.

Karmarkar, U. (1978): "Subjectively weighted utility: A descriptive extension of the expected utility model," *Organizational Behavior & Human Performance*, 21, 67–72.

Koopmans, T. and O. Reiersol (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21, 165–181.

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

McFadden, D. (1974): "The measurement of urban travel demand," *Journal of Public Economics*, 3, 303–328.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1313–1326.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 1349–1382.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Peysakhovich, A. and J. Naecker (2017): "Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity," *Journal of Economic Behavior and Organization*, 133, 373–384.

Polisson, M., J. K.-H. Quah, and L. Renou (2020): "Revealed Preferences over Risk and Uncertainty," *American Economic Review*, 110, 1782–1820.

Quiggin, J. (1982): "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization*, 3, 323–343.

Routledge, B. R. and S. E. Zin (2010): "Generalized disappointment aversion and asset prices," *The Journal of Finance*, 65, 1303–1332.

Sargan, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26, 393–415.

Selten, R. (1991): "Properties for a Measure of Predictive Success," *Mathematical Social Sciences*, 21, 153–167.

Shmaya, E. and L. Yariv (2016): "Experiments on Decisions under Uncertainty: A Theoretical Framework," *American Economic Review*, 106, 1775–1801.

Snowberg, E. and J. Wolfers (2010): "Explaining the Favorite-Long Shot Bias: Is It Risk-Love or Misperceptions?" *Journal of Political Economy*, 118, 723–746.

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

——— (1995): "On players' models of other players: Theory and experimental evidence," *Games and Economic Behavior*, 10, 218–254.

Strzalecki, T. (2020): *Decision Theory.*

Tversky, A. and D. Kahneman (1992): "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323.

Van Der Vaart, A. W. and J. A. Wellner (1996): "Weak convergence," in *Weak Convergence and Empirical Processes*, Springer, 16–28.

Varian, H. (1982): "The Nonparametric Approach to Demand Analysis," *Econometrica*, 50, 945–973.

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

Yaari, M. (1987): "The Dual Theory of Choice under Risk," *Econometrica*, 55, 95–115.

A Supplementary Material for Application 1

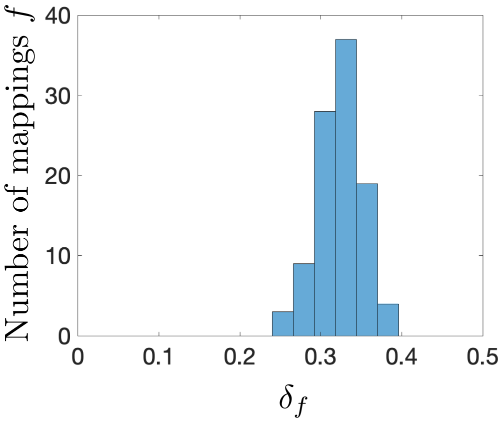
A.1 Gains Domain

	Completeness κ^*	Restrictiveness r
CPT(α, ζ, γ)	0.95	0.32
	(0.02)	(0.02)
CPT(α, γ)	0.84	0.54
	(0.02)	(0.01)
CPT(α, ζ)	0.27	0.51
	(0.06)	(0.02)
CPT(ζ, γ)	0.91	0.37
	(0.02)	(0.01)
EU(α)/CPT(α)/DA(α)	0.25	0.90
	(0.06)	(0.01)
CPT(γ)	0.66	0.59
	(0.07)	(0.02)
CPT(ζ)	0.27	0.64
	(0.06)	(0.01)
DA(α, η)	0.27	0.46
	(0.06)	(0.02)
DA(η)	0.27	0.68
	(0.06)	(0.02)

Table 5: Completeness κ^* and restrictiveness r for each model in the certainty equivalent setting. Completeness is estimated from 8906 observations; restrictiveness is estimated from 100 simulations.

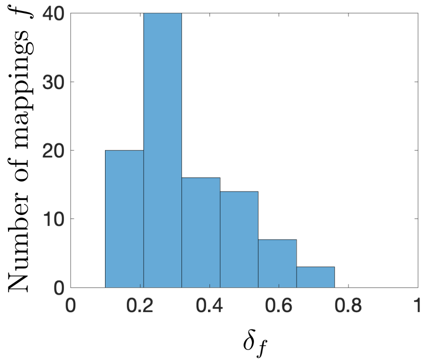
A.2 Supplementary Material to Section 5.5

Different specification of μ. We sample 100 pairs (a, b) from a uniform distribution over $[0.9, 1.1] \times [0.9, 1.1]$. For each (a, b) pair, we generate 100 mappings from a beta(a, b) distribution and evaluate restrictiveness of CPT with respect to these mappings. The figure below reports the histogram of restrictiveness values across these different beta distributions.



The estimate of restrictiveness varies very little across these different choices of μ. The max restrictiveness value, 0.40, and the min restrictiveness value, 0.25, are comparable to the restrictiveness value, 0.32, which we find for the uniform distribution.

Alternative specification of conceivable set $\mathcal{F}_{\mathcal{M}}$. Consider the alternative conceivable set of mappings consisting of all functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $f(\overline{z}, \underline{z}, p) \in [\underline{z}, \overline{z}]$. We sample 100 times from a uniform distribution over this set and report the distribution of f-discrepancies in the figure below:



The average discrepancy, 0.33, tells us that the model is more restrictive on this expanded domain of mappings, but not substantially so.[40]

[40] Even though the errors are substantially higher than when we require the conceivable mappings to respect FOSD, the estimated restrictiveness is almost the same because the naive error also increases. Specifically, the mean naive error is 343.32 (compared to 178.73 under the original $\mathcal{F}_M$), while the mean CPT error is 110.73 (compared to 58.21 under the original $\mathcal{F}_M$).

A.3 Parameter Estimates

We report below the estimated parameters for each of the models that we consider. In the first column, we report the estimated parameters on the actual data. In the second, we report the average parameter estimates for across our generated mappings.

CPT Parameters	Real Data	Generated Mappings
α, ζ, γ	(0.77,1.01,0.50)	(0.65,3.95,0.43)
α, γ	(0.77,0.50)	(0.91,0.27)
α, ζ	(1.19,0.52)	(0.43,17.74)
ζ, γ	(0.70,0.50)	(1.80,0.42)
α	0.77	0.99
γ	0.50	0.25
ζ	0.68	5.15

DA Parameters	Real Data	Generated Mappings
α, η	(1,0.47)	(0.35,-0.77)
η	0.47	-0.20

	Real Data	Generated Mappings
PCHM	$\tau = 0.5$	$\tau = 0.1$
logit level-1	$\lambda = 0.02$	$\lambda = 0.0018$
logit PCHM	$(\tau, \lambda) = (1.4, 0.11)$	$(\tau, \lambda) = (1.05, 0.02)$

A.4 Three-Outcome Lotteries

We report below the 18 three-outcome lotteries from Bernheim and Sprenger (2020) that we used in Section 5.6:

z_1	z_2	z_3	p_1	p_2	p_3
34	24	18	0.1	0.3	0.6
34	24	18	0.4	0.3	0.3
34	24	18	0.6	0.3	0.1
32	24	18	0.1	0.3	0.6
32	24	18	0.4	0.3	0.3
32	24	18	0.6	0.3	0.1
30	24	18	0.1	0.3	0.6
30	24	18	0.4	0.3	0.3
30	24	18	0.6	0.3	0.1
24	23	18	0.3	0.1	0.6
24	23	18	0.3	0.4	0.3
24	23	18	0.3	0.6	0.1
24	21	18	0.3	0.1	0.6
24	21	18	0.3	0.4	0.3
24	21	18	0.3	0.6	0.1
24	19	18	0.3	0.1	0.6
24	19	18	0.3	0.4	0.3
24	19	18	0.3	0.6	0.1

A.5 Heterogeneous Risk Preferences

Our analysis in the main text considered representative agent models. In some cases, the analyst may have auxiliary data on the subjects that can be used to improve predictions. We show now how completeness and restrictiveness can be evaluated in this case.

Specifically, we return to our first application and group subjects into three clusters identified by Bruhin et al. (2010). We fit CPT for each cluster, allowing parameter values to vary across groups. Table A.5 reports completeness measures cluster by cluster.

The performance of the naive expected value rule, the best achievable performance, and the performance of CPT, all vary substantially across clusters. For example, the behavior of subjects in cluster 1 is roughly consistent with expected value (the error of the naive rule is 39.90), while the behavior of subjects in cluster 3 departs substantially from this benchmark (the error of the naive rule is 99.94). The best achievable prediction for these groups of subjects is also very different (ranging from

	Cluster 1	Cluster 2	Cluster 3
Naive	39.90	150.10	99.94
	(4.98)	(7.24)	(7.97)
CPT	30.74	43.87	69.62
	(7.25)	(4.72)	(8.50)
Best Achievable Error	29.59	36.30	67.05
	(7.36)	(3.34)	(8.02)
Completeness	0.98	0.88	0.92
	(0.02)	(0.03)	(0.03)
N	674	1144	2641

Table 6: Completeness measures for each of three subject clusters.

29.59 to 67.05), as is the completeness of CPT (ranging from 30.74 to 69.62).

The average completeness, weighted by proportion of observations in each cluster, is 0.91, which is very close to what we found for the representative agent model. This may seem surprising at first, since allowing for parameters to vary across subjects improves the accuracy of predictions. But the best mapping from the extended feature space $\mathcal{X}' = \mathcal{X} \times \{1, 2, 3\}$ to $\mathcal{Y}$ is more predictive than the best mapping considered previously. Thus what we find is that the completeness of CPT with three clusters, *relative to the best three-cluster mapping*, is comparable to the completeness of the representative-agent version of CPT, *relative to the best representative-agent mapping.*

Similarly, when measuring restrictiveness, we extend the set of conceivable mappings to the domain $\mathcal{X}'$. Each generated pattern of behavior is thus a triple (f_1, f_2, f_3) of mappings from the original $\mathcal{F}_{\mathcal{M}}$. We ask how well these tuples can be approximated using mappings (g_1, g_2, g_3) from CPT. It is straightforward to see that the restrictiveness of the three-cluster CPT is identical to the restrictiveness of the representative-agent model.[41]

B Supplementary Material for Application 2

Table B.1 summarizes completeness and restrictiveness measures for all three models.

[41] Note that this is true for any number of exogenously specified clusters.

	Completeness κ^*	Restrictiveness r
PCHM	0.436	0.993
	(0.017)	(<0.001)
logit level-1	0.727	0.969
	(0.015)	(<0.001)
logit PCHM	0.729	0.972
	(0.014)	(0.003)

Table 7: Completeness and restrictiveness measures for each model of initial play. Completeness is estimated from 21,393 observations; restrictiveness is estimated from 100 simulations.

B.1 Varying the Conceivable Set $\mathcal{F}_{\mathcal{M}}$

We consider here a strengthening of the background constraints imposed on the conceivable set $\mathcal{F}_{\mathcal{M}}$. For each $t \in [0, 0.3)$, define the conceivable class of mappings $\mathcal{F}_{\mathcal{M}}(t)$ to be those satisfying 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 the specification of $\mathcal{F}_{\mathcal{M}}$ used in the main text. We find that across choices of $t \in [0, 0.3)$, the restrictiveness of PCHM, Logit PCHM, and Logit Level-1 do not fall below 0.89 (see Table 8 below, which reports the max and min restrictiveness measures). This suggests that the high restrictiveness of these models is rather robust to constraints imposed only on strictly dominant and strictly dominated actions.

	PCHM	Logit Level-1	Logit PCHM
max	0.993	0.969	0.972
min	0.974	0.890	0.957

Table 8: Largest and smallest restrictiveness measures as t varies over $[0, 0.3)$.

C Supplementary Material to Section 3.1

We now demonstrate the relationships in (1) and (2).

Mean-Squared Error. Suppose $S = \mathcal{Y} = \mathbb{R}$ and the loss function is $l(f,(x,y)) = (y - f(x))^2$. The following decomposition is standard:

$$\begin{aligned} e_{P^*}(f) &:= \mathbb{E}_{P^*}\left[(Y - f(X))^2\right] \\ &= \mathbb{E}_{P^*}\left[(Y - f^*(X))^2\right] + \mathbb{E}_{P^*}\left[(f(X) - f^*(X))^2\right] = e_{P^*}(f^*) + d(f, f^*) \end{aligned}$$

Negative Log-Likelihood. Suppose $S = \Delta(\mathcal{Y})$ where $\mathcal{Y}$ is a finite set, and the loss function is $l(f,(x,y)) = -\log f(y \mid x)$ for any mapping $f : \mathcal{X} \to S$. Then,

$$\begin{aligned} d(f, f^*) &= \sum_{x \in \mathcal{X}} f^*(x) \sum_{y \in \mathcal{Y}} f^*(y|\, x) \log\left(\frac{f^*(y|\, x)}{f(y|\, x)}\right) \\ &= \mathbb{E}_{P^*}\left[\log f^*(y|\, x)\right] - \mathbb{E}_{P^*}\left[\log f(y|\, x)\right] = -e_{P^*}(f^*) + e_{P^*}(f). \end{aligned}$$

So $e_{P^*}(f) = e_{P^*}(f^*) + d(f, f^*)$ as desired.

D Estimation of Completeness κ^*

D.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.

D.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 $\mathcal{F}$, we define the conditional out-of-sample error (given data set $\mathbf{Z}_N$) as

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

where $\hat{f}_{\mathbf{Z}_N} \in \mathcal{F}$ is an estimator, or an algorithm, that selects a mapping $\hat{f}_{\mathbf{Z}_N}$ within the model $\mathcal{F}$ 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

$$e_{\mathcal{F},N} := \mathbb{E}\left[e_{\mathcal{F}}\left(\mathbf{Z}_N\right)\right].$$

From the definition of the K-fold cross-validation estimator, it can be easily shown that $\mathbb{E}\left[CV\left(\mathcal{F}\right)\right] = e_{\mathcal{F},\frac{K-1}{K}N}$. As a result, the asymptotic distribution of $CV\left(\mathcal{F}\right) - e_{\mathcal{F},\frac{K-1}{K}N}$ 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 $CV\left(\mathcal{F}\right) - e_{\mathcal{F},\frac{K-1}{K}N}$.

D.1.2 Joint Parametrization of $\mathcal{F}_\Theta$ and $\mathcal{F}_\mathcal{M}$

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

For arbitrary parameters θ and β, write

$$l_\Theta\left(\theta, Z_i\right) := l\left(f_\theta, Z_i\right), \quad l_\mathcal{B}\left(\beta, Z_i\right) := l\left(f_{[\beta]}, Z_i\right).$$

We define the estimation mappings in $\mathcal{F}_\Theta$ and $\mathcal{F}_\mathcal{M}$ by

$$\begin{aligned}\hat{\theta}\left(\mathbf{Z}_N\right) &:= \arg\min_{\theta\in\Theta} \frac{1}{N}\sum l_\Theta\left(\theta, Z_i\right),\\ \hat{\beta}\left(\mathbf{Z}_N\right) &:= \arg\min_{\beta\in\mathcal{B}_\mathcal{M}} \frac{1}{N}\sum l_\mathcal{B}\left(\beta, Z_i\right).\end{aligned}$$

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

$$\begin{aligned}\hat{\alpha}\left(\mathbf{Z}_N\right) &:= \left(\hat{\theta}'\left(\mathbf{Z}_N\right), \hat{\beta}'\left(\mathbf{Z}_N\right)\right)'\\ &= \arg\min_{\theta\in\Theta,\beta\in\mathcal{B}_\mathcal{M}} \frac{1}{N}\sum_{i=1}^{N}\left[l_\Theta\left(\theta, Z_i\right) + l_\mathcal{B}\left(\beta, Z_i\right)\right],\end{aligned}$$

to be an estimator for α^*. Finally, define

$$\Delta l(\theta, \beta; Z_i) := l(f_\theta, Z_i) - l\left(f_{[\beta]}, Z_i\right) = l_\Theta(\theta, Z_i) - l_\mathcal{B}(\beta, Z_i).$$

D.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 $\mathcal{F}_\Theta$ and $\mathcal{F}$, respectively. The difference in their test errors on observation Z_i is

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

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

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

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

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

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

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

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

$$\hat{\sigma}^2_{\hat{\kappa}^*} := \frac{\hat{\sigma}^2_{\Delta_\theta} - 2\hat{\kappa}^* \hat{\sigma}_{\Delta_\theta \Delta_{\text{naive}}} + \hat{\kappa}^{*2} \hat{\sigma}^2_{\Delta_{\text{naive}}}}{\left[CV(f_{\text{naive}}) - CV(\mathcal{F})\right]^2}. \tag{D.1}$$

D.3 Assumptions and Lemmas Based on Austern and Zhou (2020)

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

1. *$l_\Theta(\theta, z)$ and $l_\mathcal{B}(\beta, z)$ are twice differentiable and strictly convex in θ and β.*
2. *$\mathbb{E}\left[\sup_{\theta\in\Theta} l_\Theta^4(\theta, Z_i)\right] < \infty$ and $\mathbb{E}\left[\sup_{\beta\in\mathcal{B}} l_\mathcal{B}^4(\beta, Z_i)\right] < \infty$.*
3. *There exist open neighborhoods $\mathcal{O}_{\theta^*}$ and $\mathcal{O}_{\beta^*}$ of $\theta^*$and β^* in Θ and $\mathcal{B}$ such that*
   (a) *$\mathbb{E}\left[\sup_{\theta\in\mathcal{O}_{\theta^*}} \|\nabla_\theta l_\Theta(\theta, Z_i)\|^{16}\right] < \infty$, $\mathbb{E}\left[\sup_{\beta\in\mathcal{O}_{\beta^*}} \|\nabla_\beta l_\mathcal{B}(\beta, Z_i)\|^{16}\right] < \infty$.*
   (b) *$\mathbb{E}\left[\sup_{\theta\in\mathcal{O}_{\theta^*}} \|\nabla_\theta^2 l_\Theta(\theta, Z_i)\|^{16}\right] < \infty$, $\mathbb{E}\left[\sup_{\beta\in\mathcal{O}_{\beta^*}} \|\nabla_\beta l_\mathcal{B}(\beta, Z_i)\|^{16}\right] < \infty$.*
   (c) *there exists $c > 0$ such that $\lambda_{min}\left(\nabla_\theta^2 l_\Theta(\theta, Z_i)\right) \geq c$, $\lambda_{min}\left(\nabla_\beta^2 l_\mathcal{B}(\beta, Z_i)\right) \geq c$ a.s. uniformly on $\mathcal{O}_{\theta^*}$ and $\mathcal{O}_{\beta^*}$.*

Lemma D.1 (Application of Proposition 5 of Austern and Zhou, 2020)**.** *Under Assumption 3:*

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

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 θ, β and $\alpha = (\theta, \beta)$, we obtain:

$$\sqrt{N}\left(CV(\mathcal{F}_\Theta) - e_{\mathcal{F}_\Theta, \frac{K-1}{K}N}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(l\left(f_{\theta^*}, Z_i\right)\right)\right),$$
$$\sqrt{N}\left(CV(\mathcal{F}_\mathcal{M}) - e_{\mathcal{F}_\mathcal{M}, \frac{K-1}{K}N}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(l\left(f^*, Z_i\right)\right)\right),$$
$$\sqrt{N}\left(CV(\mathcal{F}_\Theta) + CV(\mathcal{F}_\mathcal{M}) - e_{\mathcal{F}_\Theta, \frac{K-1}{K}N} - e_{\mathcal{F}_\mathcal{M}, \frac{K-1}{K}N}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(l\left(f_{\theta^*}, Z_i\right) + l\left(f^*, Z_i\right)\right)\right).$$

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

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

□

Lemma D.2 (Application of Proposition 1 of Austern and Zhou, 2020)**.** *Under Assumption 3,*

$$\hat{\sigma}_\Delta^2 \xrightarrow{p} \mathrm{Var}\left(\Delta l\left(f_{\theta^*}, f^*; Z_i\right)\right).$$

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

$$\begin{aligned}\hat{\sigma}^2_{CV(\mathcal{F}_\Theta)} := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N - 1}\sum_{k(i)=k}&\left(l\left(f_{\hat{\theta}^{-k}}, Z_i\right) - \frac{1}{J_N}\sum_{k(j)=k} l\left(f_{\hat{\theta}^{-k}}, Z_j\right)\right)^2 \\ &\xrightarrow{p} \mathrm{Var}\left(l\left(f_{\theta^*}, Z_i\right)\right).\end{aligned}$$

and

$$\begin{aligned}\hat{\sigma}^2_{CV(\mathcal{F}_\mathcal{M})} := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N - 1}\sum_{k(i)=k}&\left(l\left(f_{\left[\hat{\beta}^{-k}\right]}, Z_i\right) - \frac{1}{J_N}\sum_{k(j)=k} l\left(f_{\left[\hat{\beta}^{-k}\right]}, Z_j\right)\right)^2 \\ &\xrightarrow{p} \mathrm{Var}\left(l\left(f^*, Z_i\right)\right).\end{aligned}$$

and

$$\begin{aligned}&\hat{\sigma}^2_{CV(\mathcal{F}_\Theta)+CV(\mathcal{F}_\mathcal{M})} \\ :=&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{J_N - 1}\cdot\sum_{k(i)=k} \\ &\left(l\left(f_{\hat{\theta}^{-k}}, Z_i\right) + l\left(f_{\left[\hat{\beta}^{-k}\right]}, Z_i\right) - \frac{1}{J_N}\sum_{k(j)=k}\left[l\left(f_{\left[\hat{\beta}^{-k}\right]}, Z_j\right) + l\left(f_{\hat{\theta}^{-k}}, Z_i\right)\right]\right)^2 \\ \xrightarrow{p}&\mathrm{Var}\left(l\left(f_{\theta^*}, Z_i\right) + l\left(f^*, Z_i\right)\right),\end{aligned}$$

Hence:

$$\begin{aligned}\hat{\sigma}^2_{\Delta_\theta} &= 2\hat{\sigma}^2_{CV(\mathcal{F}_\Theta)} + 2\hat{\sigma}^2_{CV(\mathcal{F}_\mathcal{M})} - \hat{\sigma}^2_{CV(\mathcal{F}_\Theta)+CV(\mathcal{F}_\mathcal{M})} \\ &\xrightarrow{p} 2\mathrm{Var}\left(l\left(f_{\theta^*}, Z_i\right)\right) + 2\mathrm{Var}\left(l\left(f^*, Z_i\right)\right) - 2\mathrm{Var}\left(l\left(f_{\theta^*,Z_i}\right) + l\left(f^*, Z_i\right)\right) \\ &= \mathrm{Var}\left(\Delta l\left(f_{\theta^*}, f^*; Z_i\right)\right)\end{aligned}$$

□

D.4 Proof of Asymptotic Normality of $\hat{\kappa}^*$

Lemma D.1 characterizes the limit distribution of

$$\sqrt{N}\left[CV\left(\mathcal{F}_\Theta\right) - CV\left(\mathcal{F}_\mathcal{M}\right) - \left(e_{\mathcal{F}_\Theta, \frac{K-1}{K}N} - e_{\mathcal{F}_\mathcal{M}, \frac{K-1}{K}N}\right)\right]$$

which we now show is also the limit distribution of

$$\sqrt{N}\left[CV\left(\mathcal{F}_{\Theta}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{\mathcal{F}_{\Theta}}-e_{\mathcal{F}_{\mathcal{M}}}\right)\right].$$

To see this, notice that

$$\begin{aligned}
e_{\Theta,\frac{K-1}{K}N}-e_{\mathcal{F}_{\Theta}} &= \mathbb{E}\left[l_{\Theta}\left(\hat{\theta}^{-k(i)},Z_i\right)-l_{\Theta}\left(\theta^*,Z_i\right)\right] \\
&= \mathbb{E}\left[\nabla l_{\Theta}\left(\theta^*,Z_i\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right)+\left(\hat{\theta}^{-k(i)}-\theta^*\right)^{\prime}\nabla^2 l_{\Theta}\left(\tilde{\theta},Z_i\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= 0+\mathbb{E}\left[\left(\hat{\theta}^{-k(i)}-\theta^*\right)^{\prime}\nabla^2 l_{\Theta}\left(\tilde{\theta},Z_i\right)\cdot\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= \frac{1}{N-J_N}\mathbb{E}\left[\sqrt{N-J_N}\left(\hat{\theta}^{-k(i)}-\theta^*\right)^{\prime}\nabla^2 l_{\Theta}\left(\tilde{\theta},Z_i\right)\cdot\sqrt{N-J_N}\left(\hat{\theta}^{-k(i)}-\theta^*\right)\right] \\
&= c\frac{1}{N-J_N}+o\left(\frac{1}{N-J_N}\right)=c\frac{K}{K-1}\cdot\frac{1}{N}+o\left(\frac{1}{N}\right)
\end{aligned}$$

since $J_N=N/K$, and hence

$$\sqrt{N}\left(e_{\Theta,\frac{K-1}{K}N}-e_{\Theta}\right)=o_p\left(1\right).$$

Similarly, $\sqrt{N}\left(e_{\mathcal{F}_{\mathcal{M}},\frac{K-1}{K}N}-e_{\mathcal{F}_{\mathcal{M}}}\right)=o_p\left(1\right)$. Hence:

$$\sqrt{N}\left[CV\left(\mathcal{F}_{\Theta}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{\mathcal{F}_{\Theta}}-e_{\mathcal{F}_{\mathcal{M}}}\right)\right]\xrightarrow{d}\mathcal{N}\left(0,\operatorname{Var}\left(\Delta l\left(f_{\theta^*},f^*;Z_i\right)\right)\right).$$

Now, we replicate the previous result with f_{naive} in place of $\mathcal{F}_{\Theta}$ and obtain

$$\sqrt{N}\left[CV\left(f_{\text{naive}}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{f_{\text{naive}}}-e_{\mathcal{F}_{\mathcal{M}}}\right)\right]\xrightarrow{d}\mathcal{N}\left(0,\operatorname{Var}\left(\Delta l\left(f_{\text{naive}},f^*;Z_i\right)\right)\right).$$

and jointly

$$\sqrt{N}\left(\begin{array}{c} CV\left(\mathcal{F}_{\Theta}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{\mathcal{F}_{\Theta}}-e_{\mathcal{F}_{\mathcal{M}}}\right) \\ CV\left(f_{\text{naive}}\right)-CV\left(\mathcal{F}_{\mathcal{M}}\right)-\left(e_{f_{\text{naive}}}-e_{\mathcal{F}_{\mathcal{M}}}\right)\end{array}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\left(\begin{array}{cc}\sigma^2_{\Delta_\theta} & \sigma_{\Delta_\theta\Delta_{\text{naive}}} \\ \sigma_{\Delta_\theta\Delta_{\text{naive}}} & \sigma^2_{\Delta_{\text{naive}}}\end{array}\right)\right)$$

with

$$\begin{aligned}
\sigma^2_{\Delta_\theta} &:= \operatorname{Var}\left(\Delta l\left(f_{\theta^*},f^*;Z_i\right)\right) \\
\sigma^2_{\Delta_{\text{naive}}} &:= \operatorname{Var}\left(\Delta l\left(f_{naive},f^*;Z_i\right)\right) \\
\sigma_{\Delta_\theta\Delta_{\text{naive}}} &:= \operatorname{Cov}\left(\Delta l\left(f_{\theta^*},f^*;Z_i\right),\Delta l\left(f_{naive},f^*;Z_i\right)\right)
\end{aligned}$$

By Lemma D.2, Assumption 2 and the Delta Method, we have

$$\sqrt{N}\left(\hat{\kappa}^* - \kappa^*\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma^2_{\Delta_\theta} - 2\kappa^*\sigma_{\Delta_\theta\Delta_{\text{naive}}} + \kappa^{*2}\sigma^2_{\Delta_{\text{naive}}}}{d^2\left(f_{\text{naive}}, f^*\right)}\right)$$

Since

$$\hat{\sigma}_{\hat{\kappa}^*} \xrightarrow{p} \frac{\sigma^2_{\Delta_\theta} - 2\kappa^*\sigma_{\Delta_\theta\Delta_{\text{naive}}} + \kappa^{*2}\sigma^2_{\Delta_{\text{naive}}}}{d^2\left(f_{\text{naive}}, f^*\right)},$$

we have

$$\frac{\sqrt{N}\left(\hat{\kappa}^* - \kappa^*\right)}{\hat{\sigma}_{\hat{\kappa}^*}} \xrightarrow{d} \mathcal{N}\left(0, 1\right).$$

E Supplementary Material to Section 8

E.1 Alternative Estimator of f^*-Discrepancy

We now discuss an alternative estimator for f^*-discrepancy

$$\delta_{f^*} = \frac{d(f_{\theta^*}, f^*)}{d(f_{\text{naive}}, f^*)}$$

when the decomposability condition (10) does not hold.

We again work with the parameterization of $\mathcal{F}_{\mathcal{M}}$ via $\beta \in \mathcal{B}$. Suppose that we have access to an estimator $\hat{\beta}$ of β^* that is consistent and asymptotically normal:

$$\sqrt{N}\left(\hat{\beta} - \beta^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Sigma\right).$$

Given that $\theta^* = \arg\min_{\theta\in\Theta} d\left(f_\theta, f_{[\beta^*]}\right)$, we can construct an estimator of θ^* as

$$\hat{\theta} := \hat{\theta}\left(\hat{\beta}\right) := \arg\min_{\theta\in\Theta} d\left(f_\theta, f_{[\hat{\beta}]}\right),$$

with which we can obtain the following estimator of δ_{f^*}

$$\hat{\delta}_{f^*} := \frac{d\left(f_{\hat{\theta}(\hat{\beta})}, f_{[\hat{\beta}]}\right)}{d\left(f_{\text{naive}}, f_{[\hat{\beta}]}\right)} = \frac{\min_{\theta\in\Theta} d\left(f_\theta, f_{[\hat{\beta}]}\right)}{d\left(f_{\text{naive}}, f_{[\hat{\beta}]}\right)}.$$

We impose the following joint assumption on the dissimilarity function d and the parameterization of $\mathcal{F}_\Theta$ and $\mathcal{F}_{\mathcal{M}}$.

Assumption 4. *Define* $\overline{d}\left(\theta, \beta\right) := d\left(f_\theta, f_{[\beta]}\right)$. *Suppose that:*

(a) $\overline{d}$ *is joint differentiable with respect to* (θ, β) *in a neighborhood of* (θ^*, β^*).

(b) $\psi^* := \nabla_\beta \overline{d}\left(\hat{\theta}\left(\beta\right), \beta\right)\Big|_{\beta=\beta^*} \neq \mathbf{0}$.

The requirements in Assumption 4 are very weak. Part (a) is a standard differentiability condition, which should be satisfied in most applications. For (b), notice that by the Envelope Theorem,

$$\psi^* := \nabla_\beta \overline{d}\left(\hat{\theta}\left(\beta^*\right), \beta^*\right) = \frac{\partial}{\partial \beta}\overline{d}\left(\theta^*, \beta^*\right)$$

Hence, $\psi^* \neq \mathbf{0}$ essentially requires that the dissimilarity $d(f_{\theta^*}, f)$ between f_{θ^*} and f as f varies locally in a neighborhood of f^*.

By the Delta Method,

$$\begin{aligned} &\sqrt{N}\left(\min_{\theta\in\Theta} d\left(f_\theta, f_{[\hat{\beta}]}\right) - \min_{\theta\in\Theta} d\left(f_\theta, f_{[\beta^*]}\right)\right) \\ &= \sqrt{N}\left(\overline{d}\left(\hat{\theta}\left(\hat{\beta}\right), \hat{\beta}\right) - \overline{d}\left(\hat{\theta}\left(\beta^*\right), \beta^*\right)\right) \quad \xrightarrow{d} \quad \mathcal{N}\left(0, \psi^{*\prime}\Sigma\psi^*\right). \end{aligned}$$

Again, jointly writing

$$\sqrt{N}\left(\begin{array}{c} \overline{d}\left(\hat{\theta}\left(\hat{\beta}\right), \hat{\beta}\right) - \overline{d}\left(\hat{\theta}\left(\beta^*\right), \beta^*\right) \\ d\left(f_{\text{naive}}, f_{[\hat{\beta}]}\right) - d\left(f_{\text{naive}}, f_{\beta^*}\right) \end{array}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\begin{array}{cc} \sigma^2_{\Delta_\theta} & \sigma_{\Delta_\theta \Delta_{\text{naive}}} \\ \sigma_{\Delta_\theta \Delta_{\text{naive}}} & \sigma^2_{\Delta_{\text{naive}}} \end{array}\right)\right)$$

we have

$$\sqrt{N}\left(\hat{\delta}_{f^*} - \delta_{f^*}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma^2_{\Delta_\theta} - 2\kappa^*\sigma_{\Delta_\theta\Delta_{\text{naive}}} + \kappa^{*2}\sigma^2_{\Delta_{\text{naive}}}}{d^2\left(f_{\text{naive}}, f^*\right)}\right).$$

E.2 Example: Lack of Decomposability

Consider a setting where X is degenerate, i.e., $\mathcal{X}$ is a singleton, so that the joint distribution P is completely characterized by the distribution of Y. Furthermore, let $\mathcal{Y} := [0, 1]$.

If $f^* := \text{med}\left(Y\right) \in \mathcal{S} := \mathcal{Y} = [0, 1]$, then a mapping $f : \mathcal{X} \to \mathcal{S}$ is just a number in $[0, 1]$. When the loss function is the absolute deviation $l\left(f, y\right) := \left|y - f\right|$, and the error function is mean absolute deviation $e_{P^*}\left(f\right) := \mathbb{E}_{P^*}\left[\left|Y - f\right|\right]$, the true median f^* minimizes the error, i.e. $f^* \in \arg\min_{f\in[0,1]} e_{P^*}\left(f\right)$. However, it is not true that $\left|f - f^*\right| = e_{P^*}\left(f\right) - e_{P^*}\left(f^*\right)$ for any $f \in [0, 1]$. To see this, suppose that $Y \sim U\left[0, 1\right]$ under P^*. Then $f^* = 0.5$ and $e_{P^*}\left(f^*\right) = 0.25$. However, for $f = 0.4$, we have

$e_{P^*}(f) = 0.26.$ but $|f - f^*| = 0.1 \neq 0.01 = e_{P^*}(f) - e_{P^*}(f^*)$.

Moreover, there is no function $d : [0,1]^2 \to [0,1]$ such that decomposability (10) holds, which would require that $d(f, f_P) = e_P(f) - e_P(f_P)$ for any distribution P of Y supported on $[0,1]$. To see this, suppose that $Y \sim U[0,1]$ under P_1, we have

$$e_{P_1}(f) - e_{P_1}(f_{P_1}) = (f - 0.5)^2 = (f - f_{P_1})^2, \quad \forall f \in [0,1].$$

However, supposing that, under P_2, the probability density function of Y is given by $2y$ for $y \in [0,1]$, we have $f_{P_2} = \sqrt{2}/2$ and $e_{P_2}(f_{P_2}) = (2 - \sqrt{2})/3$ but

$$e_{P_2}(f) - e_{P_2}(f_{P_2}) = \frac{1}{3}\left(2f^3 - 3f^2 + \sqrt{2}\right) \neq (f - f_{P_2})^2.$$

F Extension to Nonparametric $\mathcal{F}_{\mathcal{M}}$

Now we consider a setting where $\mathcal{X}$, the support of X, is a continuum, and the model class $\mathcal{F}_{\mathcal{M}}$ is nonparametric. For simplicity, we focus on the case of a compact and rectangular $\mathcal{X} := [0,1]^{d_x}$.

F.1 Computing Restrictiveness r

Here the distribution μ on $\mathcal{F}_{\mathcal{M}}$ is a distribution over an infinite-dimensional functional space, so "direct" simulation from μ becomes infeasible. We propose two ways to proceed:

Simulation on a Growing Grid

For a given number of simulations M, we can restrict our attention to a finite grid of the form

$$\mathcal{X}_M := \left\{ \frac{k}{K_M} : \ k = 0, 1, ..., 2^{K_M} \right\}^{d_x},$$

where $K_M \to \infty$ and $K_M^{d_x}/M \to 0$ as $M \to \infty$. We then proceed by simulating f from a measure (say, uniform) μ_M on the restriction of $\mathcal{F}_{\mathcal{M}}$ on $\mathcal{X}_M$.

Simulation of Coefficients on Basis Functions

Alternatively, if $\mathcal{F}_{\mathcal{M}}$ satisfies certain smoothness conditions (e.g. possesss uniformly bounded derivatives up to a certain order), then we can specify a sequence of orthonormal basis functions $\{b_k(x) : k \in \mathbb{N}\}$, such as (tensor products of) power series, trigonometric series, splines, wavelets (See, for example, Chen 2007 for), such that the linear span of the basis functions is dense in $\mathcal{F}_{\mathcal{M}}$. Shape restrictions in $\mathcal{F}_{\mathcal{M}}$, such as nonnegativity, monotonicity and convexity, can also be incorporated by proper specification of the basis functions. Writing

$$\mathcal{B}_M := \left\{ f_{[\beta]} := \sum_{k=1}^{K_M} \beta_k b_k(x) : \ \beta \in \mathbb{R}^{K_M} \right\}$$

for some $K_M \to \infty$ as $M \to \infty$, we could specify simulate f from $\mathcal{B}_M$ by randomly drawing β from some measure $\mu_{\beta,M}$ on $\mathbb{R}^{K_M}$.

F.2 Estimating Completeness κ^*

The estimation of completeness κ^* can be also adapted to accommodate an infinite-dimensional $\mathcal{F}_M$, either via the growing-grid approach or the basis-function approach. We illustrate the asymptotic property of $\hat{\kappa}^*$ using the later approach, and focus on a simpler setting without no cross validations.

Specifically, for a given loss function l, define

$$\begin{aligned}
\hat{e}(f) &:= \frac{1}{N}\sum_{i=1}^{N} l(Z_i, f) \\
\hat{\theta} &:= \arg\min_{\theta\in\Theta} \hat{e}(f_\theta) \\
\hat{\beta} &:= \arg\min_{\beta\in\mathcal{B}_{\mathcal{M}}} \hat{e}\left(f_{[\beta]}\right) \\
\hat{e}(\mathcal{F}_\Theta) &:= \hat{e}(f_{\hat{\theta}}) \\
\hat{e}(\mathcal{F}_{\mathcal{M}}) &:= \hat{e}\left(f_{[\hat{\beta}]}\right).
\end{aligned}$$

and

$$\hat{\kappa}^* := 1 - \frac{\hat{e}(\mathcal{F}_\Theta) - \hat{e}(\mathcal{F}_{\mathcal{M}})}{\hat{e}(f_{\text{naive}}) - \hat{e}(\mathcal{F}_{\mathcal{M}})}.$$

Under standard regularity conditions for nonparametric estimation,

$$\left\| f_{\hat{\beta}} - f^* \right\| \xrightarrow{p} 0, \text{ as} N \to \infty.$$

Observing

$$\frac{1}{N}\sum_{i=1}^{N}\left[l\left(Z_i, f_{\hat{\theta}}\right)-\hat{e}\left(\mathcal{F}_\Theta\right)\right]=0, \quad \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta l\left(Z_i, f_{\hat{\theta}}\right)=0,$$

and, by the definition of θ^*,

$$\mathbb{E}\left[\nabla_\theta l\left(Z_i, f_{\theta^*}\right)\right]=0$$

we have, by Theorem 6.1 of Newey and McFadden (1994),

$$\sqrt{N}\left[\hat{e}\left(\mathcal{F}_\Theta\right)-e\left(\mathcal{F}_\Theta\right)\right]\xrightarrow{d}\mathcal{N}\left(0, \mathrm{Var}\left[l\left(Z_i, f_{\theta^*}\right)\right]\right).$$

Similarly, since

$$\frac{1}{N}\sum_{i=1}^{N}\left[l\left(Z_i, f_{[\hat{\beta}]}\right)-\hat{e}\left(\mathcal{F}_\mathcal{M}\right)\right]=0, \quad \frac{1}{N}\sum_{i=1}^{N}\nabla_\beta l\left(Z_i, f_{[\hat{\beta}]}\right)=0,$$

and

$$\mathbb{E}\left[\nabla_\beta l\left(Z_i, f_{[\beta^*]}\right)\right]=0,$$

we have, by Proposition 2 of Newey (1994),

$$\sqrt{N}\left[\hat{e}\left(\mathcal{F}_\mathcal{M}\right)-e\left(\mathcal{F}_\mathcal{M}\right)\right]\xrightarrow{d}\mathcal{N}\left(0, \mathrm{Var}\left[l\left(Z_i, f^*\right)\right]\right).$$

It is then straightforward to extend the above to obtain:

$$\sqrt{N}\begin{pmatrix}\hat{e}\left(\mathcal{F}_\Theta\right)-\hat{e}\left(\mathcal{F}_\mathcal{M}\right)-e\left(\mathcal{F}_\Theta\right)+e\left(\mathcal{F}_\mathcal{M}\right)\\ \hat{e}\left(f_{\text{naive}}\right)-\hat{e}\left(\mathcal{F}_\mathcal{M}\right)-e\left(f_{\text{naive}}\right)+e\left(\mathcal{F}_\mathcal{M}\right)\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(0, \begin{pmatrix}\sigma^2_{\Delta_\theta} & \sigma_{\Delta_\theta\Delta_{\text{naive}}}\\ \sigma_{\Delta_\theta\Delta_{\text{naive}}} & \sigma^2_{\Delta_{\text{naive}}}\end{pmatrix}\right)$$

and

$$\sqrt{N}\left(\hat{\kappa}^*-\kappa^*\right)\xrightarrow{d}\mathcal{N}\left(0,\ \frac{\sigma^2_{\Delta_\theta}-2\kappa^*\sigma_{\Delta_\theta\Delta_{\text{naive}}}+\kappa^{*2}\sigma^2_{\Delta_{\text{naive}}}}{\left(e(f_{\text{naive}})-e(\mathcal{F}_\mathcal{M})\right)^2}\right).$$

with $\sigma^2_{\Delta_\theta}$, $\sigma_{\Delta_\theta\Delta_{\text{naive}}}$and $\sigma^2_{\Delta_{\text{naive}}}$ defined in the same way as in earlier sections.