Model Restrictiveness in Functional and Structural Settings

Drew Fudenberg*, Wayne Yuan Gao†, and Zhiheng You‡

February 7, 2026

Abstract

We generalize the notion of model restrictiveness in Fudenberg, Gao and Liang (2023) to a wider range of economic models with semi/non-parametric and structural ingredients. We show how restrictiveness can be defined and computed in infinite-dimensional settings using Gaussian process priors (including with shape restrictions) and other alternativess in Bayesian nonparametrics. We also extend the restrictiveness framework to structural models with endogeneity, instrumental variables, multiple equilibria, and nonparametric nuisance components. We discuss the importance of the user-specific choice of discrepancy functions in the context of Rademacher complexity and GMM criterion function, and relate restrictiveness to the limit of the average-case learning curve in machine learning. We consider applications to: (1) preferences under risk, (2) exogenous multinomial choice, and (3) multinomial choice with endogenous prices: for (1), we obtain results consistent with those in Fudenberg, Gao and Liang (2023); for (2) and (3), our findings show that nested logit and mixed logit exhibit similar restrictiveness under standard parametric specifications, and that IV exogeneity conditions substantially increase overall restrictiveness while altering model rankings.

Keywords: restrictiveness, complexity, flexibility, structural model, semiparametric, nonparametric, endogeneity, Bayesian nonparametric, Gausian process

*Department of Economics, Massachusetts Institute of Technology, drewf@mit.edu
†Department of Economics, University of Pennsylvania, waynegao@upenn.edu.
‡Department of Economics, University of Pennsylvania, zhyou@sas.upenn.edu.

1 Introduction

As Box (1976) famously remarked, "All models are wrong, but some are useful." Almost all economic models are restrictive. They rule out some patterns in the data while emphasizing some others based on relevant economic theory, existing empirical knowledge, and intended usage of the model. The restrictiveness of economic models is a feature rather than a bug. Restrictions can encode economic structure, facilitate interpretation, and improve economic decision making. Yet restrictive models can be misspecified, and the nature of misspecification varies widely. Researchers often lack a quantitative sense of how much structure a given model imposes relative to plausible alternatives: for example, choosing between multinomial logit and mixed logit, or between different specifications of preferences under risk).

Fudenberg, Gao and Liang (2023) develops a framework for measuring the *restrictiveness* of a (theoretical) model, viewed as a *prediction rule* $f : \mathcal{X} \to \mathcal{Y}$ mapping (exogenous) covariates X to outcomes Y. Restrictiveness is defined relative to: (i) a user-specified set of eligible prediction rules $\mathcal{F}$, (ii) a discrepancy function d, and (iii) an evaluation distribution $\lambda_{\mathcal{F}}$ on $\mathcal{F}$. For a model class $\mathcal{F}_\Theta \subseteq \mathcal{F}$, restrictiveness measures the expected discrepancy between it and a pseudo-true rule drawn from $\lambda_{\mathcal{F}}$, normalized by the baseline discrepancy. Restrictiveness measures how much average approximation error is incurred by restricting to models in $\mathcal{F}_\Theta$ instead of using the flexible benchmark. Restrictiveness captures structural content but does not assess fit to observed data or statistical rejection-questions addressed by standard estimation, testing, and the completeness measure of Fudenberg, Kleinberg, Liang and Mullainathan (2022). Instead, restrictiveness quantifies how much structure a model imposes *a priori*.

This paper extends the restrictiveness measure to a much broader range of economic models that feature semi/non-parametric and structural ingredients. First, Fudenberg, Gao and Liang (2023) focuses mainly on a setting where $\mathcal{F}$ is a finite-dimensional compact space, for which the uniform distribution is well-defined and serves as a natural choice for $\lambda_{\mathcal{F}}$. This paper allows $\mathcal{F}$ to be an infinite-dimensional functional space, and operationalizes restrictiveness based on Bayesian nonparametric priors such as Gaussian Processes and Dirichlet Processes. We also show how to sample from $\lambda_{\mathcal{F}}$ in settings where $\mathcal{F}$ imposes shape restrictions such as monotonicity.

Second, while Fudenberg, Gao and Liang (2023) only discusses restrictiveness of

"reduced-form" models (i.e. models with explicit restrictions on prediction rules that map exogenous covariates to the outcomes), we extend the notion of restrictiveness to structural economic models, which are important in applied microeconomic areas such as industrial organization and labor economics. In structural models, researchers typically specify a structural form that implies a reduced-form distribution for Y conditional on X. We show how to define restrictiveness for such models by working with the reduced forms and their induced conditional distributions. We treat several important cases: fully parametric models with endogeneity, models defined by moment equalities, models with multiple equilibria, and semiparametric models. In some settings, such as additive-error models with endogenous regressors, we show that the infinite-dimensional optimization over functions can be reduced to a finite-dimensional optimization over the structural parameters, greatly simplifying computation.

We then relate our notion of restrictiveness to existing measures of model complexity and discrepancy in statistics, econometrics, and machine learning. In particular, we examine the connection between restrictiveness and Rademacher complexity in a binary classification setting, and clarify that the degeneracy of restrictiveness established in Ellis and Neff (2025) is tied to a specific correlation-based discrepancy. We argue that the choice of discrepancy should be guided by interpretability and context rather than by existing capacity measures designed for other purposes. We also explain why GMM criterion functions are not suitable as discrepancy functions for restrictiveness: GMM criterion functions are defined as quadratic forms that capture violations of moment conditions, rather than measures of distance between the model-implied data distribution and the (pseudo-)true distribution. We further show that, in a pure approximation-error sense and in the absence of noise, restrictiveness can be interpreted as the normalized limit of the average-case learning curve, a well-studied concept in machine learning. A central message of our analysis is that d is not fixed: it should be chosen to reflect the kind of approximation error that matters in the application, rather than inherited mechanically from existing complexity measures.

We apply our approach to three economic problems. First, we revisit the analysis of the restrictiveness of Cumulative Prospect Theory (CPT) and Disappointment Aversion (DA) models in Fudenberg, Gao and Liang (2023), replacing that paper's finite set of binary lotteries with the entire space of monotone, bounded prediction rules for arbitrary binary lotteries. Second, we study the restrictiveness of standard discrete-choice models in industrial organization-multinomial logit, nested logit, and

mixed logit, and quantify how much of mixed logit's theoretical flexibility is actually realized by the parametric forms used in practice, and how the standard IO toolkit trades off restrictiveness and completeness. Third, we extend the analysis to multinomial choice with endogenous product characteristics. Using BLP-style instruments, we compare the discrete-choice models in an IV setting, where each specification is constrained both by its functional form and by moment conditions.

Related literature

Fudenberg, Kleinberg, Liang and Mullainathan (2022)'s *completeness* measures the fraction of the predictable variation in an outcome that is captured by a given model, relative to a flexible statistical benchmark. It is implemented using machine-learning methods to approximate the best possible prediction given observables, and has been applied to models of choice under risk and other domains. Fudenberg, Gao and Liang (2023) proposes a measure of *restrictiveness* of a model, which is also the object of interest in our current paper. Fudenberg, Gao and Liang (2023) also proposes evaluating models by comparing their restrictiveness together with their completeness, which produces an empirical Pareto frontier that trades off fit on real data against the regularities ruled out by the model. Our contribution here is to: (1) develop a fully nonparametric, population-level notion of restrictiveness that applies to settings with continuum domains, (2) adapt and generalize the notion of restrictiveness to structural econometric models with endogeneity and multiple equilibria, and (3) articulate the choice of discrepancy function as a substantive modeling decision, especially in relationship with some existing concepts in econometrics and machine learning.

Ellis and Neff (2025) studies the connection between restrictiveness and Rademacher complexity in a binary classification setting. It shows that, for a particular choice of eligible set and a discrepancy inspired by Rademacher complexity, a normalized version of our restrictiveness index is an affine transformation of the limiting Rademacher complexity of the model class. In that special case, all finite-dimensional falsifiable models appear "fully restrictive" in the limit. Section 4.1, interprets the degeneracy as a critique of this particular discrepancy rather than of restrictiveness itself: when d is chosen to encode an existing capacity measure, restrictiveness inherits that measure's asymptotic behavior and limitations. In addition, Ellis and Neff (2025) also propose a finite-sample version of discrepancy function d_n (as a sample average

using the n data points on the features), which can be convenient in a variety of empirical settings. Ellis and Neff (2025) does not develop the asymptotic theory of d_n, which is required to construct confidence intervals for restrictiveness that reflects the finite-sample randomness of the covariates X_i; we provide that here.

More broadly, our framework is related to the statistical learning theory literature on model complexity and capacity, where notions such as VC dimension, Rademacher complexity, and metric entropy have been used to establish upper bounds on generalization errors under empirical risk minimization. Although conceptually related, our restrictiveness measure is an *average-case* approximation measure, defined with respect to a context-specific user-chosen discrepancy function (under an evaluation distribution on an economically meaningful eligible set), and our restrictiveness measure produces an interpretable number in the unit interval rather than rate bounds. See, also, Section 3.4 of Fudenberg, Gao and Liang (2023) for related discussions and references. In this paper, we further stress that the choice of the discrepancy function should be treated as an important modeling decision for restrictiveness to be interpretable in a context-specific manner.

The rest of the paper is organized as follows. Section 2 describes how to define and compute restrictiveness in functional settings (with continuous feature space). Section 3 discusses how to define and compute the restrictiveness of structural econometric models with endogeneity, multiple equilibria, or semiparametric specifications. Section 4 relates and compares our restrictiveness measure to a variety of related but different existing concepts in statistics and econometrics. We consider three concrete applications in Section 5, and conclude in Section 6.

2 Restrictiveness in Functional Settings

2.1 Setup

Our starting point is a random sample (X, Y), where X is a *covariate vector* and $Y \in \mathcal{Y}$ is an *outcome* variable. We use $\mathcal{X}$ to denote the support of the covariates, and P_X to denote the marginal distribution of X. Our basic setup follows that of Fudenberg, Gao and Liang (2023, "FGL" thereafter) with the exception that in this paper we do *not* restrict $\mathcal{X}$ to be finite. Instead, we assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, and P_X is either chosen by the researcher, known a priori, or estimated

from data. A *prediction rule* is a function $f : \mathcal{X} \to \mathcal{Y}$. We denote the set of all such functions by $\overline{\mathcal{F}} \equiv \mathcal{Y}^{|\mathcal{X}|}$, which is assumed to be a well-defined metric space.

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

We will evaluate the restrictiveness of a specific model class $\mathcal{F}_\Theta := \{f_\theta\}_{\theta \in \Theta} \subseteq \overline{\mathcal{F}}$, where the prediction rules f_θ depend continuously on a parameter θ from a parameter set Θ, which can be finite or infinite dimensional. Restrictiveness is defined relative to a compact set of "eligible" rules $\mathcal{F} \subseteq \overline{\mathcal{F}}$ that reflect any constraints the model is known to have. For example, if a model is known to imply that choices respect first-order stochastic dominance, we can define $\mathcal{F}$ to be all rules with this property, and measure the model's additional restrictiveness beyond this. In general, the eligible set $\mathcal{F}$ consists of all prediction rules that satisfy user-specified background constraints, where the special case of $\mathcal{F} = \overline{\mathcal{F}}$ corresponds to the question of whether $\mathcal{F}_\Theta$ imposes any restrictions at all.

Let $\lambda_\mathcal{F}$ denote a chosen evaluation distribution on $\mathcal{F}$. We define the restrictiveness of a model to be its expected discrepancy to a prediction rule f randomly drawn from $\lambda_\mathcal{F}$, normalized with respect to the expected discrepancy of a baseline prediction rule $\mathcal{F}_{\text{base}}$. This baseline prediction rule is chosen to suit the setting, and we interpret its performance as a lower bound that any sensible model should outperform: for example, in some scenarios a natural baseline is the constant model $\mathcal{F}_{\text{base}} = \{c : c \in \mathbb{R}\}$, while in others it may be a singleton set $\mathcal{F}_{\text{base}} = \{f_{\theta_0}\}$, where f_{θ_0} is the model in $\mathcal{F}_\Theta$ evaluated at baseline parameter θ_0.

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

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

$$r(\mathcal{F}_\Theta; \mathcal{F}, d) = \frac{\mathbb{E}_{\lambda_\mathcal{F}}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}_{\lambda_\mathcal{F}}[d(\mathcal{F}_{\text{base}}, f)]} \tag{1}$$

and

$$d(\mathcal{F}_\Theta, f) := \inf_{f_\theta \in \mathcal{F}_\Theta} d(f_\theta, f).$$

Note that equation (1) implies that restrictiveness is invariant to affine transformation of the discrepancy function d by the linearity of the expectation operator. In fact, restrictiveness is unitless, and lies within the unit interval $[0, 1]$ when $\mathcal{F}_\Theta$ nests $\mathcal{F}_{\text{base}}$ as a special case.

2.2 Evaluation and Numerical Implementation

Computing restrictiveness requires choosing an evaluation distribution $\lambda_\mathcal{F}$ over an infinite-dimensional functional space and sampling from it. This marks the key difference between this paper and FGL, which focuses on sampling from a distribution over a finite-dimensional space. Sampling from a distribution over an infinite-dimensional functional space has been studied and implemented with Bayesian nonparametric methods, which often require sampling from an infinite-dimensional "prior" distribution. Specifically, the Gaussian process, Dirichlet process, and their mixtures are commonly used to define such priors, and they can be configured in flexible ways for various problem setups.

Gaussian process (GP). GP is a standard tool in Bayesian nonparametric estimation for placing priors over functional spaces. The formal definition of a GP is: a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ such that for any finite collection of input points $x_1, \ldots, x_n \in \mathcal{X}$, the joint distribution $(f(x_1), \ldots, f(x_n))^\top$ follows a multivariate normal distribution:

$$(f(x_1), \ldots, f(x_n))^\top \sim \mathcal{N}(\mu, K),$$

where $\mu_i = \mathbb{E}[f(x_i)]$ and $K_{ij} = \text{Cov}(f(x_i), f(x_j))$. We denote this as:

$$f \sim \mathcal{GP}(m(x), K(x, x')).$$

A crucial modeling choice is the covariance (kernel) function, which determines smoothness, stationarity, and other structural properties of the prior. Common kernel families include squared exponential, Matérn, γ-exponential, rational quadratic, and dot-product (see Chapter 4 of Williams and Rasmussen (2006) for an overview and

properties of each class). In our applications, we consider a Matérn 3/2 kernel

$$K_{3/2}(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3}r}{l}\right) \exp\left(-\frac{\sqrt{3}r}{l}\right),$$

where $r := ||x - x'||$ is Euclidean distance, σ^2 is the variance of the Gaussian process, and l is the length scale. The Matérn 3/2 kernel yields functions that are mean-square differentiable, and balances smoothness with flexibility.

Sampling GP with Monotonicity Constraints. We may want to sample functions that satisfy particular shape restrictions, e.g., boundedness, monotonicity, or convexity. Swiler, Gulian, Frankel, Safta and Jakeman (2020) surveys common strategies for incorporating constraints within Gaussian process regression. Our purpose is to sample from a constrained GP, which is equivalent to constrained GP regression without updating the prior. In general, there are two main categories - one enforces the constraints to hold globally through, for example, transforming the output of GP (Snelson, Ghahramani and Rasmussen, 2003) or imposing constraints on the coefficients of the spline functions (Maatouk and Bay, 2017; Shively *et al.*, 2009); the other relaxes the global constraints to constraints at a finite set of "virtual" points (Riihimäki and Vehtari, 2010). In our later applications, the sampling algorithms we employ fall into the first category. We provide the details in Appendix F.

2.3 Estimation and Inference

We distinguish two cases for estimation and inference. First, if the discrepancy function $d(f_\theta, f)$ is known in closed form and does not require estimation from data, the restrictiveness estimator is

$$\widehat{r}_M = \frac{\frac{1}{M}\sum_{m=1}^{M} d(\mathcal{F}_\Theta, f_m)}{\frac{1}{M}\sum_{m=1}^{M} d(\mathcal{F}_{\text{base}}, f_m)}, \quad f_m \sim \lambda_{\mathcal{F}}.$$

In this case sampling variation is the only source of uncertainty, and the standard-error formula is exactly the one provided in Fudenberg, Gao and Liang (2023). In principle, the standard error can be made arbitrarily small by taking M sufficiently large, but in practice M is constrained by computational resources, since each draw requires evaluating the discrepancy, which in turn involves solving the associated optimization problem.

Second, in many applications the discrepancy must be estimated from an i.i.d.

sample $S_n = (X_1, \ldots, X_n)$, as in Ellis and Neff (2025). Suppose

$$d(f_\theta, f) = \mathbb{E}[g(X, \theta; f)],$$

with empirical analog

$$d_n(f_\theta, f) = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta; f), \quad d_n(\mathcal{F}_\Theta, f) = \inf_{\theta \in \Theta} d_n(f_\theta, f).$$

The resulting estimator is

$$\widehat{r}_{n,M} = \frac{\frac{1}{M}\sum_{m=1}^{M} d_n(\mathcal{F}_\Theta, f_m)}{\frac{1}{M}\sum_{m=1}^{M} d_n(\mathcal{F}_{\text{base}}, f_m)}.$$

Here, inference must account for both Monte Carlo uncertainty and sampling error in d_n. We derive the corresponding asymptotic distribution of $\widehat{r}_{n,M}$ for fixed M as $n \to \infty$ and construct a feasible variance estimator that incorporates both sources of variability. The full procedure is provided in Appendix C.

In our empirical applications, the first example falls into the known-discrepancy case, whereas the second and third examples require estimation of the discrepancy and therefore follow the second procedure.

3 Restrictiveness for Structural Models

So far we have treated $\mathcal{F}_\Theta$ as an abstract class of prediction rules. In structural economic models, however, the "structural function" mapping covariates to outcomes often involve strategic selections, interactions, or equilibrium that result in endogeneity and model incompleteness issues. In this section we explain how to define and compute restrictiveness for structural models.

3.1 Generic Structural Models with Endogeneity

We start from a generic structural equation model with potentially endogenous covariates. For simplicity, write the structural system as

$$Y_i = f_{\theta_0}(Y_i, X_i, \epsilon_i), \quad X_i \perp \epsilon_i, \tag{2}$$

for some known mapping f_θ, a random element ϵ_i with a known distribution (without loss of generality)[1], and an unknown parameter $\theta \in \Theta$. Here Y_i may be a vector

[1] If the distribution of ϵ_i is unknown, it can be absorbed into θ.

that collects all variables endogenously generated under model (2), including both the outcome variable and any endogenous covariates. In contrast, X_i collects all exogenous covariates that are independent of the structural errors ϵ_i.

To fix ideas, we will repeatedly refer to the following simultaneous equation model of demand and supply as a working example.

Example 1 (Demand and Supply)**.** Consider the classic linear demand and supply simultaneous equation model:

$$Q_i = \alpha_1 + \beta_1 P_i + \gamma_1 X_{i1} + \epsilon_{i1}$$
$$P_i = \alpha_2 + \beta_2 Q_i + \gamma_2 X_{i2} + \epsilon_{i2}$$

where Q_i is quantity, P_i is price, X_{i1} and X_{i2} are exogenous demand and supply shifters. Writing $Y_i = (Q_i, P_i)'$, $\theta_0 := (\alpha, \beta, \gamma)'$, the structural form can be summarized as

$$Y_i = BY_i + \alpha + \Gamma X_i + \epsilon_i =: f_{\theta_0}(Y_i, X_i, \epsilon_i) \tag{3}$$

with

$$B := \begin{pmatrix} 0 & \beta_1 \\ \beta_2 & 0 \end{pmatrix}, \ \Gamma := \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}.$$

We first describe how to define restrictiveness when the structural form admits a reduced-form representation.

3.1.1 Restrictiveness via Reduced-Form Representation

Assumption 2 (Reduced-Form Representation)**.** *Assume that the structural equation model* (2) *admits the following reduced form representation*

$$Y_i = f_{\theta_0}(X_i, \epsilon_i) \tag{4}$$

for some known mapping f_θ and parameter $\theta_0 \in \Theta$.

Let $\mathcal{Y}$ be the space of distributions on the range space of f_θ, and let $\mathcal{F}_{RF}$ be a given eligible class of mappings that associates each covariate value $x \in \mathcal{X}$ with a conditional distribution $P_{Y|X=x} \in \mathcal{Y}$. For each structural parameter θ and each admissible distribution of (X, ε), the reduced form (4) induces a conditional distribution $P_{Y|X}(f_\theta)$. Hence, the reduced-form model class associated with the structural model

is given by

$$\mathcal{F}_{\Theta,RF} := \left\{P_{Y|X}(f_\theta) : \theta \in \Theta\right\} \subseteq \mathcal{F}_{RF}.$$

The above coincides with the standard definition of a statistical (or reduced-form econometric) model as a constrained class of data generating processes (DGPs) with the marginal distribution of the exogenous covariates X held fixed or unrestricted.

Given a primitive discrepancy function d on (conditional) distributions, such as KL-divergence or Wasserstein distance, we may define the reduced-form discrepancy function d_{RF} induced by d:

$$d_{RF}\left(f_\theta, g\right) := \left\{d\left(P_{Y|X}\left(f_\theta\right), g\right)\right\}, \quad \forall g \in \mathcal{F}_{RF}, \tag{5}$$

We then define the restrictiveness r according to Definition 1 based on the discrepancy function d_{RF} above.

Definition 2 (Restrictiveness via Reduced-Form Representation)**.** Under Assumption 2,the restrictiveness of model (2) is defined as

$$r := r\left(\mathcal{F}_{\Theta,RF}; \mathcal{F}_{RF}, d_{RF}\right),$$

i.e., the restrictiveness of reduced-form model $\mathcal{F}_{\Theta,RF}$ under eligible set $\mathcal{F}$ based on the discrepancy function d_{RF} according to Definition 1.

Given that $\mathcal{F}_{\Theta,RF}$ as the set of reduced form conditional distributions $P_{Y|X}$ implied by the structural model, we interpret the restrictiveness r as a measure of how much the structural model (2) restricts the space of admissible reduced forms (4), relative to the eligible set $\mathcal{F}$.

Remark 1 (Reduced-Form Additivity)**.** *When the reduced-form model* (4) *has an additive-error structure of the form*

$$Y_i = f_{\theta_0}\left(X_i\right) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | X_i] = 0, \tag{6}$$

researchers may only care about the mapping $f_\theta(x)$. In this case, we can simplify the definition of restrictiveness r. Specifically, we may take

- *$\mathcal{Y}$ to be the support of the outcome variable Y_i.*
- *$\mathcal{F}$ to a given eligible set of mappings form $\mathcal{X}$ to $\mathcal{Y}$*
- *d to be the mean squared ($L_{2,X}$) distance.*

We note that this simplification has also been considered in FGL, which proposes two "canonical" discrepancy functions, one based on KL divergence for distributions, and the other based on mean squared ($L_{2,X}$) distance for conditional expectation mappings. FGL also provides a discussion about why these two discrepancy functions are nicely "paired" with error functions underlying the definitions of completeness in Fudenberg, Kleinberg, Liang and Mullainathan (2022).

Example 1 (Demand and Supply: Continued)**.** A standard argument shows that the structural form (3) yields the reduced form

$$Y_i =: f_{\theta_0}(X_i, \epsilon_i) \equiv \overline{f}_{\theta_0}(X_i) + u_{\theta_0}(\epsilon_i)$$

where $\overline{f}_\theta(x) := (I - B)^{-1}(\alpha + \Gamma x)$ and $u_\theta(\epsilon) := (I - B)^{-1}\epsilon$.

This example features additive errors, and thus we may define restrictiveness using the model class of conditional means function $\mathcal{F}_\Theta := \{\overline{f}_\theta : \theta \in \Theta\}$, a constant baseline model $\mathcal{F}_{\text{base}} = \{c : c \in \mathbb{R}^2\}$, a given eligible set $\mathcal{F}$ of mappings from demand and supply shifters to conditional expectations of prices and quantities, and the mean-squared Euclidean distance[2] for $f, g \in \mathcal{F}$. The above then induces, writing $f \equiv (f_1, f_2)$, $d_{RF}(f, g) := \mathbb{E}_X\left[\|f(X) - g(X)\|^2\right]$ for $f, g \in \mathcal{F}$,

$$d_{RF}(\mathcal{F}_\Theta, f) = \inf_{\theta \in \Theta} \mathbb{E}\left[\left\|\overline{f}_\theta(X) - f(X)\right\|^2\right], \tag{7}$$

and in particular $d(\mathcal{F}_{\text{base}}, f) = \text{Var}(f_1(X)) + \text{Var}(f_2(X))$. Given an evaluation distribution $\lambda_\mathcal{F}$, the restrictiveness is given by

$$r = \frac{\mathbb{E}_{f \sim \lambda_\mathcal{F}}\left[\inf_{\theta \in \Theta} \mathbb{E}_X\left[\left\|\overline{f}_\theta(X) - f(X)\right\|^2\right]\right]}{\mathbb{E}_{f \sim \lambda_\mathcal{F}}\left[\text{Var}(f_1(X)) + \text{Var}(f_2(X))\right]}. \tag{8}$$

3.1.2 Restrictiveness under Structural-Form Error Additivity

Many econometric models are incomplete,in the sense that the structural form specificication may not admit a reduced-form representation as in Assumption 2. One class of such models are models with multiple equilibria, in which the reduced form takes the form of a correspondence rather than a function. We treat this scenario in Section 3.2, and show how the definition of restrictiveness can be adapted accordingly. Another important class of models without explicit reduced-form representations are partial equilibrium models identified via instrumental variables (IVs): for example, a

[2]One may use other distances such as absolute distance $\mathbb{E}_X\left[|f_1(X) - g_1(X)| + |f_2(X) - g_2(X)|\right]$.

demand model identified using exogenous demand and supply shifters without an explicit specification of the supply model. Such partial equilibrium models are prevalent in applied work, and often impose additivity of the structural error as in (9) below:

Assumption 3 (Structural-Form Error Additivity)**.** *Assume that the structural equation model* (2) *admits the following outcome representation*

$$Y_{o,i} = \Lambda \left(f_{\theta_0} \left(Y_{c,i}, X_i \right) + \epsilon_i \right), \tag{9}$$

for some known mappings Λ *and* f_θ *up to the parameter* $\theta_0 \in \Theta$*, with* $Y_{o,i}$ *denoting the outcome variable and* $Y_{c.i}$ *denoting the endogenous covariates.*

Example 1 (Demand and Supply: Continued)**.** To illustrate Assumption 3, consider the demand model equation without the supply-side specification:

$$Q_i = \alpha_0 + \beta_0 P_i + \gamma_0 X_{i1} + \epsilon_i, \quad \mathbb{E}[\, \epsilon_i |\, X_i] = 0, \tag{10}$$

so that the quantity Q_i corresponds to the outcome variable $Y_{o,i}$, the price P_i to the endogenous covariate $Y_{c,i}$, and $X_i = (X_{i1}, X_{i2})$ to the exogenous demand and supply shifters (IVs). Under the exogeneity condition $\mathbb{E}[\epsilon_i | X_i] = 0$ and the standard relevance condition, the demand parameters $\theta_0 := (\alpha_0, \beta_0, \gamma_0)$ can be identified without the supply equation specification. Here we cannot invert the structural demand model (10) to obtain a reduced-form representation as in Assumption 2, precisely because there is no explicit supply-side specification.

We now show how restrictiveness can be defined under Assumption 3. Let $\mathcal{Y}_o^{\text{pre-}g}$ be the domain of Λ, and let $\mathcal{F} : \mathcal{Y}_c \times \mathcal{X} \to \mathcal{Y}_o^{\text{pre-}\Lambda}$ denote the space of eligible mappings, and let $\mathcal{F}_\Theta \subseteq \mathcal{F}$ denote the set of mappings consistent with model (9), i.e.,

$$\begin{aligned} \mathcal{F}_\Theta &:= \{ f_\theta \left(y_c, x \right) + \mathbb{E} \left[\epsilon | x \right] : \ \mathbb{E} \left[\epsilon | x \right] = 0, \theta \in \Theta \} \\ &= \{ f_\theta \left(y_c, x \right) + f \left(y_c, x \right) - \mathbb{E} \left[f \left(y_c, x \right) | x \right] : \ \theta \in \Theta, f \in \mathcal{F} \}. \end{aligned}$$

Note that even though f_θ is parametric, $\mathcal{F}_\Theta$ is an infinite-dimensional functional space, since $\mathbb{E} \left[f \left(y_c, x \right) | x \right]$ is a nonparametric function given that the distribution of ϵ is unrestricted beyond the exogeneity condition $\mathbb{E} \left[\epsilon_i | x \right] = 0$.

We then push forward $\mathcal{F}_\Theta$ under Λ to the outcome space $\mathcal{Y}_o$ and define

$$\mathcal{F}_\Theta^\Lambda := \Lambda(\mathcal{F}_\Theta), \quad \mathcal{F}^\Lambda := \Lambda(\mathcal{F}),$$

and define restrictiveness based on any given discrepancy function on $\mathcal{F}^\Lambda$.

Definition 3 (Restrictiveness under Structural-Form Error Additivity)**.** Under Assumption 3, let d be any given discrepancy function on $\mathcal{F}^{\Lambda}$. We define the restrictiveness of model (9) as

$$r := r\left(\mathcal{F}_{\Theta}^{\Lambda}; \mathcal{F}^{\Lambda}, d\right),$$

i.e., the restrictiveness of pushed-forward model $\mathcal{F}_{\Theta}^{\Lambda}$ under eligible set $\mathcal{F}^{\Lambda}$ based on the discrepancy function d according to Definition 1.

In typical scenarios with scalar-valued Y_o, we can set the discrepancy function d as the mean-squared distance on the outcome space $\mathcal{Y}_o$,

$$d\left(f^{\Lambda}, g^{\Lambda}\right) := \mathbb{E}_{P_{X,Z}}\left[\left(f^{\Lambda}\left(Y_{c,i}, X_i\right) - g^{\Lambda}\left(Y_{c,i}, X_i\right)\right)^2\right], \quad \forall f^{\Lambda}, g^{\Lambda} \in \mathcal{F}^{\Lambda}.$$

Remark 2 (Simplification under Linearity of Λ)**.** *When Λ is linear, we can simplify the definition of restrictiveness r in a similar manner as in Remark 1, since the expectation $\mathbb{E}$ can be moved inside of Λ. In addition, when $\mathcal{F}$ and $\mathcal{F}_{\Theta}$ are linear spaces (as they typically are), we also have $\mathcal{F}^{\Lambda} = \mathcal{F}$ and $\mathcal{F}_{\Theta}^{\Lambda} = \mathcal{F}_{\Theta}$.*

Proposition 1. *Under 3, suppose that $\mathcal{F}_{\Theta}, \mathcal{F}$ are linear spaces, and Λ is linear. Let d be the mean-squared distance on $\mathcal{F}$. Then, for any $g \in \mathcal{F}$,*

$$d\left(\mathcal{F}_{\Theta}, g\right) = \inf_{\theta \in \Theta} \overline{d}\left(\overline{f}_{\theta}, \overline{g}\right),$$

where $\overline{f}_{\theta}(x) := \mathbb{E}[m_{\theta}(Y_c, x)|x]$, $\overline{g}(x) := \mathbb{E}\left[g\left(Y_c, x\right)|x\right]$, and

$$\overline{d}\left(\overline{f}, \overline{g}\right) := \mathbb{E}_{P_X}\left[\left(\overline{f}\left(X_i\right) - \overline{g}\left(X_i\right)\right)^2\right]. \tag{11}$$

Example 1 (Demand and Supply: Continued)**.** Applying Proposition 1 to the demand equation model in (10), we note that $\overline{d}$ is effectively a version of d_{RF} in (7) defined on the demand component only. In particular, $d(\mathcal{F}_{\text{base}}, \overline{f}) = \text{Var}(\overline{f}(X)))$ and thus

$$r = \frac{\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}\left[\inf_{\theta \in \Theta} \mathbb{E}_X\left[\left(\overline{f}_{\theta}(X) - f(X)\right)^2\right]\right]}{\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}\left[\text{Var}(\overline{f}(X))\right]},$$

a "subvector analog" of (8).

3.2 Structural Model with Multiple Equilibria

Some structural models may not produce a unique equilibrium, so that the reduced form of the model cannot be written as a regular function. Instead, the reduced form becomes a correspondence:

$$\overline{f}_{\theta_0} : \mathcal{X} \times \mathcal{E} \to 2^{\mathcal{Y}}.$$

which maps the covariates and errors (X_i, ϵ_i) to the set of equilibrium outcomes Y_i under parameter θ_0. In such cases the structural model does not pin down a single conditional distribution of Y_i given X_i: for a given (x, ϵ) there may be multiple admissible equilibria y.

To define restrictiveness in this context, let

$$\overline{\mathcal{F}}_\Theta := \left\{ f \in \mathcal{F} : \ f(x,\epsilon) \in \overline{f}_\theta(x,\epsilon) \ \forall (x,\epsilon), \ \theta \in \Theta \right\}$$

denote the space of selection mappings. Given a discrepancy function d between two conditional distributions of Y_i conditional on X_i, and define

$$d\left(\overline{f}_\theta, g\right) := \inf_{f \in \overline{\mathcal{F}}_\Theta} d\left(f, g\right) \tag{12}$$

This is the best approximation error when the equilibrium selection rule s is chosen optimally for each pseudo-truth g. Notice that we still just need to simulate g from $\lambda_{\mathcal{F}}$, a distribution of conditional distributions of Y_i given X_i; there is no need to simulate the reduced-form correspondences.

Remark 3 (Equilibrium Selection and Model Completion)**.** *Definition in* (12) *implicitly assumes that the model f has incorporated all relevant restrictions. Thus any $f \in \overline{\mathcal{F}}_\Theta$ is consistent with all stated model restrictions, which is why the definition uses the infimum approximation error. In some scenarios, one may want to impose an equilibrium selection restriction that picks a unique equilibrium. This additional equilibrium selection restriction effectively converts an incomplete model into a complete one and makes the minimization in equation* (12) *trivial: in other words, restrictiveness can be defined in the same way as in Definition 2.*

We illustrate the above with the following example.

Example 2 (Entry Game)**.** Consider the two-firm entry game as Tamer (2003):

$$y_{i1} = \mathbb{1}\left\{\alpha_1 + \beta_1 y_{i2} + \gamma_1 x_{i1} \geqslant \epsilon_{i1}\right\}$$
$$y_{i2} = \mathbb{1}\left\{\alpha_2 + \beta_2 y_{i1} + \gamma_2 x_{i2} \geqslant \epsilon_{i2}\right\}$$

where $\beta \leqslant 0$ to capture strategic substitutability between the two firms. The reduced form of the equilibrium of this entry game features multiple equilibrium. Writing

$$\pi_1(y_{i2}) := \alpha_1 + \beta_1 y_{i2} + \gamma_1 x_{i1},$$
$$\pi_2(y_{i1}) := \alpha_2 + \beta_2 y_{i1} + \gamma_2 x_{i2},$$

we have

$$y_i = \overline{f}_{\theta_0}(x_i, \epsilon_i) := \begin{cases} (0,0), & \epsilon_{i1} > \pi_1(0), \epsilon_{i2} > \pi_2(0) \\ (1,1), & \epsilon_{i1} \leqslant \pi_1(1), \epsilon_{i2} \leqslant \pi_2(1) \\ (1,0), & (\epsilon_{i1} \leqslant \pi_1(1), \epsilon_{i2} > \pi_2(1)) \\ & \text{or } (\pi_1(1) < \epsilon_{i1} \leqslant \pi_1(0), \epsilon_{i2} > \pi_2(0)) \\ (0,1), & \epsilon_{i1} > \pi_1(1), \epsilon_{i2} \leqslant \pi_2(1) \\ & \text{or } (\epsilon_{i1} > \pi_1(0), \pi_2(1) < \epsilon_{i2} \leqslant \pi_2(0)) \\ \{(0,1),(1,0)\} & \pi_1(1) < \epsilon_{i1} < \pi_1(0), \pi_2(1) < \epsilon_{i2} < \pi_2(0) \end{cases}$$

Let $\mathcal{F}$ be the set of all conditional choice-probability mappings

$$g : (x_1, x_2) \mapsto \big(p_{00}(x), p_{01}(x), p_{10}(x), p_{11}(x)\big),$$

with $p_{jk}(x) \geqslant 0$ and $\sum_{j,k} p_{jk}(x) = 1$. For each θ and each admissible selection rule s, the model implies a CCP vector

$$f_{\theta,s}(x) := \big(\mathbb{P}_\theta(y = (0,0) \mid X = x, s), \ldots, \mathbb{P}_\theta(y = (1,1) \mid X = x, s)\big),$$

and hence a conditional distribution $P^{\theta,s}_{Y|X}$. We may then choose an L^2 discrepancy between CCPs,

$$d(f,g) := \mathbb{E}\left[\sum_{j,k} (f_{jk}(X) - g_{jk}(X))^2\right],$$

and define $d(\overline{f}_\theta, g)$ via (12). Restrictiveness $r(\overline{f}_\theta, \mathcal{F})$ reflects how much the equilibrium structure and strategic interaction restrict the feasible CCPs, after optimally choosing an equilibrium selection rule for each pseudo-true mapping g. Given the discrepancy function d, restrictiveness can then be defined correspondingly.

3.3 Semiparametric Structural Models

Some structural models are semiparametric: some primitives are parametrically modeled, while others are left nonparametric. Often we may not wish to impose parametric

assumptions on the unobserved error terms. A generic representation of such models takes the form of

$$Y_i = f_{\theta_0,h_0}(X_i, \varepsilon_i), \tag{13}$$

where $f_{\theta,h}$ is a known mapping, h_0 is an infinite-dimensional nuisance parameter that captures the nonparametric distribution of ϵ_i and/or other nonparametric components of the model. Typically h_0 is restricted to lie in some function space H with encoded shape restrictions and regularity conditions.

From our perspective, this can be viewed as a special case of the incomplete-model framework above: for each θ there is a whole *family* of reduced forms indexed by h. Given a discrepancy d on $\mathcal{F}$ and an eligible pseudo-truth $g \in \mathcal{F}$, we can write

$$\overline{f}_\theta := \left\{ P_{Y|X}\left(f_{\theta,h}\right) : h \in \mathcal{H} \right\}$$

and define

$$d\left(\overline{f}_\theta, g\right) := \inf_{f \in \overline{f}_\theta} d\left(f, g\right).$$

Conceptually, $d(\overline{f}_\theta, g)$ measures how far the structural parameter θ alone restricts the reduced form in the presence of the flexible nonparametric component $h \in H$. The semiparametric structural model is then more or less restrictive depending on how tightly the union over θ and h of induced reduced forms $P_{Y|X}^{\theta,h}$ sits inside $\mathcal{F}$.

Example 3 (The BLP Multinomial Choice Model)**.** Consider the baseline BLP model (Berry, Levinsohn and Pakes, 1995) with market-level data on (s_{jt}, x_{jt}, z_{jt}), where j indexes a product, t indexes a market, and s_{jt} denotes market share of product j. The underlying model, parametrized by $\theta = (\beta, \sigma^2)$, is given by

$$y_{jt} := \mathbb{1}\left\{ \delta_{jt} + x_{jt}^{'}\nu_i + \epsilon_{ijt} \geqslant \max_k \left(\delta_{kt} + x_{kt}^{'}\nu_i + \epsilon_{ikt} \right) \right\}$$

with $\delta_{jt} = x_{jt}^{'}\beta + \xi_{jt}$, $\nu_i \sim \mathcal{N}\left(0, \sigma^2\right)$, and $\epsilon_{ijt} \sim TIEV\left(0, 1\right)$. Then

$$S_j\left(\delta_{jt}, x_t\right) := \mathbb{P}\left(y_{jt} = 1 | \, x_t, \xi_t\right) = \int \frac{\exp\left(\delta_{jt} + x_{jt}^{'}\nu\right)}{\sum_k \exp\left(\delta_{kt} + x_{kt}^{'}\nu\right)} \phi\left(\nu\right) d\nu$$

with $\delta_{jt} = S^{-1}\left(s_t; x_t, \sigma^2\right)$ and $\xi_{jt} = S^{-1}\left(s_t; x_t, \sigma^2\right) - x_{jt}^{'}\beta$ and $\mathbb{E}\left[\xi_{jt} | \, z_t\right] = 0$. Again, let $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^J$, and $\mathbb{E}\left[\xi_{jt} | \, x_t, z_t\right] = h_j - \overline{h}_j$. Hence, the BLP model class can be characterized by

$$\mathcal{F}_\Theta = \left\{ S\left(x_t^{'}\beta + h - \overline{h}; x_t, \sigma^2 \right) : h \in \mathcal{H}, \theta \in \Theta \right\}$$

where $\mathcal{H}$ is an appropriately chosen functional space. The eligible set $\mathcal{F}$ in this case is all mappings $\mathcal{F} = \left\{f : \mathcal{X} \times \mathcal{Z} \to \Delta^{J-1}\right\}$ where Δ^{J-1} denotes the $(J-1)$-dimensional simplex. We can set the discrepancy function as

$$d(f, g) := \mathbb{E}\left[\|f - g\|^2\right]$$

and define restrictiveness of $\mathcal{F}_\Theta$ based on

$$\inf_{f \in \mathcal{F}_\Theta} d(f, g) = \inf_{\theta \in \Theta, h \in \mathcal{H}} \mathbb{E}\left[\left\|S\left(x_t'\beta + h - \overline{h}; x_t, \sigma^2\right) - g\right\|^2\right]$$

To numerically approximate $\mathcal{H}$ by simulating M copies of h from $\lambda_{\mathcal{H}}$, a specified space of functions mapping from $\mathcal{X} \times \mathcal{Z}$ to $\mathbb{R}^J$, and define

$$\mathcal{F}_\theta^{(M)} := \left\{S\left(x_t'\beta + h^{(m)} - \overline{h}^{(m)}; x_t, \sigma^2\right) : m = 1, ..., M_1\right\}$$

which is a finite set. Based on the above, we compute

$$\inf_{f \in \mathcal{F}_\Theta} d^{(M)}(f, g) = \inf_{\theta \in \Theta} \min_{m=1,...,M} \mathbb{E}\left[\left\|S\left(x_t'\beta + h^{(m)} - \overline{h}^{(m)}; x_t, \sigma^2\right) - g\right\|^2\right]$$

which we use as an approximation for $d(\mathcal{F}_\Theta, g)$.

4 Related Concepts

A key aspect of our framework is to treat the discrepancy d as a *design choice*, to be selected based on the intended interpretation and the empirical context, rather than inherited from existing complexity measures that were built for different goals. Concretely:

- In the certainty-equivalent and discrete-choice applications, we use squared L^2 discrepancies on certainty-equivalent functions or choice probabilities, so that $d(f, g)$ can be read as an average prediction error with clear economic units.
- In structural models with endogeneity or multiple equilibria, we work with discrepancies between implied conditional distributions (in reduced form or structural forms), again chosen for their interpretability in terms of approximation error rather than for reproducing any particular capacity bound.

Our restrictiveness measure is defined at the level of prediction rules and a user-chosen discrepancy function d, with no reference to a particular sample size or estimation

procedure. This section discusses and develops formal connections between restrictiveness and related concepts in econometrics, statistics and machine learning.

4.1 Rademacher Complexity and VC Dimensions

This subsection explains how standard capacity measures from learning theory fit within our framework, and why their asymptotic behavior is driven by the discrepancies they encode.

Following Ellis and Neff (2025), suppose that $\mathcal{X}$ is a closed and bounded infinite subset of $\mathbb{R}^m$ for some finite $m \in \mathbb{N}$, X be a random vector of $\mathcal{X}$ with distribution P_X, and that the outcome space is $\mathcal{Y} = \{-1, 1\}$[3]. Let $\bar{\mathcal{F}}_\sigma = \{f : \mathcal{X} \to \{-1, 1\}\}$ be the set of all deterministic binary classifiers, and let $\mathcal{F}_\Theta \subseteq \bar{\mathcal{F}}_\sigma$ be a model class (for example, a parametric class indexed by $\theta \in \Theta$). The paper then takes the eligible set to be the full binary class $\mathcal{F} := \bar{\mathcal{F}}_\sigma$ with some probability measure $\lambda_\mathcal{F}$; and the discrepancy to be the correlation-based quantity

$$d_{\text{Rad}}(f, g) := \mathbb{E}_{X \sim P_X}\left[1 - f(X)g(X)\right], \quad f, g \in \bar{\mathcal{F}}_\sigma. \tag{14}$$

As we show in Appendix D, since f and g are binary-valued, the discrepancy $d_{\text{VC}}(f, g) = \mathbb{P}(f(X) \neq g(X))$ satisfies $d_{\text{Rad}} = 2d_{\text{VC}}$, so it induces exactly the same restrictiveness ordering. Ellis and Neff (2025) establishes that under, this discrepancy, all model classes with finite VC dimension are fully restrictive in the limit. Of course, Rademacher complexity and VC dimension are designed to control worst-case generalization error under adversarial labelings, so it is natural that produce degenerate restrictiveness measures in infinite domains. This degeneracy should thus be interpreted as a consequence of the underlying discrepancy rather than as a limitation of restrictiveness itself. An important point of our paper is that the the discrepancy function d should be chosen to maximize interpretability in a user-driven context-specific manner.

We discuss a related, though slightly different point, about the choice of discrepancy function in the next subsection.

[3]Ellis and Neff allow the outcome Y to take values in $[-1, 1]$. For expositional purposes, we restrict attention to the canonical binary case, since the most standard definitions of both the Rademacher complexity and the VC dimension are formulated for binary-valued function classes.

4.2 Discrepancy Associated with GMM Criterion Function

In many structural econometric models, parameters are estimated by minimizing the sample analog of a moment-based criterion function, such as the generalized method of moments (GMM) criterion. It is natural to ask whether such criteria implicitly define a discrepancy that can be used to evaluate restrictiveness. This subsection shows that the discrepancy function implicitly associated with the GMM criterion is focused on measuring violations of moment conditions, but not on the distance between prediction rules. Thus, although the GMM criterion is convenient for econometric identification and estimation, the resulting discrepancy is hard to interpret for our purposes.

Consider the additive-error model with instruments

$$Y_i = m_\theta(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid Z_i = z] = 0, \tag{15}$$

where X_i and Z_i are observed covariates and instruments, respectively. For simplicity, suppose that $m_\theta : \mathcal{X} \to \mathbb{R}$ is a parametric family indexed by $\theta \in \Theta$, and that the true conditional mean is given by some $g : \mathcal{X} \to \mathbb{R}$, which is not necessarily in $\{m_\theta : \theta \in \Theta\}$.

We proposed a *discrepancy* of the form

$$d(f, g) := \mathbb{E}\left[(f(X) - g(X))^2\right], \tag{16}$$

and defined restrictiveness of $\mathcal{F}_\Theta = \{m_\theta : \theta \in \Theta\}$ relative to an eligible set $\mathcal{F}$ using the induced quantity

$$d(\mathcal{F}_\Theta, g) := \inf_{f \in \mathcal{F}_\Theta} d(f, g).$$

The interpretation is straightforward: $d(\mathcal{F}_\Theta, g)$ is the best achievable mean-squared prediction error when approximating the pseudo-true rule g with elements of $\mathcal{F}_\Theta$.

By contrast, a standard population GMM criterion for (15) takes the form

$$Q_{\mathrm{GMM}}(\theta) := \mathbb{E}\left[g_i(\theta)\right]' W \, \mathbb{E}\left[g_i(\theta)\right], \quad g_i(\theta) := Z_i\big(Y_i - m_\theta(X_i)\big),$$

for some positive semi-definite weighting matrix W. Under correct specification there exists θ_0 such that

$$\mathbb{E}[g_i(\theta_0)] = 0, \quad Q_{\mathrm{GMM}}(\theta_0) = \min_{\theta \in \Theta} Q_{\mathrm{GMM}}(\theta) = 0.$$

In finite samples, the GMM estimator is defined as

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \widehat{Q}_{\mathrm{GMM}}(\theta), \quad \widehat{Q}_{\mathrm{GMM}}(\theta) := g_n(\theta)' W g_n(\theta), \quad g_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} g_i(\theta).$$

When the model is misspecified with $Q_{\mathrm{GMM}}(\theta) > 0$ for all $\theta \in \Theta$, the *GMM pseudo-true parameter* is defined by

$$\theta^\circ := \arg\min_{\theta\in\Theta} Q_{\mathrm{GMM}}(\theta).$$

A natural temptation is to interpret $Q_{\mathrm{GMM}}(\theta)$ as defining a discrepancy between m_θ and the true regression function g, and hence to define

$$d_{\mathrm{GMM}}(m_\theta, g) := Q_{\mathrm{GMM}}(\theta), \quad d_{\mathrm{GMM}}(\mathcal{F}_\Theta, g) := \inf_{\theta\in\Theta} Q_{\mathrm{GMM}}(\theta). \tag{17}$$

However, there are three reasons this interpretation is problematic. First, Q_{GMM} measures violations of the *moment condition*

$$\mathbb{E}\left[Z_i(Y_i - m_\theta(X_i))\right] = 0,$$

rather than predictive performance. Even in the correctly specified case, $Q_{\mathrm{GMM}}(\theta)$ is invariant to transformations of m_θ that leave the conditional moments $\mathbb{E}[Z_i(Y_i - m_\theta(X_i))]$ unchanged, and in general

$$Q_{\mathrm{GMM}}(\theta) = 0 \quad \nRightarrow \quad m_\theta(x) = g(x) \text{ for all } x, \tag{18}$$

unless the instruments are sufficiently rich to identify g pointwise. Thus as shown in the next example Q_{GMM} does not satisfy $d(f, g) = 0 \Rightarrow f = g$ (almost surely) which we require of our discrepancies.

Example 4 (Example via Irrelevant Instruments)**.** Let $W = (Y, X, Z)$ with $Y = \theta_0 X + \varepsilon$, $\mathbb{E}[\varepsilon \mid X, Z] = 0$, and $\mathbb{E}[X^2] > 0$. Consider $\mathcal{F}_\Theta = \{f_\theta(x) = \theta x : \theta \in \mathbb{R}\}$ and the linear IV/GMM moment $\psi(W, \theta) = Z\,(Y - \theta X)$ with population criterion $Q(\theta) = \big(\mathbb{E}[\psi(W,\theta)]\big)' \,\Omega^{-1}\, \big(\mathbb{E}[\psi(W,\theta)]\big)$ for any positive definite Ω. Take instruments that are valid but irrelevant: $\mathbb{E}[Z] = 0$ and $Z \perp (X, \varepsilon)$, so $\mathbb{E}[Z\varepsilon] = 0$ and $\mathbb{E}[ZX] = 0$. Then, for every θ, $\mathbb{E}[\psi(W,\theta)] = \mathbb{E}[ZY] - \theta\mathbb{E}[ZX] = \mathbb{E}[Z(\theta_0 X + \varepsilon)] - \theta\mathbb{E}[ZX] = 0$, hence $Q(\theta) = 0$ for all θ and the set of population minimizers is all of Θ. In contrast, the predictive L^2 discrepancy between f_θ and the pseudo-truth $g(x) = \theta_0 x$ is $d_2(f_\theta, g) = \mathbb{E}[(f_\theta(X) - g(X))^2] = (\theta - \theta_0)^2\mathbb{E}[X^2]$, which is strictly positive whenever $\theta \neq \theta_0$. Thus $Q_{\mathrm{GMM}}(\theta) = 0$ does not imply $f_\theta = g$, and the GMM value depends on the choice of instruments rather than on predictive distance.

Second, Q_{GMM} depends not only on the prediction rule m_θ and the data-generating process, but also on the choice of instruments Z_i and the weighting matrix W. Two researchers analyzing the same model class $\mathcal{F}_\Theta$ and the same set of pseudo-truths g but

using different instruments or weights would obtain different values of $d_{\text{GMM}}(\mathcal{F}_\Theta, g)$, even though the underlying space of prediction rules is unchanged. In our framework, by contrast, restrictiveness is a property of $\mathcal{F}_\Theta$ relative to an eligible set $\mathcal{F}$ and an evaluation distribution $\lambda_{\mathcal{F}}$, not of auxiliary choices that are convenient or optimal for estimation.

Third, Q_{GMM} is not directly interpretable in the unit of prediction error. By definition,

$$Q_{\text{GMM}}(\theta) = \mathbb{E}\left[Z_i\big(g(X_i) - m_\theta(X_i)\big)\right]' W\, \mathbb{E}\left[Z_i\big(g(X_i) - m_\theta(X_i)\big)\right], \tag{19}$$

which is a quadratic form in averaged and *instrumented* residuals. In general there is no simple relationship between $Q_{\text{GMM}}(\theta)$ and the predictive discrepancy $d(m_\theta, g)$ in (16), even up to monotone transformations.

The same critique also applies to the minimum-distance (MD) estimation criterion, where the objective functions are quadratic forms in deviations of low-dimensional summaries from their targets, and the scale is driven by arbitrary choices of normalization and weighting. These features make the criterion convenient for estimation and inference but are ill-suited to serve as discrepancies.

For these reasons, we do not use GMM (and MD) criterion functions as discrepancies for the purposes of restrictiveness or completeness. Our proposal for moment-equality models is instead to: (i) use moment conditions to define the *eligible set* of admissible prediction rules, and (ii) evaluate restrictiveness with respect to discrepancies such as (16) that are explicitly defined on prediction rules and have a clear interpretation in terms of approximation or prediction error.

4.3 Limit of the Average-Case Learning Curve

We now connect restrictiveness to the limit of the *average-case learning curve* for a given estimation procedure. We show that restrictiveness can be interpreted as the normalized limit of a "noise-free" version of the average-case learning curve that focuses on functional-space approximation errors.

Specifically, consider the following setup. Given a probability measure P_X on $\mathcal{X}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, let the discrepancy d be given by, for all $f, g \in \mathcal{F}$,

$$d(f, g) := \mathbb{E}_{P_X}\left[\ell(f(X), g(X))\right]. \tag{20}$$

Suppose that ℓ is continuous, nonnegative, and uniformly bounded so that, for any

model class $\mathcal{F}_\Theta$ and evaluation distribution $\lambda_\mathcal{F}$, the random variable $d(\mathcal{F}_\Theta, f)$ is also uniformly bounded and thus

$$d(\mathcal{F}_\Theta, f) = \inf_{f_\theta \in \mathcal{F}_\Theta} d(f_\theta, f) = \inf_{f_\theta \in \mathcal{F}_\Theta} \mathbb{E}_{X \sim P_X}[\ell(f_\theta(X), f(X))],$$

is well-defined and uniformly bounded.

For a given pseudo-truth $f \in \mathcal{F}$, define the sample

$$S_n(f) := \{(X_i, Y_i)\}_{i=1}^n, \quad X_i \overset{\text{i.i.d.}}{\sim} P_X, \quad Y_i = f(X_i). \tag{21}$$

Let $\mathcal{A}$ be an estimation algorithm that maps the sample into a parameter estimate $\hat{\theta}_n = \mathcal{A}(S_n(f)) \in \Theta$, thus producing an estimated prediction rule $\hat{f}_n := f_{\hat{\theta}_n} \in \mathcal{F}_\Theta$.

We then define the average-case learning curve associated with $(\mathcal{F}_\Theta, \mathcal{A})$, where the average is taken over pseudo-truths $f \sim \lambda_\mathcal{F}$ and over samples $S_n(f)$.

Definition 4 (Noise-Free Average-Case Learning Curve)**.** For each $n \geqslant 1$, the *average-case learning curve* of $(\mathcal{F}_\Theta, \mathcal{A})$ is

$$L_n(\mathcal{F}_\Theta, \mathcal{A}) := \mathbb{E}_{f \sim \lambda_\mathcal{F}} \mathbb{E}_{S_n(f)} \left[R(\hat{f}_n; f) \right], \quad R(h; f) := \mathbb{E}_{P_X}[\ell(h(X), f(X))], \tag{22}$$

We impose the following risk-consistency assumption.

Assumption 4 (Risk-Consistency)**.** *The estimation algorithm $\mathcal{A}$ is risk-consistent for $\mathcal{F}_\Theta$ (relative to $\lambda_\mathcal{F}$), i.e., for $\lambda_\mathcal{F}$-almost every $f \in \mathcal{F}$,*

$$\lim_{n \to \infty} \mathbb{E}_{S_n(f)} \left[R(\hat{f}_n; f) \right] = \inf_{f_\theta \in \mathcal{F}_\Theta} R(f_\theta; f) = d(\mathcal{F}_\Theta, f). \tag{23}$$

Risk-consistency is satisfied by a wide range of regularized empirical risk minimization procedures under suitable conditions on $\mathcal{F}_\Theta$, P_X, and ℓ, and hence should be interpreted as a very mild requirement.

We now present our main result of this subsection, establishing the connection of restrictiveness to the limit of the average-case learning curve.

Proposition 2 (Restrictiveness as Normalized Limit of the Learning Curve)**.** *Under the setup of this subsection and, in particular, Assumption 4, we have*

$$\lim_{n \to \infty} L_n(\mathcal{F}_\Theta, \mathcal{A}) = \mathbb{E}_{\lambda_\mathcal{F}} \left[d(\mathcal{F}_\Theta, f) \right].$$

Moreover, if $\mathcal{A}_{base}$ is a risk-consistent estimator for $\mathcal{F}_{base}$, then

$$r(\mathcal{F}_\Theta, \mathcal{F}) = \frac{\lim_{n \to \infty} L_n(\mathcal{F}_\Theta, \mathcal{A})}{\lim_{n \to \infty} L_n(\mathcal{F}_{base}, \mathcal{A}_{base})}.$$

Proposition 2 shows that, under mild assumptions, restrictiveness can be interpreted as the ratio of two long-run average-case function-space *approximation errors*: one achieved by the best-fitting member of the model class $\mathcal{F}_\Theta$, the other by the baseline model class $\mathcal{F}_{\text{base}}$. Low restrictiveness corresponds to a model whose best-fitting members achieve low asymptotic risk across pseudo-truths $f \sim \lambda_{\mathcal{F}}$, while high restrictiveness corresponds to a model that, on average, cannot reduce risk much relative to the baseline. This is analogous to but also different from the standard notion of learning curve in machine learning, which features irreducible errors arising from the noise term ϵ_i in the data generating process $Y_i = f(X_i) + \epsilon_i$, which contrasts with our noise-free setup (22). Hence, restrictiveness can be viewed as a normalized noise-free analog of the limit of the standard average-case learning curve. This again illustrates our point that restrictiveness seeks to reveal that functional-form flexibility of the model in question.

See Appendix E for a more detailed explanation of the relationship and differences between restrictiveness and the standard learning curve, in the context of Gaussian process regressions.

5 Applications

This section applies our restrictiveness framework to three settings: certainty equivalents, multinomial choice with exogenous product characteristics, and multinomial choice with endogenous prices. In the certainty-equivalent application, the relative importance of parameters mirrors that in FGL, but restrictiveness is uniformly higher when models are evaluated over the full continuum of lotteries. In discrete-choice models without endogeneity, restrictiveness is driven primarily by flexibility in mean utility, and the restrictiveness of commonly used empirical specifications differ meaningfully despite the theoretical generality of mixed logit. When prices are endogenous, moment restrictions substantially increase restrictiveness and alter model rankings. Together, the first two applications illustrate how the framework evaluates restrictiveness for functional models on continuum domains, while the third shows how it extends naturally to structural, semiparametric models with endogeneity.

5.1 Cumulative Prospect Theory

In the first example, we revisit the evaluation of "CPT", a popular three-parameter specification of Cumulative Prospect Theory (Tversky and Kahneman, 1992), and "DA," a two-parameter specification of Disappointment Aversion (Gul, 1991), which were studied in FGL. That paper evaluated model restrictiveness in predicting the certainty equivalents for a set of 25 binary lotteries from Bruhin, Fehr-Duda and Epper (2010). Here we are able to evaluate model restrictiveness in predicting certainty equivalents for any lottery, using our general framework.

Setting

Each lottery is characterized by a tuple $x = (\bar{z}, \underline{z}, p)$, where $\bar{z} > \underline{z} \geqslant 0$ are the possible prizes, and p is the probability of the larger prize. A prediction rule is a function $f : \mathcal{D} \to \mathbb{R}$ that maps a lottery to its certainty equivalent, where $\mathcal{D} = \{(\bar{z}, \underline{z}, p) \in [0,1]^3 : \bar{z} \geqslant \underline{z}\}$.

Both CPT and DA specify prediction rules of the form

$$f(\bar{z}, \underline{z}, p) = v^{-1}(w(p)v(\bar{z}) + (1 - w(p))v(\underline{z}))\,,$$

where $v(z) = z^\alpha$, and the two models differ in their probability weighting functions. For CPT,

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}, \quad (\alpha, \gamma, \delta) \in [0,1]^2 \times \mathbb{R}_+,$$

while for DA,

$$w(p) = \frac{p}{1 + (1-p)\eta}, \quad (\alpha, \eta) \in [0,1] \times (-1, \infty).$$

The baseline model class is a singleton that fixes parameters at their benchmark values: $(\alpha, \gamma, \delta) = (1,1,1)$ for CPT and $(\alpha, \eta) = (1, 0)$ for DA.

To study the contribution of individual parameters, we consider submodels obtained by fixing one or more parameters at their baseline values while allowing the remaining parameters to vary. For example, CPT(α, γ) denotes the CPT model with δ fixed at its baseline value. Analogous variants are considered for DA. We refer to the unrestricted models as CPT(α, γ, δ) and DA(α, η).

Eligible Set, Evaluation Distribution, and Discrepancy

We define the eligible set $\mathcal{F}$ to be all prediction rules $f : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} = \{(\bar{z}, \underline{z}, p) \in [0,1]^3 : \bar{z} \geqslant \underline{z}\}$, satisfying two criteria: (i) $\underline{z} \leqslant f(\bar{z}, \underline{z}, p) \leqslant \bar{z}$; and (ii) $f(\bar{z}, \underline{z}, p)$ is monotone increasing with respect to a partial order $\succ$ on vectors $\mathbf{x} = (\bar{z}, \underline{z}, p)$. Specifically, we define $\mathbf{x}_1 \succ \mathbf{x}_2$ if $\bar{z}_1 \geqslant \bar{z}_2, \underline{z}_1 \geqslant \underline{z}_2$, and $p_1 \geqslant p_2$, where $\mathbf{x}_1 = (\bar{z}_1, \underline{z}_1, p_1)$ and $\mathbf{x}_2 = (\bar{z}_2, \underline{z}_2, p_2)$. A function $f(\bar{z}, \underline{z}, p)$ is monotone increasing if $\mathbf{x}_1 \succ \mathbf{x}_2 \Rightarrow f(\mathbf{x}_1) \geqslant f(\mathbf{x}_2)$.

We define a constrained GP prior $\lambda_{\mathcal{F}}$ on the prediction rule f. To sample f, we proceed in two steps. First, we draw a monotone increasing function $g(\bar{z}, \underline{z}, p)$ from a constrained Gaussian process with a Matérn 3/2 kernel. [4] Second, we map g to f via the sigmoid transformation $f(\bar{z}, \underline{z}, p) = \underline{z} + (\bar{z} - \underline{z})\sigma(g(\bar{z}, \underline{z}, p))$, ensuring that f takes values in $[\underline{z}, \bar{z}]$. We use $M = 2000$ draws from $\lambda_{\mathcal{F}}$ in our implementation. Random samples from the constrained GP prior satisfy the required monotonicity constraints and exhibit non-flat behavior; see Online Appendix F.1 for illustrative plots. The discrepancy between f_1 and f_2 is defined as L^2-norm of their difference.

Results

Table 1 compares our restrictiveness estimates with those reported in FGL. For the CPT specification, two main observations emerge. First, the relative contribution of the three parameters is the same as in FGL–δ contributes the most, followed by γ, and then α. Comparing the full CPT model (0.56) with models that drop each parameter shows this ordering: dropping δ increases restrictiveness to 0.77, dropping γ increases it to 0.67, and dropping α increases only to 0.59. Second, the absolute level of restrictiveness is uniformly higher in our estimates. For every CPT specification, our restrictiveness estimates exceed those in FGL, often by a large margin–e.g., 0.56 vs. 0.28 for the full model, and 0.77 vs. 0.51 for (α, γ). In addition, the range of restrictiveness across CPT specifications becomes much narrower in our results (0.56 to 0.92) than in FGL's (0.28 to 0.91). Both patterns are consistent with the conceptual difference between the two approaches: we measure restrictiveness over the entire eligible functional space, whereas FGL evaluate restrictiveness only over predictions on a finite-sample dataset of lotteries. When the eligible set expands, models naturally

[4] As robustness checks, we (i) replace the Matérn 3/2 kernel with a squared exponential kernel, and (ii) replace the GP draws with spline basis draws; the qualitative ranking of restrictiveness across models is unchanged. See Online Appendix F.1.

Table 1: Restrictiveness for Certainty Equivalents

	#Param	New	Old
CPT Spec.			
α, δ, γ	3	0.56	0.28
		(0.00)	(0.00)
α, γ	2	0.77	0.51
		(0.00)	(0.01)
γ, δ	2	0.59	0.37
		(0.00)	(0.00)
α, δ	2	0.67	0.49
		(0.01)	(0.01)
α	1	0.92	0.91
		(0.00)	(0.01)
γ	1	0.86	0.59
		(0.00)	(0.01)
δ	1	0.69	0.68
		(0.01)	(0.01)
DA Spec.			
α, η	2	0.67	0.47
		(0.01)	(0.01)
η	1	0.69	0.69
		(0.01)	(0.01)

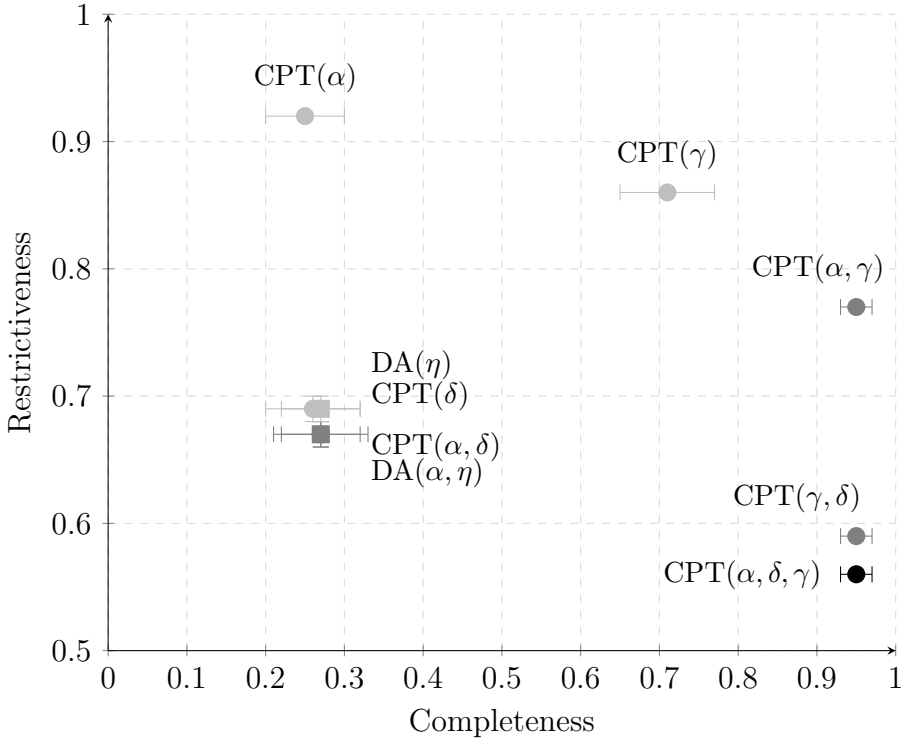
Notes: New: estimates from this paper. Old: estimates from Fudenberg, Gao, and Liang (2023).

appear more restrictive.

For the DA specification, our results suggest that adding parameter α makes the model only slightly more flexible–lowering restrictiveness from 0.69 (with only η) to 0.67 (with both α and η). In contrast, FGL report a much larger decrease, from 0.69 to 0.47. Hence, while both papers agree that α increases flexibility in DA, our framework finds its effect to be quantitatively much smaller.

Figure 1 illustrates our results by comparing model restrictiveness and completeness across various CPT and DA specifications. As in FGL, several specifications lie strictly inside the restrictiveness–completeness Pareto frontier, meaning that there exists another model that is both more complete and more restrictive. The undominated models are preferred: they rule out more regularities, yet capture the regularities that are present in real data. In FGL, the interior models are CPT(α, δ) and DA(α, η). In

Figure 1: Model Comparison by Their Completeness and Restrictiveness



our results, these two remain interior, but we find that CPT(δ) and DA(η) also fall inside the frontier.

5.2 Multinomial Choice Models

In the second example, we evaluate the restrictiveness of three multinomial choice models commonly used in industrial organization: multinomial logit (MNL), nested logit (NL), and mixed logit (MXL). Theoretical work, going back to McFadden and Train (2000), shows that mixed logit models can approximate any random-utility model arbitrarily well under mild conditions. More recently, Chang *et al.* (2022) provide necessary and sufficient conditions under which mixed logit models span the full nonparametric random-utility class. While these results are theoretically powerful, empirical applications of mixed logit typically impose strong parametric structure– most commonly assuming normally distributed random coefficients and linear utility in product characteristics. As a result, the mixed logit models used in practice are far

less flexible than the general formulation studied in theory. This makes it meaningful to compare the restrictiveness of the specifications actually estimated in empirical industrial organization.

Setting

We consider markets indexed by $m = 1, \ldots, M$. In each market m, the choice set is $\mathcal{J} = \{0, 1, \ldots, J\}$, where $j = 0$ denotes the outside option ("no purchase"). Let s_{jm} denote the market share of product j in market m, with the outside share given by $s_{0m} = 1 - \sum_{j \geqslant 1} s_{jm}$. Each product j in market m has a vector of observed characteristics $x_{jm} \in \mathbb{R}^K$, and we write $X_m = (x_{1m}, \ldots, x_{Jm})$ for the collection of covariates for all products in market m. In this application, we focus on a purely exogenous setting; see our third application in Section 5.3 for a structural multinomial choice model with endogeneity.

Define the model as the prediction rule p_θ that maps covariates X_m in market m to a $J \times 1$ vector of product share $p_m(X_m; \theta) = (p_{1m}(X_m; \theta), \ldots, p_{Jm}(X_m; \theta))'$. The baseline model class is a singleton consisting of the uniform share vector $p_{base} = 1/(J+1) \cdot \iota$, where ι is a $J \times 1$ vector. We evaluate the restrictiveness of three widely used discrete-choice specifications: multinomial logit (MNL), nested logit (NL), and mixed logit (MXL), which differ in the substitution patterns they allow: MNL implies independence of irrelevant alternatives, NL allows within-nest correlation, and MXL incorporates random taste heterogeneity. Formal model definitions are given in Appendix F.

We use the cereal dataset from Nevo (2000), which contains $M = 94$ markets and $J = 24$ products. Product characteristics include a continuous price variable and a binary indicator for "mushy" cereals. Here we treat prices as exogenous; consequently, restrictiveness is measured solely with respect to the functional form of $p_m(.)$.

Eligible Set, Evaluation Distribution, and Discrepancy

In this setting, an eligible set $\mathcal{F}$ is a collection of prediction rules s mapping covariates X_m to market-share vectors s_m. Our specification of the eligible set here is motivated by theoretical work on the flexibility of mixed logit models. Starting from a parametric MXL structure, we consider three variants that differ in which components of utility are allowed to be general functions of product characteristics x_{jm}, possibly subject to monotonicity restrictions. Specifically, we consider: (i) an eligible set in

which both the common mean utility and individual heterogeneity are allowed to be general functions ("NP Both"), (ii) an eligible set in which the mean utility is allowed to be a general function while individual heterogeneity remains parametric ("NP Mean"), and (iii) an eligible set in which the mean utility remains parametric while individual heterogeneity is allowed to be a general function ("NP Individual"). We introduce an evaluation distribution $\lambda_{\mathcal{F}}$ over each eligible set, defined using (constrained) Gaussian process priors[5], and draw 100 functions from this distribution. Full details are provided in Appendix F.

For the discrepancy, we use the squared L^2-norm as our distance metric between product shares $p_m(X_m;\theta)$ and $s_m(X_m)$, taking expectation over covariates X_m:

$$d\left(p_\theta, s\right) = \mathbb{E}_{X_m}\left[\sum_{j=1}^{J}\left|p_{jm}(X_m;\theta) - s_{jm}(X_m)\right|^2\right].$$

Results

Table 2 reports restrictiveness for the multinomial choice models under three eligible sets. When both the mean and the individual heterogeneity are modeled nonparametrically (NP Both), all three models remain restrictive, with MNL the most restrictive (0.154) and NL and MXL very similar and less restrictive (0.113 and 0.112). Although mixed logit is theoretically the most flexible, the parametric structure imposed in empirical implementations makes NL and MXL exhibit very similar restrictiveness in our application. Our results also show that the restrictiveness of these models is driven almost entirely by the mean utility component. When the eligible set permits a nonparametric mean utility but keeps individual heterogeneity parametric (NP Mean), restrictiveness is essentially the same as when the eligible set allows both components to be nonparametric (NP Both), and the ranking across models is unchanged. In contrast, when the eligible set keeps the mean utility parametric but allows nonparametric heterogeneity (NP Individual), restrictiveness is near zero for all models, with MXL being the least restrictive. In the "NP Individual" eligible set, flexibility enters only through individual-specific terms and is largely integrated out when forming market shares; as a result, the eligible set is not materially expanded relative to parametric mixed logit. By contrast, allowing a nonparametric mean utility relaxes the functional form of the common utility component and effectively enlarges the eligible

[5]Robustness checks that vary the kernel choice and the function-draw procedure yield similar results; see Online Appendix F.2.

Table 2: Restrictiveness of Multinomial Choice Models (No Endogeneity)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.154	0.008
NP Both	NL	0.113	0.005
NP Both	MXL	0.112	0.005
NP Individual	MNL	0.002	0.000
NP Individual	NL	0.002	0.000
NP Individual	MXL	0.002	0.000
NP Mean	MNL	0.157	0.009
NP Mean	NL	0.116	0.005
NP Mean	MXL	0.120	0.005

Table 3: Completeness of Multinomial Choice Models (No Endogeneity)

Model	**Complete.**	**SE**
MNL	0.396	0.039
NL	0.396	0.039
MXL	0.397	0.040

set beyond the parametric specification.

Table 3 reports completeness measures, which quantify the fraction of predictable variation in market shares captured by each model relative to a flexible statistical benchmark.[6] Using the squared prediction error of market shares as the loss function, we find that models exhibit nearly identical completeness when applied to the cereal data, indicating that the additional parameters in NL and MXL models do not provide meaningful additional predictive power for this dataset.

5.3 Multinomial Choice with Endogeneity

In the third example, we compare the restrictiveness of multinomial choice models in a setting where product characteristics (such as price) may be endogenous. Although the parametric functional forms coincide with those in the previous section, the presence of endogeneity fundamentally changes the evaluation: valid instruments are required, and the resulting specifications incorporate additional moment conditions. These moment restrictions make the models semiparametric, so we need to employ the procedures outlined in Section 3.3. We find that endogeneity raises re-

[6] In this setting, a fully nonparametric estimation approach will perfectly fit the observed shares.

strictiveness for all models. Unlike in the case without endogeneity, MXL becomes the least restrictive, while NL delivers restrictiveness nearly identical to MNL.

Setting

Unlike in Section 5.2, we now allow price to be endogenous. We address this endogeneity using the BLP instruments constructed in Nevo (2000), selecting the two instruments that are most strongly correlated with price from the full set of 20. [7] The model class is defined as in Section 3.3. The structural mapping S corresponds to the mixed logit functional form in the BLP framework, while for the multinomial logit and nested logit specifications we replace S with the corresponding functional forms used in the previous section. See Appendix F for details.

Eligible Set, Evaluation Distribution, and Discrepancy

To make the results comparable to the no-endogeneity case in Section 5.2, we use the same definition for the eligible sets–"NP Both," "NP Mean," and "NP Individual," which differ in whether the mean utility and/or individual heterogeneity are allowed to be nonparametric–along with its corresponding evaluation distribution, to generate the pseudo-true prediction rules. We draw 50 functions from the evaluation distribution $\lambda_{\mathcal{F}}$.

The discrepancy function follows Section 3.3 and is defined as

$$\inf_{f \in \mathcal{F}_\Theta} d(f, g) = \inf_{\theta \in \Theta, h \in \mathcal{H}} \mathbb{E}\left[\left\| S\left(x_t' \beta_0 + h - \overline{h}; x_t, \sigma^2 \right) - g \right\|^2 \right],$$

where $\bar{h}(Z) = \mathbb{E}[h(X, Z)|Z]$. In the implementation, $\bar{h}$ is obtained by projecting h onto higher-order polynomial functions of the instruments Z. When computing the infimum over h, we draw 20 candidate h-functions from a Gaussian process prior with Matérn-3/2 kernel. [8]

Results.

Table 4 reports restrictiveness for the discrete-choice models when price is treated as endogenous. Relative to the no-endogeneity case, restrictiveness is substantially

[7] In Online Appendix F.3, we also consider selecting the three instruments most strongly correlated with price; the qualitative ranking of restrictiveness across the three models is unchanged.

[8] Robustness checks that vary the kernel choice and the function-draw procedure yield similar results; see Online Appendix F.3.

Table 4: Restrictiveness of Multinomial choice Models (With Endogeneity)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.758	0.012
NP Both	NL	0.758	0.012
NP Both	MXL	0.679	0.020
NP Individual	MNL	0.631	0.009
NP Individual	NL	0.631	0.009
NP Individual	MXL	0.587	0.014
NP Mean	MNL	0.729	0.012
NP Mean	NL	0.729	0.012
NP Mean	MXL	0.631	0.011

Table 5: Completeness of Multinomial Choice Models (With Endogeneity)

Model	**Complete.**	**SE**
MNL	0.301	0.035
NL	0.302	0.035
MXL	0.335	0.040

higher for each model under each eligible set. For example, under eligible set "NP Both", restrictiveness for MNL increases from 0.154 to 0.758, and for MXL from 0.112 to 0.679. This reflects the additional constraints imposed by the moment conditions: each structural model must now match both its functional form and the exogeneity restrictions, which makes it harder to approximate the pseudo-true share.

Across all three eligible sets, MXL is the least restrictive model, while MNL and NL yield nearly identical restrictiveness. This contrasts with the no-endogeneity case, where both NL and MXL are less restrictiveness than MNL; here, the nesting parameter in NL does not play a role in relaxing restrictiveness once endogeneity restrictions are imposed. Moreover, as in the no-endogeneity case, the role of the mean utility component remains central. Restrictiveness under the "NP Mean" eligible set is close to that under "NP Both" for all models. By contrast, the "NP Individual" eligible set delivers somewhat lower restrictiveness, but the levels remain high, ranging from 0.587 to 0.631. Moment restrictions prevent a parametric mean utility component from yielding restrictiveness values close to zero.

Table 5 reports completeness for the three models. Mixed logit achieves slightly higher completeness than MNL and NL, while MNL and NL remain nearly identical.

6 Conclusion

In this paper, we further develop the notion of restrictiveness proposed in Fudenberg, Gao and Liang (2023) to functional and structural settings, which substantially enlarges the scope under which the restrictiveness measure can be used to evaluate the a priori economic structures imposed by a specific economic or econometric model. Restrictiveness, together with completeness (Fudenberg, Kleinberg, Liang and Mullainathan, 2022), provide a general-purpose toolkit for theoretical and applied researchers to evaluate the "value added" by economic theory, or more generally any restrictions imposed by any domain-specific structures, beyond purely statistical aspects of their models.

Our work also suggests several directions of future research: (1) Investigate the restrictiveness-completeness frontier in a larger variety of economic applications, and provide a more comprehensive view and understanding of the "value added" by economic theory in various empirical settings. (2) Investigate how to incorporate recent advances in the Bayesian nonparametrics and machine learning literature to make the procedure of sampling from the prior distribution $\lambda_{\mathcal{F}}$ more computationally efficient and robust, especially under various shape restrictions. (3) Explore the use of restrictiveness as a regularization device in machine learning.

We think that direction (3) is particularly interesting. The idea is to extend our framework by regularizing empirical risk minimization with a penalty term $(1 - r)$, where r measures the model's restrictiveness. Unlike syntactic complexity measures (such as parameter counting in AIC or BIC), such a penalty would favor models based on their population-level structural content and adherence to economic constraints. This approach could be applied both to selection across non-nested model classes and to regularization within parametric families, favoring regions of the parameter space that induce more restrictive prediction rules. An important empirical question is whether such restrictiveness-based regularization lead to better finite-sample performance in estimation and prediction than standard complexity penalties.

References

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, **63** (4), 841–890.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, **71** (356), 791–799.

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

Chang, H., Narita, Y. and Saito, K. (2022). Approximating choice data by discrete choice models. *arXiv preprint arXiv:2205.01882*.

Ellis, K. and Neff, S. (2025). Model complexity and restrictiveness. *Working Paper*.

Fudenberg, D., Gao, W. and Liang, A. (2023). How flexible is that functional form? quantifying the restrictiveness of theories. *Review of Economics and Statistics*, pp. 1–50.

—, Kleinberg, J., Liang, A. and Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, **130** (4), 956–990.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, **59** (3), 667–686.

Le Gratiet, L. and Garnier, J. (2015). Asymptotic analysis of the learning curve for gaussian process regression. *Machine learning*, **98** (3), 407–433.

Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, **49** (5), 557–582.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, **15** (5), 447–470.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The Rand journal of economics*, pp. 395–421.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, pp. 645–652.

Shively, T. S., Sager, T. W. and Walker, S. G. (2009). A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71** (1), 159–175.

Snelson, E., Ghahramani, Z. and Rasmussen, C. (2003). Warped gaussian processes. *Advances in neural information processing systems*, **16**.

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C. and Jakeman, J. D. (2020). A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, **1** (2).

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, **70** (1), 147–165.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, **5** (4), 297–323.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA.

Appendix

A Proof of Proposition 1

Proof. Any element $f \in \mathcal{F}_\Theta$ can be written as

$$f(y_c, x) = f_\theta(y_c, x) + h(y_c, x) - \overline{h}(x),$$

for some $\theta \in \Theta$ and $h \in \mathcal{F}$, where $\overline{h}(x) := \mathbb{E}[h(Y_c, x)|x]$. Similarly we write $\overline{f}(x) := \mathbb{E}[f(Y_c, x)|x]$, and by the linearity of the conditional expectation we have

$$\overline{f}(x) = \mathbb{E}[f_\theta(Y_c, x)|x] + \mathbb{E}[h(Y_c, x)|x] - \overline{h}(x) = \overline{f}_\theta(x).$$

Thus, we can decompose f into:

$$f(y_c, x) = \overline{f}_\theta(x) + u(y_c, x), \quad \text{where } u(y_c, x) = f_\theta(y_c, x) - \overline{f}_\theta(x) + h(y_c, x) - \overline{h}(x).$$

Similarly, decompose the target function $g(y_c, x)$ as $g(y_c, x) = \overline{g}(x) + v(y_c, x)$, where $v(y_c, x) := g(y_c, x) - \overline{g}(x)$ satisfies $\mathbb{E}[v|x] = 0$. Then

$$\begin{aligned} d(f, g) &= \mathbb{E}_{P_{Y_c, X}}\left[\left((\overline{f}_\theta(X) + u(Y_c, X)) - (\overline{g}(X) + v(Y_c, X))\right)^2\right] \\ &= \mathbb{E}_{P_{Y_c, X}}\left[\left((\overline{f}_\theta(X) - \overline{g}(X)) + (u(Y_c, X) - v(Y_c, X))\right)^2\right]. \end{aligned}$$

Expanding the square, we observe that the cross-term vanishes:

$$\mathbb{E}\left[(\overline{m}_\theta(X) - \overline{g}(X))(u(Y_c, X) - v(Y_c, X))\right] = \mathbb{E}_X\left[(\overline{f}_\theta - \overline{g})\underbrace{\mathbb{E}[u - v|X]}_{=0}\right] = 0.$$

Therefore, the discrepancy separates into two additive terms:

$$d(f, g) = \mathbb{E}_{P_X}\left[(\overline{f}_\theta(X) - \overline{g}(X))^2\right] + \mathbb{E}_{P_{Y_c, X}}\left[(u(Y_c, X) - v(Y_c, X))^2\right].$$

Hence,

$$\inf_{f \in \mathcal{F}_\Theta} d(f, g) = \inf_{\theta \in \Theta} \mathbb{E}_{P_X}\left[(\overline{f}_\theta(X) - \overline{g}(X))^2\right] + \inf_{h \in \mathcal{F}} \mathbb{E}_{P_{X,X}}\left[(u(Y_c, X) - v(Y_c, X))^2\right]$$

$$= \inf_{\theta\in\Theta} \mathbb{E}_{P_X}\left[(\overline{f}_\theta - \overline{g})^2\right] + \inf_{h\in\mathcal{F}} \mathbb{E}\left[\left((f_\theta - \overline{f}_\theta + h - \overline{h}) - (g - \overline{g})\right)^2\right]$$
$$= \inf_{\theta\in\Theta} \mathbb{E}_{P_X}\left[(\overline{f}_\theta - \overline{g})^2\right] + 0$$

since we may choose $h(y_c, x) = g(y_c, x) - f_\theta(y_c, x)$ so that that $\overline{h} = \overline{g} - \overline{f}_\theta$ and thus $\mathbb{E}\left[\left((f_\theta - \overline{f}_\theta + h - \overline{h}) - (g - \overline{g})\right)^2\right] = 0$. Thus,

$$\inf_{f\in\mathcal{F}_\Theta} d(f, g) = \inf_{\theta\in\Theta} \mathbb{E}_{P_X}\left[(\overline{f}_\theta(X) - \overline{g}(X))^2\right] + 0 = \inf_{\theta\in\Theta} \overline{d}(\overline{f}_\theta, \overline{g}). \qquad \square$$

B Structural-Form Restrictiveness

Alternatively, we may want to focus on the *structural form* as a mapping from endogenous and exogenous covariates to outcomes instead of the reduced form. In addition, there are many partially specified structural models (often identified via the use of IV exogeneity conditions) that do not admit a reduced-form representation. In such cases, we a notion of structural-form restrictiveness under the following assumption.

Assumption 5 (Outcome-Covariate Representation)**.** *Assume that the structural equation model* (2) *admits the following outcome representation*

$$Y_{o,i} = f_{\theta_0}\left(Y_{c,i}, X_i, \epsilon_i\right) \tag{24}$$

for some known mapping f_θ and parameter $\theta_0 \in \Theta$, with $Y_{o,i}$ denoting the outcome variable and $Y_{c.i}$ denoting the endogenous covariates.

Structural-form restrictiveness measures constraints on counterfactual mappings, not on observed moments. The subtlety of the structural-form restrictiveness lies in how we define the class of conditional distributions generated by f_θ, which we now describe. Let $\mathcal{Y}_o$ be the space of *distributions* on the domain of $Y_{o,i}$, let $\overline{\mathcal{X}} := \mathcal{Y}_c \times \mathcal{X}$ be the joint domain of augmented covariates $(Y_{c,i}, X_i)$, and let $\mathcal{F}$ be a given eligible class of mappings that associates with each covariate vector $(y_c, x) \in \overline{\mathcal{X}}$ the conditional distribution $P_{Y_o|y_c,x} \in \mathcal{Y}_o$.

For each structural parameter θ and each admissible distribution of ϵ_i, we define the (counterfactual) conditional distribution of outcome as:

$$P^*_{Y_o|Y_c,X}(f_\theta) := P_{f_\theta(Y_c,X,\tilde{\epsilon})}, \quad \text{with } \tilde{\epsilon} \sim \epsilon \text{ and } \tilde{\epsilon} \perp (Y_c, X) \tag{25}$$

Importantly, in the expression $f_\theta(y_c, X, \tilde{e})$, the error argument $\tilde{e}$ is an *independent* copy of the structural error ϵ and independent of all the covariates (Y_c, X).

The endogenous covariate Y_c is (counterfactually) held fixed at value y_c as a constant, and the randomness of X is conditioned upon (which is nevertheless irrelevant for the distribution of ϵ due to assumed independence between X and ϵ). Crucially, notice that

$$f_\theta(y_c, X, \epsilon)|X = x \quad \nsim \quad f_\theta(Y_c, X, \epsilon)|(Y_c, X) = (y_c, x),$$

since the conditional distribution of ϵ given $Y_c = c$ is generally different from the (unconditional) distribution of ϵ, precisely due to the endogeneity of Y_c. Economists are interested in the structural form precisely because of the need for counterfactual analysis as encoded in the definition of $P^*_{Y_o|y_c,X}(f_\theta)$.

Hence, the class of structural form mappings associated with (2) is given by

$$\mathcal{F}_{\Theta,SF} := \left\{P^*_{Y_o|Y_c,X}(f_\theta) : \theta \in \Theta\right\} \subseteq \mathcal{F}.$$

Given a primitive discrepancy function d on (conditional) distributions, such as KL-divergence or Wasserstein distance, we may define the structural-form discrepancy function d_{SF} induced by d:

$$d_{SF}\left(f_\theta, g\right) := \left\{d\left(P^*_{Y_o|Y_c,X}\left(f_\theta\right), g\right)\right\}, \quad \forall g \in \mathcal{F}, \tag{26}$$

We then define the *structural-form restrictiveness* r_{SF} as follows.

Definition 5 (Structural-Form Restrictiveness)**.** Under Assumption 5, we define the structural-form restrictiveness of model (2) as

$$r_{SF} := r(\mathcal{F}_{\Theta,SF}; \mathcal{F}, d_{SF}),$$

based on the discrepancy function d_{SF} according to Definition 1.

The structural-form restrictiveness r_{SF} captures how tightly the model specification (2) constrains the space of mappings from the space of endogenous and exogenous covariates to the outcome space, independently of the form of endogeneity between the structural error ϵ and the endogenous covariate Y_c and how the endogeneity issue is dealt with to achieve econometric identification. The endogeneity issue "disap-

pears" when we replace the structural error ϵ with a completely i.i.d. copy of it $\tilde{\epsilon}$. However, it should be pointed out that this should be by no means interpreted as a "solution" of the endogeneity issue; rather, the construction with an independent $\tilde{\epsilon}$ is intended to provide the structural-form restrictiveness with an interpretation as a counterfactual relationship that economists often hope to obtain in structural models.

Remark 4 (Simplification under Additive Errors)**.** *When the reduced-form model* (4) *has an additive-error structure of the form* (6)*, we can again simplify the definition of the structural-form restrictiveness r_{SF} in a similar manner as in Remark 1. Specifically, we may take:*

- *$\mathcal{Y}$ to be the support of the outcome variable $Y_{o,i}$.*
- *$\mathcal{F}$ to a given eligible set of mappings from the support of (Y_c, X) to $\mathcal{Y}$.*
- *d to be the mean squared distance, i.e., $L_{2,(Y_c,X)}$ distance.*

Example 1 (Demand and Supply: Continued)**.** To illustrate the structural form restrictiveness in this example, focus on the demand equation

$$Q_i = \alpha_1 + \beta_1 P_i + \gamma_1 X_{i1} + \epsilon_{i1}, \quad \mathbb{E}[\,\epsilon_{i1}|\,X_i] = 0,$$

so that the quantity Q_i is the outcome variable $Y_{o,i}$, the price P_i is the endogenous covariate $Y_{c,i}$, while $X_i = (X_{i1}, X_{i2})$ are the exogenous demand and supply shifters. Then the structural-form discrepancy function on the demand function is given by

$$d_{SF}(\mathcal{F}_\Theta, f) = \inf_{\alpha_1,\beta_1,\gamma_1} \mathbb{E}\left[(\alpha_1 + \beta_1 P_i + \gamma_1 X_{i1} - f(P_i, X_i))^2\right]$$

and the structural-form restrictiveness is given by

$$r_{SF} = \frac{\mathbb{E}_{f\sim\lambda_\mathcal{F}}\left[\inf_{\alpha_1,\beta_1,\gamma_1} \mathbb{E}\left[(\alpha_1 + \beta_1 P_i + \gamma_1 X_{i1} - f(P_i, X_i))^2\right]\right]}{\mathbb{E}_{f\sim\lambda_\mathcal{F}}\left[\mathrm{Var}(f(P_i, X_i))\right]}$$

One may similarly define the discrepancy function d_{SF} and the structural-form restrictiveness r_{SF} for the supply equation as well, or for the demand and supply equations together.

Notice that structural form restrictiveness coincides with the restrictiveness of the

following demand model with a "misspecified" exogeneity condition:

$$Q_i = \alpha_1 + \beta_1 P_i + \gamma_1 X_{i1} + \epsilon_{i1}, \quad \mathbb{E}[\,\epsilon_{i1} |\, P_i, X_i] = 0.$$

which imposes a linear additive structure on the demand equation, together with the exclusion restriction that the supply shock X_{i2} does not enter into the structural demand equation. This is clearly different, both in terms of mathematical definition and economic interpretations, from the reduced-form restrictiveness defined earlier for this example.

Reduced-Form vs. Structural-Form Restrictiveness

Formally, the structural primitives are mapped into the reduced form via an equilibrium operator

$$T : \mathcal{S}_{\text{SF}} \longrightarrow \mathcal{S}_{RF},$$

where $\mathcal{S}_{\text{SF}}$ is a space of admissible structural equations, and T assigns to each structural specification its implied reduced form (4), viewed as an element in $\mathcal{S}_{RF}$.

The reduced-form restrictiveness r_{RF} of a model is therefore a property of the *image* of a structural form model (2) parametrized by $\theta \in \Theta$ under the equilibrium operator T, while the structural-form restrictiveness r_{SF} is a property of the structural class structural form model (2) itself, before applying the equilibrium operator T. In general, these two need not coincide.

More generally, one can view r_{SF} and r_{RF} as complementary diagnostics. Structural-form restrictiveness r_{SF} answers the question: *How strongly does the economic structure constrain the mapping from endogenous variables and shifters to outcomes (e.g. demand curves)?* Reduced-form restrictiveness r_{RF} answers the question: *How strongly does the combination of structure, equilibrium, and error assumptions constrain the observable mapping from instruments to equilibrium outcomes?* In applications, it may be informative to report both, to separate the contributions of structural assumptions from those of equilibrium and reduced-form structure.

C Inference with Estimated Discrepancies

As in Ellis and Neff (2025), we assume that for $\lambda_{\mathcal{F}}$-almost every $f \in F$ the discrepancy admits a moment representation $d(f_\theta, f) = \mathbb{E}_{P_X}[g(X, \theta; f)]$, for a measurable function $g : \mathbf{X} \times \Theta \times \mathcal{F} \to \mathbb{R}$, and similarly $d(f_{\text{base}}, f) = \mathbb{E}_{P_X}[g_{\text{base}}(X; f)]$ for some measurable $g_{\text{base}} : \mathbf{X} \times \mathcal{F} \to \mathbb{R}$.

Assumption 6 (Finite Moments)**.** $\sup_{\theta,f} \mathbb{E}\left[(g(\cdot, \theta; f))^{2+\epsilon}\right] < \infty$ *and* $\sup_{\theta,f} \mathbb{E}\left[(g_{\text{base}}(\cdot; f))^{2+\epsilon}\right] < \infty$ *for some* $\epsilon > 0$.

Given S_n we define the empirical analogs

$$d_n(f_\theta, f) := \frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta; f), \quad d_n(f_{\text{base}}, f) := \frac{1}{n}\sum_{i=1}^{n} g_{\text{base}}(X_i; f),$$

and the profiled empirical discrepancy

$$d_n(F_\Theta, f) := \inf_{\theta\in\Theta} d_n(f_\theta, f) = \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta; f).$$

Let $f_1, \ldots, f_M$ be independent draws from $\lambda_{\mathcal{F}}$, independent of S_n. Recall that for each f we denote the population profiled discrepancy by

$$d(F_\Theta, f) := \inf_{\theta\in\Theta} d(f_\theta, f) = \inf_{\theta\in\Theta} \mathbb{E}_{P_X}[g(X, \theta; f)].$$

The Monte Carlo estimator of the numerator and denominator of the restrictiveness index is

$$\widehat{\mu}_{1,n,M} := \frac{1}{M}\sum_{m=1}^{M} d_n(F_\Theta, f_m), \quad \widehat{\mu}_{0,n,M} := \frac{1}{M}\sum_{m=1}^{M} d_n(f_{\text{base}}, f_m),$$

and the plug-in estimator of restrictiveness is $\widehat{r}_{n,M} := \frac{\widehat{\mu}_{1,n,M}}{\widehat{\mu}_{0,n,M}}$. For fixed M this targets the finite-M quantity

$$\mu_{1,M} := \frac{1}{M}\sum_{m=1}^{M} d(F_\Theta, f_m), \quad \mu_{0,M} := \frac{1}{M}\sum_{m=1}^{M} d(f_{\text{base}}, f_m), \quad r_M := \frac{\mu_{1,M}}{\mu_{0,M}},$$

with $r_M \to r(\mathcal{F}_\Theta, \mathcal{F})$ almost surely as $M \to \infty$ by the LLN.

We now seek to establish a central limit theorem for $\widehat{r}_{n,M}$ as $n \to \infty$, for fixed M, that explicitly incorporates sampling uncertainty in $(X_1, \ldots, X_n)$, and to propose a feasible variance estimator.

Assumption 7 (Asymptotic Linearity)**.** *For $\lambda_{\mathcal{F}}$-almost every $f \in \mathcal{F}$ there exists a measurable function $\phi_1(\cdot; f) : \mathbf{X} \to \mathbb{R}$ with $\mathbb{E}_{P_X}[\phi_1(X; f)] = 0$ and $\mathbb{E}_{P_X}[\phi_1(X; f)^2] < \infty$ such that, as $n \to \infty$,*

$$d_n(F_\Theta, f) - d(F_\Theta, f) \;=\; \frac{1}{n}\sum_{i=1}^{n} \phi_1(X_i; f) + o_p(n^{-1/2}). \tag{27}$$

For the baseline discrepancy $d_n(f_{\text{base}}, f)$, an explicit expansion is always available:

$$d_n(f_{\text{base}}, f) - d(f_{\text{base}}, f) = \frac{1}{n}\sum_{i=1}^{n} \Big(g_{\text{base}}(X_i; f) - \mathbb{E}_{P_X}[g_{\text{base}}(X; f)] \Big) := \frac{1}{n}\sum_{i=1}^{n} \phi_0(X_i; f),$$

Given Assumption 7, the estimator $\widehat{r}_{n,M}$ has a simple influence-function representation. For each observation X_i define

$$\Phi_i^{(1)} := \frac{1}{M}\sum_{m=1}^{M} \phi_1(X_i; f_m), \quad \Phi_i^{(0)} := \frac{1}{M}\sum_{m=1}^{M} \phi_0(X_i; f_m),$$

and the combined influence function

$$\psi_M(X_i) := \frac{1}{\mu_{0,M}} \big(\Phi_i^{(1)} - r_M \Phi_i^{(0)} \big). \tag{28}$$

Note that $\mathbb{E}_{P_X}[\psi_M(X_1)] = 0$ and $\mathrm{Var}(\psi_M(X_1)) < \infty$ by Assumption 7.

Theorem 1 (Asymptotic Normality of $\widehat{r}_{n,M}$)**.** *Under Assumption 7. Fix $M \geqslant 1$ and independent draws $f_1, \ldots, f_M \sim \lambda_{\mathcal{F}}$, and define $\mu_{1,M}$, $\mu_{0,M}$ and r_M as above. Then, conditional on $(f_1, \ldots, f_M)$,*

$$\sqrt{n}\big(\widehat{r}_{n,M} - r_M\big) \;\xrightarrow{\;d\;}\; N(0, \sigma_M^2), \quad \sigma_M^2 := \mathit{Var}\big(\psi_M(X_1)\big),$$

as $n \to \infty$, where ψ_M is given by (28)*.*

Proof. By Assumption 7,

$$\begin{aligned}\widehat{\mu}_{1,n,M} &= \frac{1}{M}\sum_{m=1}^{M} d_n(F_\Theta, f_m) \\ &= \frac{1}{M}\sum_{m=1}^{M}\Big(d(F_\Theta, f_m) + \frac{1}{n}\sum_{i=1}^{n}\phi_1(X_i; f_m) + o_p(n^{-1/2})\Big) \\ &= \mu_{1,M} + \frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(1)} + o_p(n^{-1/2}),\end{aligned}$$

and similarly

$$\widehat{\mu}_{0,n,M} = \mu_{0,M} + \frac{1}{n}\sum_{i=1}^{n}\Phi_i^{(0)} + o_p(n^{-1/2}).$$

Therefore

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{1,n,M} - \mu_{1,M} \\ \widehat{\mu}_{0,n,M} - \mu_{0,M}\end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\Phi_i^{(1)} \\ \Phi_i^{(0)}\end{pmatrix} + o_p(1).$$

Conditional on $(f_1, \ldots, f_M)$, the vector $(\Phi_i^{(1)}, \Phi_i^{(0)})$ is i.i.d. in i with finite second moments, so by the multivariate central limit theorem,

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}_{1,n,M} - \mu_{1,M} \\ \widehat{\mu}_{0,n,M} - \mu_{0,M}\end{pmatrix} \xrightarrow{d} N(0, \Sigma_M),$$

where Σ_M is the 2×2 covariance matrix of $(\Phi_1^{(1)}, \Phi_1^{(0)})$.

View $\widehat{r}_{n,M}$ as $\widehat{r}_{n,M} = h(\widehat{\mu}_{1,n,M}, \widehat{\mu}_{0,n,M})$ with $h(u, v) := u/v$. The gradient of h at $(\mu_{1,M}, \mu_{0,M})$ is

$$\nabla h(\mu_{1,M}, \mu_{0,M}) = \Big(\frac{1}{\mu_{0,M}},\ -\frac{\mu_{1,M}}{\mu_{0,M}^2}\Big) = \Big(\frac{1}{\mu_{0,M}},\ -\frac{r_M}{\mu_{0,M}}\Big).$$

By the delta method,

$$\sqrt{n}\big(\widehat{r}_{n,M} - r_M\big) \xrightarrow{d} N(0, \sigma_M^2),$$

with $\sigma_M^2 = \nabla h^\top \Sigma_M \nabla h$. Then, $\nabla h^\top (\Phi_1^{(1)}, \Phi_1^{(0)})^\top = \psi_M(X_1)$, so $\sigma_M^2 = \mathrm{Var}(\psi_M(X_1))$, as claimed. □

We now consider variance estimation. The baseline term has the exact represen-

tation $\phi_0(X; f_m) = g_{\text{base}}(X; f_m) - d(f_{\text{base}}, f_m)$, so a natural plug-in estimator is

$$\widehat{\phi}_0(X_i; f_m) := g_{\text{base}}(X_i; f_m) - d_n(f_{\text{base}}, f_m).$$

For the model term, we use $\widehat{\phi}_1(X_i; f_m) := g\big(X_i, \widehat{\theta}_n(f_m); f_m\big) - d_n(F_\Theta, f_m)$, where $\widehat{\theta}_n(f_m)$ is any approximate minimizer of $d_n(f_\theta, f_m)$ (for instance, the output of the numerical optimization used to compute $d_n(F_\Theta, f_m)$). Under the conditions that justify (27), this plug-in is consistent for ϕ_1 in $L^2(P_X)$.

Define the empirical analogs

$$\widehat{\Phi}_i^{(1)} := \frac{1}{M}\sum_{m=1}^{M} \widehat{\phi}_1(X_i; f_m), \quad \widehat{\Phi}_i^{(0)} := \frac{1}{M}\sum_{m=1}^{M} \widehat{\phi}_0(X_i; f_m),$$

and

$$\widehat{\psi}_i := \frac{1}{\widehat{\mu}_{0,n,M}}\Big(\widehat{\Phi}_i^{(1)} - \widehat{r}_{n,M}\,\widehat{\Phi}_i^{(0)}\Big).$$

The plug-in variance estimator is

$$\widehat{\sigma}_M^2 := \frac{1}{n-1}\sum_{i=1}^{n}\big(\widehat{\psi}_i - \overline{\widehat{\psi}}\big)^2, \quad \overline{\widehat{\psi}} := \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i.$$

Under Assumption 7 and mild additional moment conditions (e.g. $\mathbb{E}[\psi_M(X_1)^4] < \infty$), a standard law-of-large-numbers argument and the continuous mapping theorem yield $\widehat{\sigma}_M^2 \xrightarrow{p} \sigma_M^2$, so that $\frac{\sqrt{n}\left(\widehat{r}_{n,M} - r_M\right)}{\widehat{\sigma}_M} \xrightarrow{d} N(0,1)$. Alternatively, we can use non-parametric bootstrap in X: resampling $X_1, \ldots, X_n$ with replacement while holding $f_1, \ldots, f_M$ fixed.

D Rademacher Complexity and VC Dimension

Let $\mathcal{X}$ be a closed and bounded infinite subset of $\mathbb{R}^m$ for some finite $m \in \mathbb{N}$, let X be a random vector on $\mathcal{X}$ with distribution P_X, and let the outcome space be $\mathcal{Y} = \{-1, 1\}$. Let $\bar{\mathcal{F}}_\sigma = \{f : \mathcal{X} \to \{-1, 1\}\}$ denote the set of all deterministic binary classifiers, and let $\mathcal{F}_\Theta \subseteq \bar{\mathcal{F}}_\sigma$ be a model class. Take the eligible set to be $\mathcal{F} = \bar{\mathcal{F}}_\sigma$ with evaluation distribution $\lambda_{\mathcal{F}}$, and define the discrepancy

$$d_{\text{Rad}}(f, g) := \mathbb{E}_{X \sim P_X}[1 - f(X)g(X)], \quad f, g \in \bar{\mathcal{F}}_\sigma. \tag{29}$$

Since $f(X), g(X) \in \{-1, 1\}$, d_{Rad} is proportional to the misclassification rate under labels $g(X)$.

For this choice of $(\mathcal{F}, d)$, the population approximation error is

$$e(\mathcal{F}_\Theta, \mathcal{F}, d_{\text{Rad}}) = \mathbb{E}_{f \sim \lambda_{\mathcal{F}}} \left[d_{\text{Rad}}(\mathcal{F}_\Theta, f) \right], \quad d_{\text{Rad}}(\mathcal{F}_\Theta, f) = \inf_{h \in \mathcal{F}_\Theta} d_{\text{Rad}}(h, f).$$

Under standard regularity conditions and with a singleton baseline class $\mathcal{F}_{\text{base}} = \{f_{\text{base}}\}$, Ellis and Neff (2025) show that

$$r(\mathcal{F}_\Theta, \mathcal{F}, d_{\text{Rad}}) = 1 - \lim_{n \to \infty} \mathfrak{R}_n(\mathcal{F}_\Theta), \tag{30}$$

where $\mathfrak{R}_n(\mathcal{F}_\Theta)$ denotes the Rademacher complexity of $\mathcal{F}_\Theta$. Since $\mathfrak{R}_n(\mathcal{F}_\Theta) \to 0$ for any class with finite VC dimension, it follows that $r(\mathcal{F}_\Theta, \mathcal{F}, d_{\text{Rad}}) = 1$ for all such model classes. The resulting degeneracy follows from the correlation-based discrepancy d_{Rad}; it does not arise for many alternative discrepancies that measure approximation error directly.

VC dimension is associated with an equivalent discrepancy. For a finite set $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, define $\hat{d}_{\text{VC}}(f, g) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{f(x_i) \neq g(x_i)\}$, with population analog $d_{\text{VC}}(f, g) = \mathbb{E}[\mathbf{1}\{f(X) \neq g(X)\}]$. This discrepancy induces the same notion of restrictiveness studied by Ellis and Neff (2025). The VC dimension of $\mathcal{F}$ is the largest m for which such a finite set exists.

Since $f(X), g(X) \in \{-1, 1\}$, $1 - f(X)g(X) = 2\,\mathbf{1}\{f(X) \neq g(X)\}$, which yields the following equivalence.

Proposition 3. *In the binary classification setting,* $d_{\text{Rad}}(f, g) = 2\, d_{\text{VC}}(f, g)$. *Hence* $r(\mathcal{F}_\Theta, \mathcal{F}, d_{\text{Rad}}) \equiv r(\mathcal{F}_\Theta, \mathcal{F}, d_{\text{VC}})$, *since restrictiveness is invariant to rescaling of the discrepancy.*

Thus, Rademacher complexity and VC dimension correspond to the same underlying discrepancy in this setting, and both lead to degenerate restrictiveness under average-case approximation when the covariate space is infinite.

E Limit of Gaussian-Process Learning Curve

The analysis of the limit of the learning curve in Proposition 2 is related to the literature on learning curves for Gaussian process (GP) regression. To articulate the

connection and the difference, it is useful to separate the roles of approximation, estimation, and irreducible noise. In GP regression, one typically takes squared loss $\ell(y, y') = (y-y')^2$, so $d(f, g)$ is an $L^2(P_X)$ prediction error. The average generalization error of the posterior mean (or BLUP) can then be written in exactly the form of our learning curve $L_n(F_\Theta, A)$, with the average taken over both training samples and draws of the pseudo-truth function f from a GP prior.

Le Gratiet and Garnier (2015) consider this setting with $X_i \sim \mu$ on $\mathbb{R}^d$, a zero-mean GP prior with covariance kernel k, and noisy observations $Z(x_i) + \varepsilon_i$ where the noise variance scales as $\mathrm{Var}(\varepsilon_i) = n\tau$. They study the integrated mean squared error (IMSE) of the BLUP,

$$\mathrm{IMSE}_n = \int \sigma_n^2(x)\, d\mu(x),$$

where $\sigma_n^2(x)$ is the posterior MSE of the predictor for the latent function value $Z(x)$. Their main result shows that, for a broad class of kernels,

$$\mathrm{IMSE}_n \xrightarrow{p} \sum_{p \geqslant 0} \frac{\tau \lambda_p}{\tau + \lambda_p}, \text{ as } n \to \infty. \tag{31}$$

where (λ_p, ϕ_p) are the eigenvalues and eigenfunctions of the covariance operator associated with k and the design measure μ.[9]

In our notation, this corresponds to the following specialization. Let $\mathcal{F}$ be the set of sample paths of the GP, let $\lambda_{\mathcal{F}}$ be the GP prior, let $P_X = \mu$, and take $\ell(y, y') = (y - y')^2$, so that $d(f, g)$ is squared $L^2(\mu)$ distance between functions. Let $\mathcal{A}$ be the BLUP estimator based on noisy observations. Then $L_n(\mathcal{F}_\Theta, \mathcal{A})$ is exactly the GP learning curve (IMSE) when we average over both $f \sim \lambda_{\mathcal{F}}$ and samples. By Proposition 2, in a noise-free ($\tau = 0$) and risk-consistent setting, $\lim_n L_n(\mathcal{F}_\Theta, A)$ must equal $\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}[d(\mathcal{F}_\Theta, f)]$, i.e. the average approximation error of $\mathcal{F}_\Theta$ relative to $(\mathcal{F}, \lambda_{\mathcal{F}}, d)$.

The key difference lies in what is kept in the limit. In Proposition 2 we consider a noise-free design ($Y_i = f(X_i)$) and a risk-consistent estimator, so that the finite-sample estimation error vanishes in the limit and the learning-curve limit contains only the *population approximation error* $d(\mathcal{F}_\Theta, f)$. By contrast, Le Gratiet and Garnier (2015) let the observation noise variance grow proportionally to n, i.e., $\mathrm{Var}(\varepsilon_i) = n\tau$, so as to produce a nontrivial limit. Even when the GP prior is correctly

[9] See Section 3 of Le Gratiet and Garnier (2015) for a general statement and examples.

specified so that the true function lies in the model class (and hence $d(\mathcal{F}_\Theta, f) = 0$ in our notation), the limit is strictly positive: it reflects the uncertainty due to noisy observations (asymptotically balanced with the learning algorithm) rather than any lack of flexibility of the model class.

F Details and Additional Results for Section 5

F.1 Details for Section 5.1

Algorithms to Sample from Constrained GP. To sample a multi-dimensional monotonic increasing function, the starting point is to initially sample a function h from a Gaussian process $\mathcal{GP}\left(0, K\left(x, x'\right)\right)$. Since the drawn sample h is not necessarily monotonic, we apply an algorithm to enforce monotonic increasing properties and obtain a function g that satisfies the required monotonicity. The full procedure is detailed in Algorithms 1 and 2 below. To enforce monotonic increasing behavior, we discretize the domain $\mathcal{D}$ into grid points. When a pair of points $(\mathbf{x}_i, \mathbf{x}_j)$ violates monotonicity-that is, $\mathbf{x}_j > \mathbf{x}_i$ and $h_j < h_i$-we update their values to the average: $h'_j = h'_i = (h_j + h_i)/2$. We then apply linear interpolation between grid points to extend the function over $\mathcal{D}$ while preserving global monotonicity.

Figure A.1 plots five random samples drawn based on the algorithm. We see that they all satisfy the required monotonicity constraints and exhibit non-flat behavior.

Additional Results Figure A.2 illustrates the model fit of $\mathrm{CPT}(\alpha, \delta, \gamma)$ and $\mathrm{DA}(\alpha, \eta)$ for one randomly drawn functional sample. The "Optimal CPT/DA" (red) line shows the closest function within each model class to the pseudo-truth (blue). The discrepancy is measured as the integrated squared distance between the blue and red curves. For this particular draw, CPT delivers a visibly closer approximation than DA.

As a robustness check, we consider two deviations from the baseline Matérn–3/2 GP used to draw the latent function g in Section 5.1. In both cases, we enforce monotonicity exactly as in the baseline.

Squared exponential kernel: we replace the Matérn–3/2 kernel with the squared exponential (RBF) kernel,

$$k_{\mathrm{RBF}}(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right),$$

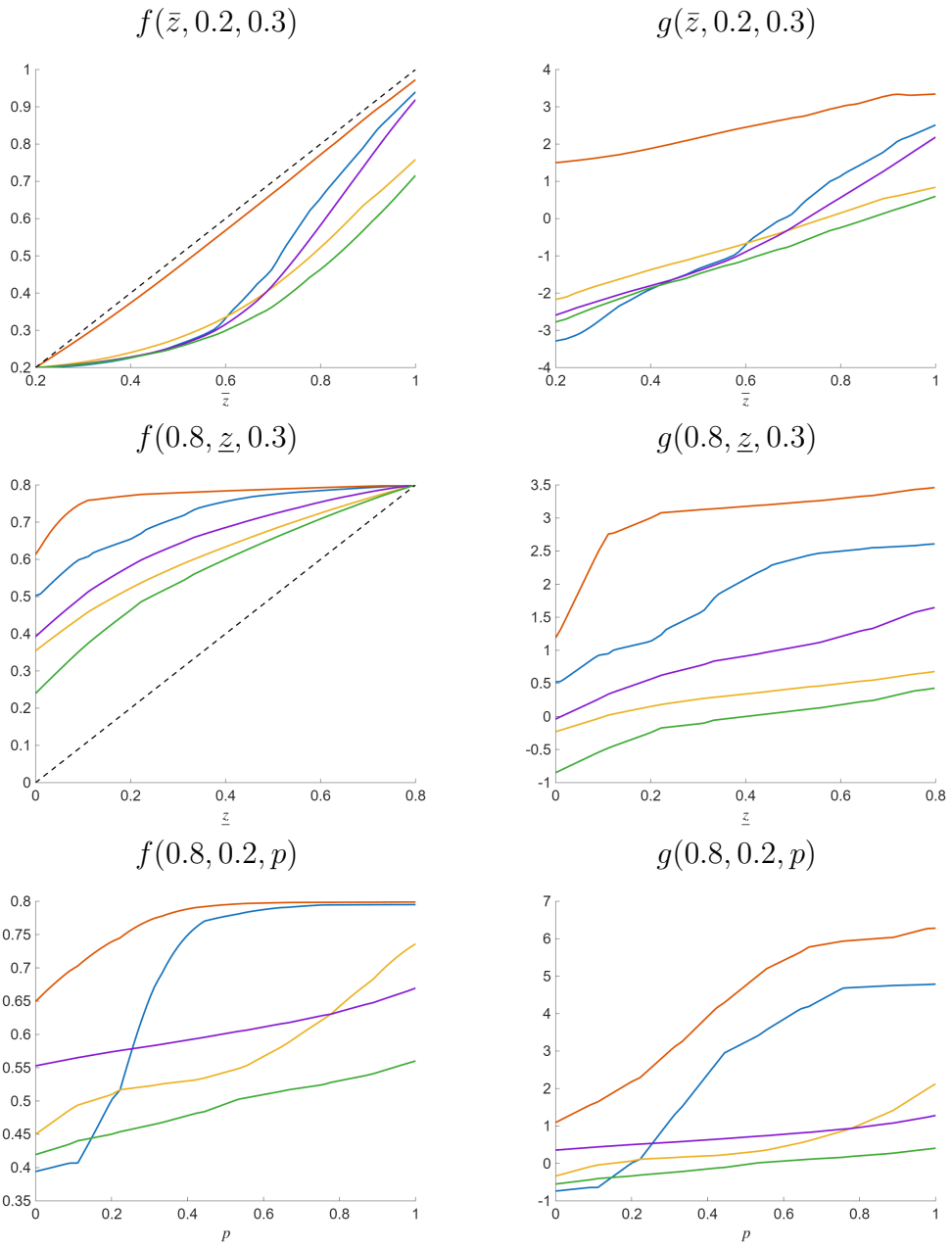
Algorithm 1 Monotonic Function Sampler

Input: Number of samples N, kernel parameters (σ^2, ℓ)
Output: Monotonic increasing function samples $\{g_1, g_2, \ldots, g_N\}$
Generate constrained grid $\mathcal{G} \subset \mathcal{D}$ with n points
Compute covariance matrix $K \in \mathbb{R}^{n \times n}$ using Matérn 3/2 kernel
for $i = 1$ to N **do**
  Sample function values on grid: $\mathbf{h}^{(i)} \sim \mathcal{N}(\mathbf{0}, K)$
  Compute increasing monotonic approximation
  $\mathbf{g}_+^{(i)} \leftarrow \mathrm{MCE}(\mathbf{h}^{(i)}, \mathcal{G})$
  Compute decreasing monotonic approximation
  $\mathbf{g}_-^{(i)} \leftarrow -\mathrm{MCE}(-\mathbf{h}^{(i)}, \mathcal{G})$
  Compute approximation errors
  $e_+^{(i)} \leftarrow \|\mathbf{h}^{(i)} - \mathbf{g}_+^{(i)}\|^2, \quad e_-^{(i)} \leftarrow \|\mathbf{h}^{(i)} - \mathbf{g}_-^{(i)}\|^2$
  Compute adaptive weight
  $w_+^{(i)} \leftarrow e_-^{(i)} / (e_+^{(i)} + e_-^{(i)})$
  Create mixed monotonic increasing function
  $\mathbf{g}_{\text{mix}}^{(i)} \leftarrow w_+^{(i)} \cdot \mathbf{g}_+^{(i)} - (1 - w_+^{(i)}) \cdot \mathbf{g}_-^{(i)}$
  Create linear interpolant
  $g_i \leftarrow \mathrm{Interpolate}(\mathbf{g}_{\text{mix}}^{(i)}, \mathcal{G})$
end for
return $\{g_1, g_2, \ldots, g_N\}$

Algorithm 2 Monotonic Constraint Enforcement (MCE) via Iterative Averaging

Input: Function values $\mathbf{f} \in \mathbb{R}^n$, grid points $\mathcal{G} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
Output: Monotonic function values $\mathbf{f}^*$
$\mathbf{f}^* \leftarrow \mathbf{f}$
repeat
  $\mathbf{f}_{\text{old}} \leftarrow \mathbf{f}^*$
  for all pairs of grid points (i, j) where $i, j \in \{1, \ldots, n\}$ **do**
    if $\mathbf{x}_j \succ \mathbf{x}_i$ and $f_j^* < f_i^*$ **then**
      $f_i^*, f_j^* \leftarrow (f_i^* + f_j^*)/2$
    end if
  end for
until $\mathbf{f}^* = \mathbf{f}_{\text{old}}$ or max iterations reached
return $\mathbf{f}^*$

Figure A.1: Monotonicity Check: Five Samples of f and g

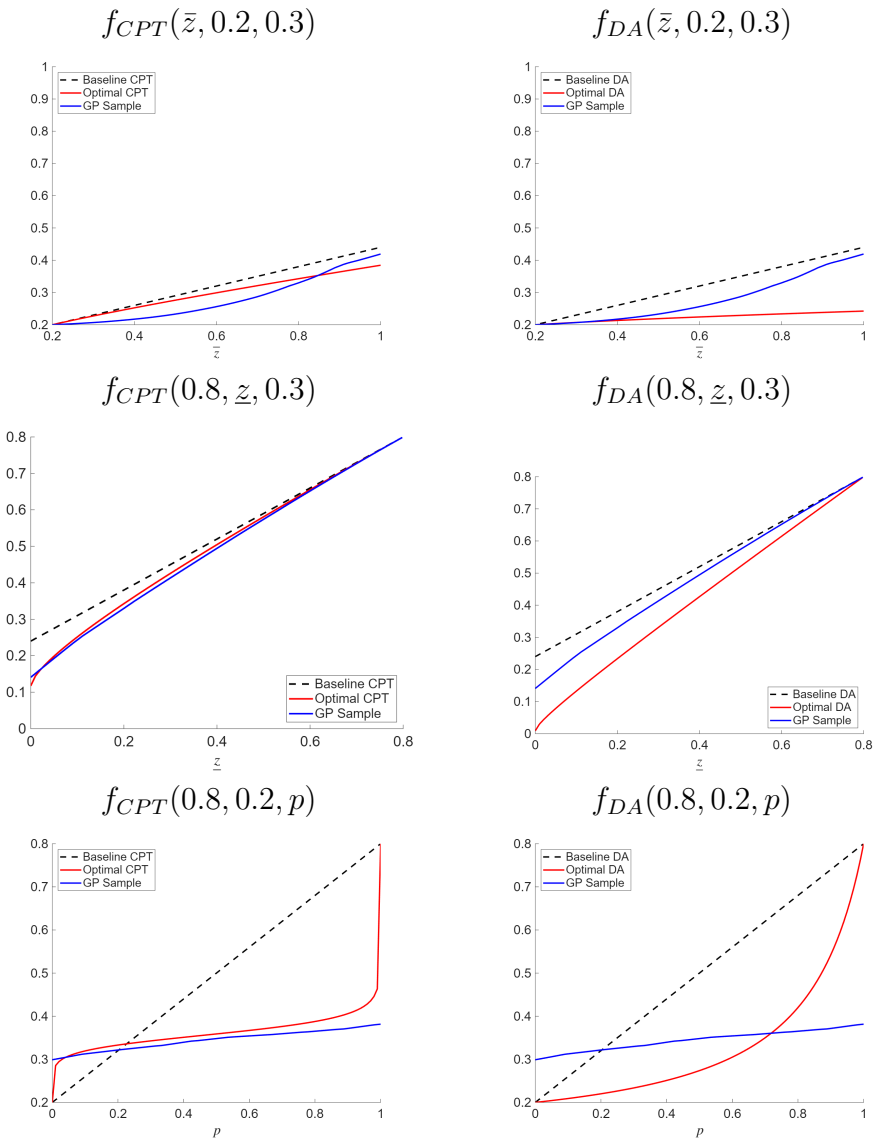


Notes: Five drawn functional samples, f and g plotted against one parameter, holding the other two fixed.

holding (σ^2, ℓ) fixed at their baseline values.

Spline basis: we replace the GP draw for g with an additive cubic spline basis

Figure A.2: CPT(α, δ, γ) is More Flexible Than DA(α, η)



Notes: GP Sample (red solid line) refers to a functional sample f from $\lambda_{\mathcal{F}}$. Baseline CPT/DA (black dashed line) denotes the CPT/DA model f_{base}. Optimal CPT/DA (blue solid line) refers to the function f_θ within the CPT/DA parametric family that achieves the closest fit to the drawn sample. The optimal parameters for the CPT model are $\hat{\alpha} = 0.65$, $\hat{\gamma} = 0.17$, $\hat{\delta} = 0.46$. The optimal parameters for DA model are $\hat{\alpha} = 0.57$, $\hat{\eta} = 4.16$. L^2-norm of $|f_{CPT} - f_m|$ is 0.05, while L^2-norm of $|f_{DA} - f_m|$ is 0.07.

draw,

$$g(x) = B(x)\alpha, \qquad \alpha \sim \mathcal{N}\big(0,\, (I + \lambda D'D)^{-1}\big)\,,$$

Table A.1: Restrictiveness for Certainty Equivalents (Robustness)

	#Param	Baseline	RBF	Spline
CPT Spec.				
α, δ, γ	3	0.56	0.50	0.56
		(0.00)	(0.00)	(0.00)
α, γ	2	0.77	0.75	0.90
		(0.00)	(0.00)	(0.00)
γ, δ	2	0.59	0.52	0.58
		(0.00)	(0.00)	(0.00)
α, δ	2	0.67	0.60	0.60
		(0.01)	(0.00)	(0.00)
α	1	0.92	0.91	1.00
		(0.00)	(0.00)	(0.00)
γ	1	0.86	0.85	0.90
		(0.00)	(0.00)	(0.00)
δ	1	0.69	0.61	0.61
		(0.01)	(0.00)	(0.00)
DA Spec.				
α, η	2	0.67	0.60	0.60
		(0.01)	(0.00)	(0.00)
η	1	0.69	0.61	0.61
		(0.01)	(0.00)	(0.00)
α	1	0.94	0.92	1.00
		(0.00)	(0.00)	(0.00)

where $B(x)$ stacks cubic B-spline bases for each dimension (with $K = 4$ internal knots), and we orthogonalize the non-intercept columns of B for numerical stability. We use a P-spline penalty with $\lambda = 10$, scale the coefficients by 0.6, and then rescale the draw to match the baseline GP variance.

The results are reported in Table A.1. Relative to the baseline result, the resulting restrictiveness estimates are broadly similar and preserve the qualitative ranking across specifications, indicating that our conclusions are not sensitive to the sampling scheme.

F.2 Details for Section 5.2

Model The multinomial choice models we compare are

- Multinomial Logit (MNL): The market share of product j in market m is

$$p_{jm}(X_m; \beta_0) = \frac{\exp\left(x_{jm}^{'}\beta_0\right)}{1 + \sum_{k=1}^{J} \exp\left(x_{km}^{'}\beta_0\right)}$$

- Nested Logit (NL): Products are partitioned into nests indexed by $g \in \mathcal{G}$, and $g(jm)$ denotes the nest containing product j in market m. The market share of product j in market m is

$$p_{jm}(X_m; \beta_0, \rho_0) = p_{j|g(jm),m}\, P_{g(jm),m},$$

  where the within-nest conditional probability is

$$p_{j|g(jm),m} = \frac{\exp\big(x_{jm}'\beta_0/(1-\rho_0)\big)}{\sum_{k\in\mathcal{J}_{g(jm)}(m)} \exp(x_{km}'\beta_0/(1-\rho_0))},$$

  and the nest-level probability is

$$P_{g,m} = \frac{\left[\sum_{k\in\mathcal{J}_g(m)} \exp(x_{km}'\beta_0/(1-\rho_0))\right]^{1-\rho_0}}{1 + \sum_{h\in\mathcal{G}} \left[\sum_{k\in\mathcal{J}_h(m)} \exp(x_{km}'\beta_0/(1-\rho_0))\right]^{1-\rho_0}}.$$

  When $\rho_0 = 0$, the NL model collapses to the standard multinomial logit.

- Mixed Logit (MXL): The market share of product j in market m is

$$p_{jm}(X_m; \beta_0, \Sigma_0) = \mathbb{E}\left[\left.\frac{\exp\left(x_{jm}^{'}\left(\beta_0 + \Sigma_0\nu_i\right)\right)}{1 + \sum_k \exp\left(x_{km}^{'}\left(\beta_0 + \Sigma_0\nu_i\right)\right)}\right| X_m\right],$$

  where $\nu_i \sim N(0, I)$ and $\Sigma_0 = diag(\sigma_1, \ldots, \sigma_d)$.

Eligible Set $\mathcal{F}$ and Evaluation Distribution $\lambda_{\mathcal{F}}$ Our specification of the eligible set is motivated by theoretical work on the flexibility of mixed logit models. The parametric MXL model we evaluate has mean utility component $x_{jm}^{'}\beta_0$ and individual heterogeneity component $x_{jm}^{'}\Sigma\nu_i$. The three alternative eligible sets we consider

differ in which components of utility are allowed to be general functions of product characteristics x_{jm}, possibly subject to monotonicity restrictions.

The first eligible set "NP Both" relaxes both utility components to be general functions. The shares follow

$$s_{jm}(X_m; f) = \mathbb{E}_{f_i}\left[\frac{\exp\left(f(x_{jm}) + f_i(x_{jm})\right)}{1 + \sum_k \exp\left(f(x_{km}) + f_i(x_{km})\right)}\middle| X_m\right] \quad \text{(NP Both)}.$$

Here, $f(.)$ determines the product-level utilities common across individual i, and the f_i determine the individual-specific product-level utilities. We restrict f to be monotonic decreasing in price p_{jm}.

To construct the evaluation distribution $\lambda_{\mathcal{F}}$, individual-specific components $f_i(\cdot)$ are drawn from zero-mean Gaussian processes defined over observed product characteristics $x_{jm} = (p_{jm}, d_{jm})$, where p_{jm} denotes price and d_{jm} is a binary category indicator. The covariance kernel takes a product form

$$K(x, x') = K_p(p, p')\, K_d(d, d'),$$

where K_p is a Matérn kernel with smoothness parameter $\nu = 3/2$,

$$K_p(p, p') = \sigma^2\left(1 + \sqrt{3}\frac{|p - p'|}{\ell}\right)\exp\left(-\sqrt{3}\frac{|p - p'|}{\ell}\right),$$

and $K_d(d, d') = 1\{d = d'\} + \rho\, 1\{d \neq d'\}$. We fix $\sigma^2 = 10$, $\ell = 10$, $\rho = 0.6$. Individual-specific functions $f_i(\cdot)$ are drawn independently across $N_s = 2000$ simulated consumers.

To draw common component $f(\cdot)$ and enforce monotonicity in price, we adopt a derivative-based construction. For each market, we first draw an unconstrained latent Gaussian process $h(\cdot)$ with the same kernel. Products are sorted by price within category $d_{jm} \in \{0, 1\}$, and the latent draw is transformed into a strictly positive derivative magnitude via $\dot{f} = \log(1 + \exp(h))$. The monotonic function is then obtained by cumulative integration over price differences,

$$f(p_{(k)}, d) = -\sum_{r=2}^{k} \dot{f}_{(r),d}\ \left(p_{(r)} - p_{(r-1)}\right),$$

which guarantees that $f(\cdot)$ is weakly decreasing in price. Finally, $f(\cdot)$ is centered to

have zero mean across products in each market.

Given draws of $f(\cdot)$ and $f_i(\cdot)$, market shares are approximated by Monte Carlo integration,

$$\hat{s}_{jm}(X_m; f) = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\exp(f(x_{jm}) + f_i(x_{jm}))}{1 + \sum_k \exp(f(x_{km}) + f_i(x_{km}))}.$$

We also consider two variants, each relaxing one of the utility components to be nonparametric. Specifically,

$$s_{jm}(X_m; f) = \mathbb{E}_{\nu_i} \left[\frac{\exp\left(f(x_{jm}) + x'_{jm}\Sigma\nu_i\right)}{1 + \sum_k \exp\left(f(x_{km}) + x'_{km}\Sigma\nu_i\right)} \middle| X_m \right] \quad \text{(NP mean)},$$

and

$$s_{jm}(X_m; f) = \mathbb{E}_{f_i} \left[\frac{\exp\left(x'_{jm}\beta + f_i(x_{jm})\right)}{1 + \sum_k \exp\left(x'_{km}\beta + f_i(x_{km})\right)} \middle| X_m \right] \quad \text{(NP individual)}.$$

To sample from the evaluation distribution, we use the same Gaussian process priors for f or f_i as in the "NP Both" case, and impose diffuse priors on the parametric components β (NP individual) and Σ (NP mean). Specifically,

$$\beta \sim \mathcal{N}(0, \Omega) \mid \{\beta_{x^1} < 0\}, \quad \Omega = \text{Diag}\left(20^2, 20^2, 20^2\right),$$

where the truncation is imposed only on the coefficient of the price variable, and the remaining coefficients are unrestricted. We assume that Σ is diagonal, with each diagonal element independently distributed as $IG(2, 1)$.

Squared exponential GP kernel As a robustness check, we report restrictiveness results using the squared exponential kernel for Gaussian process draws. As with the Matérn–3/2 kernel, the squared exponential kernel is governed by two parameters,

$$k_{\text{SE}}(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right),$$

which control the marginal variance and length scale, respectively. We use the same parameterization of σ^2 and ℓ as in the baseline specification.

Table A.2 reports restrictiveness under this alternative kernel. Relative to the baseline results in Table 2, restrictiveness decreases for all three models. Importantly,

Table A.2: Multinomial Choice Models (Endogeneity, Squared Exponential Kernel)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.132	0.009
NP Both	NL	0.089	0.005
NP Both	MXL	0.089	0.005
NP Individual	MNL	0.000	0.000
NP Individual	NL	0.000	0.000
NP Individual	MXL	0.001	0.000
NP Mean	MNL	0.133	0.009
NP Mean	NL	0.091	0.005
NP Mean	MXL	0.095	0.005

the qualitative ranking of models remains unchanged: MNL is the most restrictive model, while NL and MNL have similar restrictiveness.

Spline basis As a robustness check, we replace the GP draws for both the common component f and the individual component f_i with cubic spline basis draws. Specifically, in the sampling procedure outlined in F.2, the latent process h is now generated from a spline-basis draw. We construct a truncated-power cubic basis

$$b(x) = \left[1,\ x,\ x^2,\ x^3,\ (x-\kappa_1)_+^3, \ldots, (x-\kappa_K)_+^3\right]',$$

with $K = 4$ internal knots equally spaced over the observed range, and orthogonalize all non-intercept columns. Stacking these basis functions across observations yields a matrix B. We draw the basis coefficients from a P-spline prior, $\alpha \sim \mathcal{N}(0,\ (I + \lambda D'D)^{-1})$ with $\lambda = 10$, and form $h = B\alpha + c$, where the category effect c follows an equicorrelated Gaussian with correlation $\rho = 0.6$. Finally, we rescale h to match the standard deviation of the original GP process.

Table A.3 reports the resulting restrictiveness. Relative to the baseline results, restrictiveness increases for all three models. The qualitative ranking of models is unchanged.

F.3 Details for Section 5.3

Model For each market m and product $j \in \{1, \ldots, J\}$, let $x_{jm} \in \mathbb{R}^K$ denote observed product characteristics and $z_{jm} \in \mathbb{R}^L$ denote excluded instruments. Let $s_{0m} \in (0, 1)$

Table A.3: Multinomial Choice Models (Endogeneity, Spline Basis)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.251	0.009
NP Both	NL	0.211	0.006
NP Both	MXL	0.211	0.007
NP Individual	MNL	0.002	0.000
NP Individual	NL	0.002	0.000
NP Individual	MXL	0.002	0.000
NP Mean	MNL	0.244	0.008
NP Mean	NL	0.205	0.006
NP Mean	MXL	0.206	0.006

denote the outside-option share in market m, which is treated as fixed and observed.

Let $\mathcal{H}$ be a prescribed function space. As in Section 3.3, the systematic utility index is

$$u_{jm} = x'_{jm}\beta + \tilde{h}(x_{jm}, z_{jm}),$$

where

$$\tilde{h}(x_{jm}, z_{jm}) = h(x_{jm}, z_{jm}) - \mathbb{E}[h(x_{jm}, z_{jm})|z_{jm}].$$

In the presence of h, the scale of $\exp(u_{jm})$ can vary substantially across draws, which makes normalizing the outside-option utility undesirable. For this reason, we treat the outside-option share s_{0m} as fixed and compute model-implied shares only for inside goods, normalizing them to sum to $1 - s_{0m}$ in each market.

Given a vector of inside-good utilities $u_m = (u_{1m}, \ldots, u_{Jm})$, define the structural share map for multinomial logit

$$S(u_m; x_m) = (1 - s_{0m}) \left(\frac{\exp(u_{1m})}{\sum_{k=1}^{J} \exp(u_{km})}, \ldots, \frac{\exp(u_{J-1,m})}{\sum_{k=1}^{J} \exp(u_{km})} \right),$$

with analogous definitions for nested logit and mixed logit, where the softmax is taken over inside goods and then rescaled by $1 - s_{0m}$.

The baseline model class restricts the parametric component to $\beta = 0$ while retaining the same control-function flexibility

$$\mathcal{F}_{\text{base}} = \left\{ S\Big(\tilde{h}(x, z); x\Big) : \ h \in \mathcal{H} \right\}.$$

The multinomial choice model classes we compare are:

- Multinomial Logit (MNL): The MNL model class is

$$\mathcal{F}_{\text{MNL}} = \left\{ S\Big(x'\beta + \tilde{h}(x,z); x\Big) : \ \beta \in \mathbb{R}^K, \ h \in \mathcal{H} \right\}.$$

- Nested Logit (NL): The NL model class is

$$\mathcal{F}_{\text{NL}} = \left\{ S_{\text{NL}}\Big(x'\beta + \tilde{h}(x,z); x, \rho\Big) : \ \beta \in \mathbb{R}^K, \ \rho \in [0,1), \ h \in \mathcal{H} \right\},$$

  where S_{NL} denotes the nested-logit share map with fixed outside share.

- Mixed Logit (MXL): Let $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$ denote the random-coefficient scale matrix. The MXL model class is

$$\mathcal{F}_{\text{MXL}} = \left\{ S_{\text{MXL}}\Big(x'\beta + \tilde{h}(x,z); x, \Sigma\Big) : \ \beta \in \mathbb{R}^K, \ \Sigma \in \Theta_\Sigma, \ h \in \mathcal{H} \right\},$$

  where S_{MXL} integrates the inside-good logit shares over $\nu \sim N(0, I)$ and rescales by $1 - s_{0m}$.

Drawing h We draw a *global* function h using Random Fourier Features (RFF). Let

$$\xi_{jm} := \big(\tilde{p}_{jm}, \tilde{z}_{1,jm}, \tilde{z}_{2,jm}\big)$$

collect the arguments entering h. For each draw, we generate

$$h(\xi) = \sqrt{\frac{2}{D}} \sum_{d=1}^{D} a_d \cos(w_d'\xi + b_d),$$

where $b_d \sim \text{Unif}(0, 2\pi)$, $a_d \sim \mathcal{N}(0, \sigma_h^2)$, and the frequency vectors

$$w_d \sim t_3/\ell$$

are drawn independently from a scaled Student-t distribution with three degrees of freedom. This choice of spectral distribution corresponds to a Matérn–3/2 kernel with length scale ℓ, and D controls the accuracy of the approximation. The same realization of h is shared across all markets, ensuring that $\tilde{h}$ represents a common nonparametric component.

Table A.4: Multinomial Choice Models (Endogeneity, 3 IVs)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.789	0.012
NP Both	NL	0.789	0.012
NP Both	MXL	0.682	0.021
NP Individual	MNL	0.670	0.008
NP Individual	NL	0.670	0.008
NP Individual	MXL	0.605	0.014
NP Mean	MNL	0.799	0.010
NP Mean	NL	0.799	0.010
NP Mean	MXL	0.663	0.010

To construct the control-function residual, we project h onto the space of functions of instruments. Let $z_{jm} = (z_{1,jm}, z_{2,jm})$ denote the selected BLP instruments. We compute

$$\bar{h}(z) = \Pi\big(h(\xi) \mid \text{span}\{1, z_1, z_2, z_1^2, z_2^2, z_1 z_2\}\big)\,,$$

where $\Pi(\cdot)$ denotes least-squares projection. The control-function component entering the utility index is then

$$\tilde{h}(x_{jm}, z_{jm}) = h(\xi_{jm}) - \bar{h}(z_{jm}).$$

Three instruments As a robustness check, we report restrictiveness results using three instruments that are most strongly correlated with the endogenous regressor in Table A.4. Relative to the specification with fewer instruments (Table 4), restrictiveness increases across all models. This is consistent with theory, as adding an additional moment condition further constrains the model's ability to fit the pseudo data. Importantly, the qualitative ranking of models remains unchanged: MNL and NL exhibit nearly identical restrictiveness, while MXL remains less restrictive than the other two.

Squared exponential GP kernel As in example 2, we report restrictiveness results using the squared exponential kernel for Gaussian process draws in Table A.5. The restrictiveness of the three models are close to the baseline results.

Table A.5: Multinomial Choice Models (Endogeneity, Squared Exponential Kernel)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.754	0.013
NP Both	NL	0.754	0.013
NP Both	MXL	0.672	0.021
NP Individual	MNL	0.636	0.009
NP Individual	NL	0.636	0.009
NP Individual	MXL	0.588	0.014
NP Mean	MNL	0.730	0.013
NP Mean	NL	0.730	0.013
NP Mean	MXL	0.634	0.010

Spline basis As in example 2, we report restrictiveness results using the spline basis draws in Table A.6. Compared to the baseline results, the restrictiveness of the three models are lower under the "NP Individual" eligible set, while remaining similar under the other two eligible sets. The qualitative ranking of models is unchanged.

Table A.6: Multinomial Choice Models (Endogeneity, Spline)

Eligible Set	**Model**	**Restr.**	**SE**
NP Both	MNL	0.745	0.012
NP Both	NL	0.745	0.012
NP Both	MXL	0.652	0.010
NP Individual	MNL	0.462	0.006
NP Individual	NL	0.462	0.006
NP Individual	MXL	0.442	0.005
NP Mean	MNL	0.746	0.012
NP Mean	NL	0.746	0.012
NP Mean	MXL	0.655	0.011