Refined Deep Causal Inequalities: An Extended Study on Estimators in Differentiated Products Markets

Jian YI[1] , Yichen LIU[2]

Abstract.

The estimation of demand is a fundamental aspect of various economic problems, especially when dealing with extensive and complex data. In this context, two key objectives have emerged:incorporating unstructured data into demand estimation settings, and addressing endogeneity concerns inherent in observational data. Traditional models, such as Berry's (reference), heavily rely on parameterization and the availability of exogenous variations for identification. However, Edvard et al. introduced a novel approach inspired by the partial identification literature, termed the Deep Causal Inequalities (RDeepCI) estimator, which effectively tackles these challenges. Unlike conventional methods, the RDeepCI estimator leverages inferred moment inequalities derived from agents' observed behavior, allowing us to overcome endogeneity issues associated with explanatory variables. The authors provide theoretical guarantees, demonstrating the estimator's consistency under mild conditions. Additionally, they propose an enhanced analytical and statistical estimation rate applicable to diverse hypothesis spaces, showcasing its capability to achieve precise and efficient estimation. Finally, the practical application of the refined estimator is demonstrated in demand estimation scenarios involving both low-dimensional and high-dimensional unstructured data. The code is available Here.
Keywords: Partial Identification, Adversarial Estimation, Non-parameterics, Causal Learning, Demand Estimation

1. Introduction

Supervised machine learning algorithms aim to learn output labels based on a set of explanatory variables. However, there are cases where the explicit output variable is unavail-

able, and researchers must infer the output labels based on their discretion. For example, when consumers choose among multiple products, the researcher may only have access to the final choices and product characteristics. In such scenarios, a common task is to construct a model that predicts the probability of a product being chosen based on its attributes. Conventional approaches use binary choices as output labels and product characteristics as explanatory variables. However, these blackbox methods often overlook important information, such as the characteristics of other products in the choice set, leading to subpar results. (Andrews and Guggenberger, 2009) To address these issues, researchers (McFadden and Train, 2000) have relied on parametric assumptions about the choice-making process. However, these assumptions limit the range of explainable data generating processes and fail to address endogeneity issues with explanatory variables (Berry et al., 1995). Recent studies in partial identification (Andrews and Guggenberger, 2009)(Ciliberto and Tamer, 2009) have shown that utilizing the revealed preferences of agents can yield improved predictions. (Pakes et al., 2015) propose an approach where inequalities are constructed based on the observed behavior of agents to estimate their preference parameters. For instance, in the consumer product choice example, if a consumer chooses a particular product from a given set of options, it implies that the expected utility from that product is higher than that of every other product in the set. However, existing partial identification literature relies on strong parametric assumptions about target functions and scales poorly with the dimensionality of explanatory variables.

Another common issue with observational data is the endogeneity of explanatory variables. Recent studies have demonstrated the limitations of supervised machine learning algorithms when dealing with endogeneity. For instance, the XGBoost method exhibits significant bias and struggles to learn the true underlying function in the presence of substantial endogeneity .

To overcome these challenges, we propose a nonparametric approach called deep causal inequalities. This method utilizes an adversarial estimator to learn the choice function and constructs inequalities from observed consumer choices to address endogeneity issues effectively. Our approach outperforms existing parametric partial identification methods and is applicable to high-dimensional unstructured data.

2. Literature Review

Our work intersects with three streams of literature. Firstly, we contribute to the literature on partial identification and moment inequalities. Partial identification approaches offer more flexibility in applied problems by relaxing restrictive assumptions. They address issues such as multiple equilibria in static games without specifying equilibrium selection rules (Tamer, 2003)(Ciliberto and Tamer, 2009), accommodate flexible specifications of

fixed effects (Ho and Pakes, 2014), handle interval data (Manski and Tamer, 2002), and enhance the robustness of econometric models in a data-driven manner. General methods for set inference with inequality restrictions as moments have been proposed by (Bugni, 2010), (Andrews and Soares, 2010) (Pakes et al., 2015) discuss the application of moment inequalities and the inference of economically interpretable estimators in industrial organization problems.

Secondly, our paper contributes to the emerging literature on causal machine learning. Recent machine learning studies have focused on addressing the issue of endogenous explanatory variables by utilizing available exogenous data (instrumental variables) and solving nonparametric instrumental variable regression problems. Approaches such as linear projections on basis functions (Blundell et al., 2007), nonparametric estimation of conditional distributions (Darolles et al., 2011)(Hall and Horowitz, 2005), and deep generative models(Hartford et al., 2017) have been proposed.(Singh Amandeep and Jiding, 2021) introduce a kernel IV estimator based on conditional mean embedding, and (Muandet et al., 2020) focus on the dual problem and employ a single kernel ridge regression. However, both approaches suffer from the curse of dimensionality.

Lastly, our work is connected to the broader literature on supervised machine learning. In applied work, it has been noted that implicit labeling can exclude valuable information and lead to inferior predictive outcomes. Additionally, if the data includes endogenous explanatory variables, the performance of supervised machine learning algorithms can be further degraded. We demonstrate how our method overcomes these challenges, yielding superior predictive outcomes that can also be interpreted causally.

The rest of the paper is organized as follows: Section 2 provides a brief overview of examples where inequalities can be derived from observed data and how careful construction can mitigate endogeneity issues. Section 3 outlines the formal problem setup and the RDeepCI procedure. The theoretical properties of our RDeepCI estimator are presented in Section 4. The numerical performance of our algorithms is illustrated in Section 5. Finally, Section 6 concludes.

3. Constructing Moment Inequalities: Preference and Discrete Choices

Estimating individuals' preferences holds significant importance across various domains. In marketing, demand analysis plays a crucial role in quantifying consumer responses to policies and forecasting the impact of interventions. This heavily relies on understanding the underlying consumer preference model (Chintagunta and Nair, 2011). Discrete choice models, in particular, have gained popularity due to their ability to capture category choices effectively and offer straightforward interpretations.

The fundamental concept behind discrete choice models is to align observed choice

probabilities with individuals' utility models, accounting for unobserved errors that are unknown to the econometrician. Traditional models of discrete choice heavily rely on assumptions regarding the structure of these unobserved errors. For instance, the **logit model** assumes that the unobserved error follows the type I extreme value distribution, which imposes restrictions on the model.

Products possess various characteristics that individuals can observe but may go unnoticed by researchers. For instance, when analyzing ready-to-eat cereal choices, customers consider factors beyond nutritional information, price, size, and brand, which are typically controlled for explicitly. Additionally, detailed product information may not be accessible to researchers due to confidentiality concerns.

3.1 The Utility Model for Choices

Assume the utility of consumer i choosing product j in market t is

$$u_{ijt} = kX_{ijt}^{(1)} + h^{(k)}(X_{ijt}^{(2)}) + \eta_{jt} + \epsilon_{jt} \tag{1}$$

We model the effect of observable product features (depending on the customer-product pair) $X_{ijt}^{(1)}$ on the utility of purchasing. We divide such features into two channels: $X_{ijt}^{(1)}$ and $X_{ijt}^{(2)}$. $X_{ijt}^{(1)} \in \mathbb{R}$ enters the utility function in a linear way. A typical example is the personalized price of products. $X_{ijt}^{(2)} \in \mathcal{X}$represent the rest of features that contributes to the utility in a potentially nonlinear way through $h^{(k)}$ as a relative measure to k. We drop the superscript (k) in the rest of the paper for notational simplicity.

In the utility model, there are two terms capturing the idiosyncratic errors that are observed by customers when making choices but not by the researcher: η_{jt} and ϵ_{ijt}. η_{jt} could be interpreted as the collection of customer-invariant latent product characteristics, and it is allowed to be correlated with the observable product characteristics. ϵ_{ijt} captures the rest of the randomness, and we only impose rather weak assumptions on its distribution. We defer the discussion on such assumptions to later sections.

3.2 Constructing Moment Inequalities through Revealed Preference

Revealed preference from customer i choosing j over $j^{'}$ in market t implies:

$$\begin{aligned} u_{ijt} \geq u_{ij^{'}t} &\iff kX_{ijt}^{(1)} + h(X_{ijt}^{(2)}) + \eta_{jt} + \epsilon_{ijt} \geq kX_{ij^{'}t}^{(1)} + h(X_{ij^{'}t}^{(2)}) + \eta_{j^{'}t} + \epsilon_{ij^{'}t} \\ &\iff k(X_{ijt}^{(1)} - X_{ij^{'}t}^{(1)}) + h(X_{ijt}^{(2)}) - h(X_{ij^{'}t}^{(2)}) + (\eta_{jt} - \eta_{j^{'}t}) \\ &\quad + (\epsilon_{ijt} - \epsilon_{ij^{'}t}) \geq 0 \end{aligned} \tag{2}$$

Note the inequalities above could not be used to construct moment conditions directly, as η_{jt}'s are unobserved and could be correlated with X_{ijt}'s. To deal with the unknown $\eta_{jt} - \eta_{j't}$, the key idea is to find another potentially different customer $i^{'}$ such that customer $i^{'}$ chooses $j^{'}$ over j in the same market t:

$$k(X^{(1)}_{i'j't} - X^{(1)}_{i'jt}) + (h(X^{(2)}_{i'j't}) - h(X^{(2)}_{i'jt})) + (\eta_{j't} - \eta_{jt}) + (\epsilon_{i'j't} - \epsilon_{i'jt}) \geq 0 \quad (3)$$

Summing up inequalities 2 and 3 effectively differences out the product unobservables ηjt and $\eta j^{'} t$:

$$\begin{aligned} &k(X^{(1)}_{i'j't} - X^{(1)}_{i'jt} + X^{(1)}_{ijt} - X^{(1)}_{ij't}) + (h(X^{(2)}_{i'j't}) - h(X^{(2)}_{i'jt}) + h(X^{(2)}_{ijt}) - h(X^{(2)}_{ij't})) + \\ &(\epsilon_{i'j't} - \epsilon_{i'jt} + \epsilon_{ijt} - \epsilon_{ij't}) \geq 0 \end{aligned} \quad (4)$$

Further taking expectation on both sides and utilize the calculation principle of expectation, we get:

$$\begin{aligned} &k(\mathbb{E}(X^{(1)}_{i'j't} - X^{(1)}_{i'jt} + X^{(1)}_{ijt} - X^{(1)}_{ij't})) + \mathbb{E}(h(X^{(2)}_{i'j't}) - h(X^{(2)}_{i'jt}) + h(X^{(2)}_{ijt}) - \\ &h(X^{(2)}_{ij't}) + \mathbb{E}(\epsilon_{i'j't} - \epsilon_{i'jt} + \epsilon_{ijt} - \epsilon_{ij't}) \geq 0 \end{aligned} \quad (5)$$

we claim that:

$$\mathbb{E}(\epsilon_{i'j't} - \epsilon_{i'jt} + \epsilon_{ijt} - \epsilon_{ij't}) \leq 0, \quad (6)$$

holds, with the intuition that choosing someone else's product increases randomness is higher when individuals choose products not of their own creation. Hence we reduce (4 to our key inequality in the population to construct moments for inference on f :

$$\mathbb{E}(kX^{(1)}_{i'j't} - kX^{(1)}_{i'jt} + kX^{(1)}_{ijt} - kX^{(1)}_{ij't}) + \mathbb{E}(hX^{(2)}_{i'j't} - hX^{(2)}_{i'jt} + hX^{(2)}_{ijt} - hX^{(2)}_{ij't}) \geq 0 \quad (7)$$

We note that, there is no need of further structural assumptions of ϵ_{ijt}, and the assumption of $\mathbb{E}(\epsilon_{i'j't} - \epsilon_{i'jt} + \epsilon_{ijt} - \epsilon_{ij't}) \leq 0$ is rather weak. For instance, $\mathbb{E}(\epsilon_{ijt}) = 0$, for all i, j, t would satisfy this condition, and it is very common to assume the idiosyncratic error has (unconditional) mean zero.

Constructing moment inequalities through revealed preference can be applied to broader settings where the structure of unobserved characteristics are more complicated. For example, in a hospital choice setting, (Ho and Pakes, 2014 allows for qualities of each hospital to differ according to the severity of patients. The inequality method handles such settings by finding pairs to difference out the fixed effects and constructing inequalities as moments inferred from individuals' choices.

3.3 Instruments and Selection of Moments

In traditional settings of using moment inequalities for demand estimation, one major challenge is under-identification. In these settings, we find $\hat{h}$ such that the empirical analog of(7 holds, with h replaced by $\hat{h} \in \mathcal{H}$

$$\begin{aligned}&\frac{1}{|\{i,i',j,j'\}|}\sum_{(i,i',j,j')}(kX^{(1)}_{i'j't}-kX^{(1)}_{i'jt}+kX^{(1)}_{ijt}-kX^{(1)}_{ij't})\\+&\frac{1}{|\{i,i',j,j'\}|}\sum_{(i,i',j,j')}(\hat{h}X^{(2)}_{i'j't}-\hat{h}X^{(2)}_{i'jt}+\hat{h}X^{(2)}_{ijt}-\hat{h}X^{(2)}_{ij't})\geq 0\end{aligned} \tag{8}$$

However, even if we restrict $\mathcal{H}$ to be the class of linear functions(i.e., let $h(X;\theta) = X\theta$) there can be a very large set of parameters that satisfy the moment conditions. To see this, notice the empirical analog

$$\begin{aligned}&\frac{1}{|\{i,i',j,j'\}|}\sum_{(i,i',j,j')}(kX^{(1)}_{i'j't}-kX^{(1)}_{i'jt}+kX^{(1)}_{ijt}-kX^{(1)}_{ij't})\\+&\frac{1}{|\{i,i',j,j'\}|}\sum_{(i,i',j,j')}(X^{(2)}_{i'j't}-X^{(2)}_{i'jt}+X^{(2)}_{ijt}-X^{(2)}_{ij't})\theta\geq 0\end{aligned} \tag{9}$$

defines a hyperplane. Hence, the set of θ's that satisfy the condition above is a half space, which is unbounded and hard to interpret from an empirical standpoint.

Practically, a common solution is to introduce "instruments" to create more moments (i.e., resctrictions). Consider instruments set $\{Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}\}$, such that

$$\mathbb{E}\left[\epsilon_{ijt}-\epsilon_{ij't}+\epsilon_{i'j't}-\epsilon_{i'jt}|(Z_{ijt},Z_{ij't},Z_{i'jt},Z_{i'j't})\right]\leq 0 \tag{10}$$

Such condition is analogous to the standard exclusion restriction $\mathbb{E}(\epsilon|Z) = 0$ for instruments in Generalized Methods of Moments (GMM), but is weaker.

For any $f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}) \geq 0$, we multiply $f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't})$ on both sides of (4 and take the expectation w.r.t. the data generating process, this inequality should still hold:

$$\begin{aligned}&\mathbb{E}\Big[f(Z_{ijt},Z_{ij't},Z_{i'jt},Z_{i'j't})\times[(kX^{(1)}_{i'j't}-kX^{(1)}_{i'jt}\\&+kX^{(1)}_{ijt}-kX^{(1)}_{ij't})+(hX^{(2)}_{i'j't}-hX^{(2)}_{i'jt}+hX^{(2)}_{ijt}-hX^{(2)}_{ij't})]\Big]\\&+\mathbb{E}[f(Z_{ijt},Z_{ij't},Z_{i'jt},Z_{i'j't})(\epsilon_{ijt}-\epsilon_{ij't}+\epsilon_{i'j't}-\epsilon_{i'jt})]\geq 0\end{aligned} \tag{11}$$

Given 11, we have

$$m(h,f) := \mathbb{E}[f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}) \times ((kX^{(1)}_{i'j't} - kX^{(1)}_{i'jt} + kX^{(1)}_{ijt} - kX^{(1)}_{ij't}) + (hX^{(2)}_{i'j't} - hX^{(2)}_{i'jt} + hX^{(2)}_{ijt} - hX^{(2)}_{ij't}))] \geq 0 \tag{12}$$

Since the inequality above holds for any $f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}) \geq 0$, there are essentially infinite number of moments that f should satisfy. In applied research, however, usually specific instruments and structural forms of h functions are chosen for estimation. For example, let $Z_{ijt} = X_{ijt}$, $Z_{ij't} = X_{ij't}$ and $f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}) = (Z_{ijt} - Z_{ij't})_+$ (Ho and Pakes, 2014). Such choices, while reasonable, can be subjective and may fail to incorporate informative bounds.

Our approach uses a minimax objective to reduce the subjectivity of choosing f. For $\mathcal{F} \to \mathbb{R}_+$, we define the identified function f_0 as:

$$f_0 = arg \inf_{h \in \mathcal{H}} \sup_{f \in \mathcal{F}} m(h,f)_- \tag{13}$$

where $x_- = max\{-x, 0\}$ is the negative part to penalize the violation of inequality **??**.

3.4 The RDeepCI Estimator

3.4.1 Prelude

We set $X = (X_{ijt}, X_{ij't}, X_{i'jt}, X_{i'j't})$ and $Z = (Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't})$.Please note that X and Z are not necessarily synonymous, as supply shifters can be considered as a component of Z. The critical radius of a function class $\mathcal{F}$ with range in $[-1, 1]$ is defined as any solution δ_n to the inequality:

$$\mathcal{R}(\delta; \mathcal{F}) \leq \delta^2 \; with : \mathcal{R}(\delta; \mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leq \delta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i)\right] \tag{14}$$

where $\epsilon_{1:n}$ is independent Rademacher random variable drawn in $\{-1, 1\}$ with equal probability. For pairs of individuals and products $(i_j, i_l^{'}, j_l, j_l^{'})_{l=1}^n$ that are constructed from the procedure in Section 3.2(???) and the corresponding product characteristics and instruments, we define the sample moment as

$$m(h,f) := \frac{1}{n} \sum_{i=1}^{n} \Big[f(Z_{ijt}, Z_{ij't}, Z_{i'jt}, Z_{i'j't}) \times \big((kX^{(1)}_{i'j't} - kX^{(1)}_{i'jt} + kX^{(1)}_{ijt} - kX^{(1)}_{ij't}) + (hX^{(2)}_{i'j't} - hX^{(2)}_{i'jt} + hX^{(2)}_{ijt} - hX^{(2)}_{ij't})\big)\Big] \tag{15}$$

which is the empirical analog of (12). Our RDeepCI estimator optimizes the empirical analog of (13), potentially adding norm-based penalties $\Phi : \mathcal{F} \to \mathbb{R}_+$ and $\mathcal{R} : \mathcal{H} \to \mathbb{R}_+$.

3.4.2 Assumptions

Given these assumptions, we can resonably deduce $\forall f \in \mathcal{F}$ is differentiable.
Assumption 1 $\mathcal{F}$ has vanishing Rademacher complexity:

$$E\left[\sup_{f\in\mathcal{F}:\|f\|_2\leq\delta} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right] \to 0$$

Assumption 2 $The\ moment\ function\ m\ is\ Lipschitz$

$$\exists\gamma_j : \forall f_1, f_2 \in \mathcal{F}, |m_j(f_1) - m_j(f_1)| \leq \gamma_j |f_1 - f_2|.$$

Assumption 3 (i)$\mathcal{F}$ is a class of bounded functions, i.e. $sup_{f\in\mathcal{F}}\|f\|_{\mathcal{F}}$ is a bounded random variable.

3.4.3 Formal definition:

The RDeepCI estimator is defined as:

$$\hat{h} := \underset{h\in\mathcal{H}}{arg\,inf}\ \underset{f\in\mathcal{F}}{sup}\, m(h,f)_- - \lambda_n\Phi(f) + \mu_n R(h) \tag{16}$$

4. Simulations

In this section, we evaluate and present the performance of our deep causal inequalities method. We conduct simulations of consumer choices based on the utility model 1, where k is normalized to one

$$u_{ijt} = X_{ijt}^{(1)} + f(X_{ijt}^{(2)}) + \eta_{jt} + \epsilon_{ijt}. \tag{17}$$

Subsequently, we employ the RDeepCI method to estimate the latent utility functions of consumers, which were previously unknown. Consequently, we are able to predict the likelihood of an item being chosen based on its observable attributes. This approach allows

us to gain practical insights into the impact of price discounts or markups on consumer demand.

In this example, we describe how demand estimation can be carried out in markets with differentiated products with highly unstructured data like images. To conduct this simulation, we use the MNIST dataset (LeCun et al., 1998). It is a labelled dataset of 60,000 small gray-scale images with hand-written digits within them, where the size of each image is 28×28. To test the efficiency of our method, we consider the following simulation design:

We assume the utility a consumer i gets by purchasing a good j in market t is given by

$$u_{ijt} = X_{ijt}^{(1)} + \phi(Image_{ijt}) + \eta_{jt} + \epsilon_{ijt} \tag{18}$$

For each product, consumers view some attributes and the product's image. In our simulation, there is a one-dimensional product feature ($X_{ijt}^{(1)}$), e.g., price. The product image, on the other hand, is randomly drawn from the MNIST dataset.

5. Experiments

Our experimental procedures align with the defined Utility functions. To execute these experiments, we apply the RDeepCI estimator in conjunction with the PPHI estimator, as outlined in (Pakes et al., 2015), utilizing a linear specification for the moment inequalities. Additionally, the PolyPPHI estimator incorporates a polynomial function of degree 2 as the moment function. In order to compare the performance of the RDeepCI estimator with the PPHI and PolyPPHI estimators, we evaluate various metrics including RMSE, MSE, MAE, MAPE, and bias. Across all examined specifications, the RDeepCI estimator consistently exhibits significantly superior performance compared to the other estimators. Moreover, as the number of products in the choice set increases, the estimators demonstrate improved performance.

5.1 Deriving RDeepCI exstimator for simulation

The RDeepCI estimator is trained and derived using a Generative Adversarial Network (GAN) network, as discussed in (Creswell et al., 2018).

The employed framework comprises two neural networks: one network aims to learn image classification by generating output labels, while the other network endeavors to generate deceptive images. Given the simplicity of the data, both the learner and adversarial networks are constructed using Multilayer Perceptrons, incorporating activation and softmax layers.

Figure 1: Sturcture for learner net. The learner will receive input in 1x784 shape, and output it's classification. For MNIST dataset, the output is 0-9, along with a fake indicator, indicating the image is not representing any number.

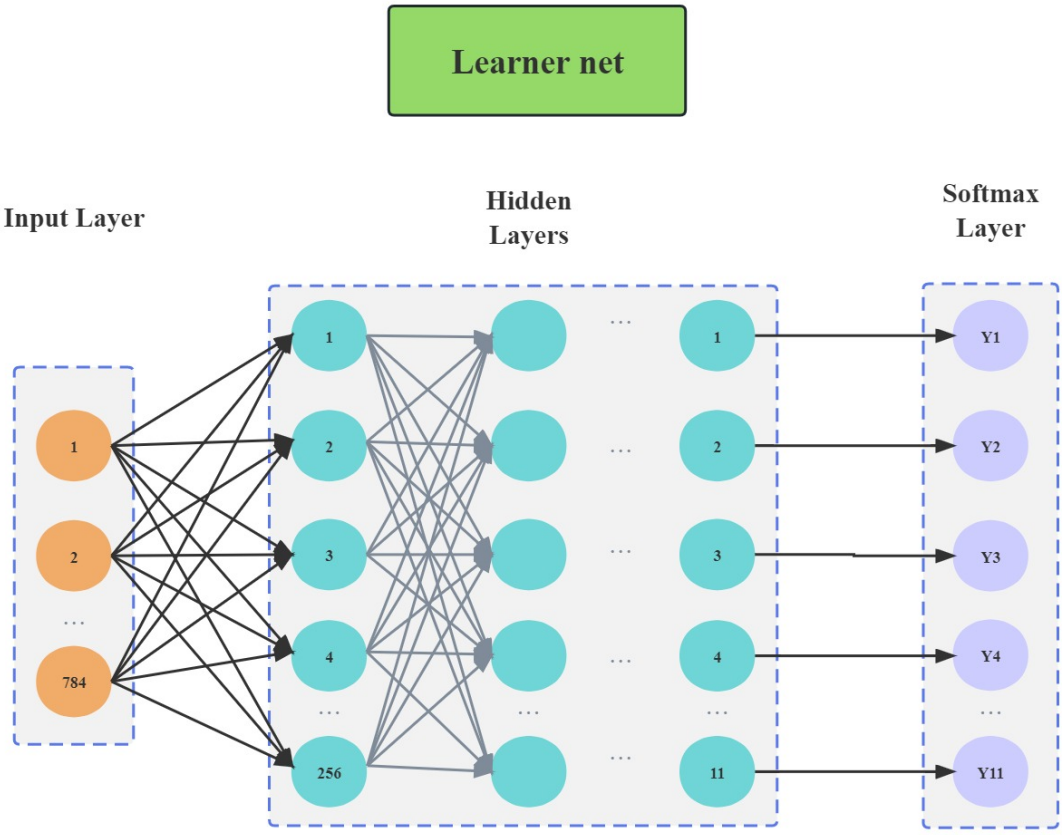


Figure 2: Structure for adversary net. Given a random noise, It will try to generate a fake image.

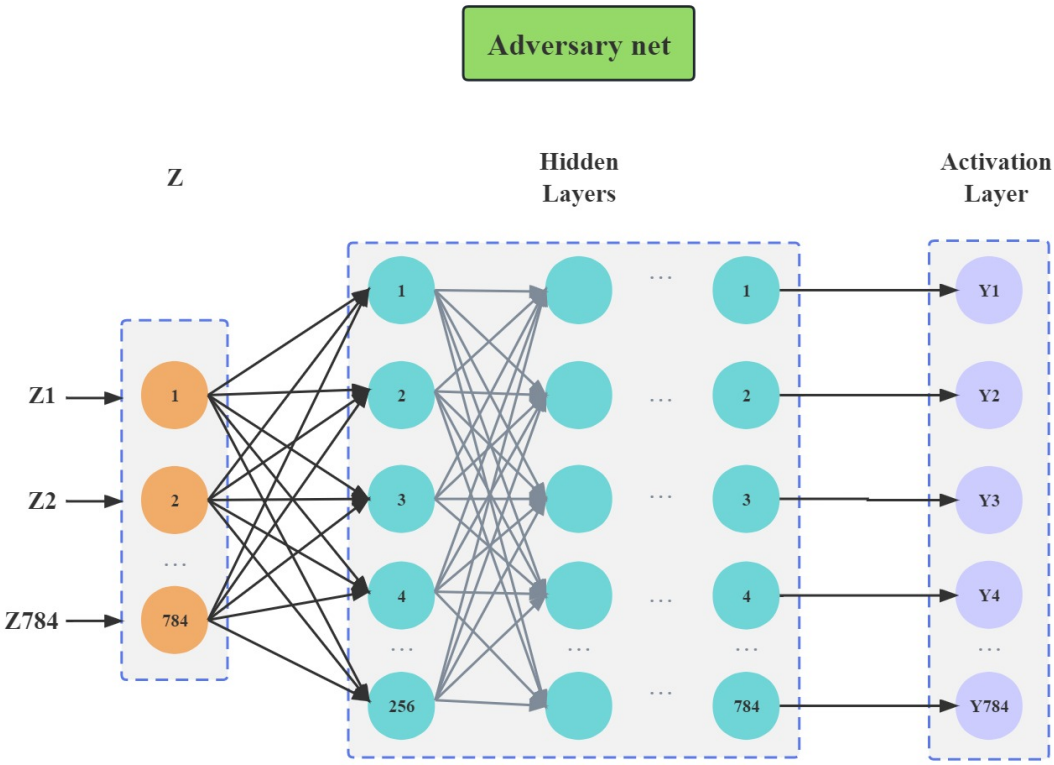
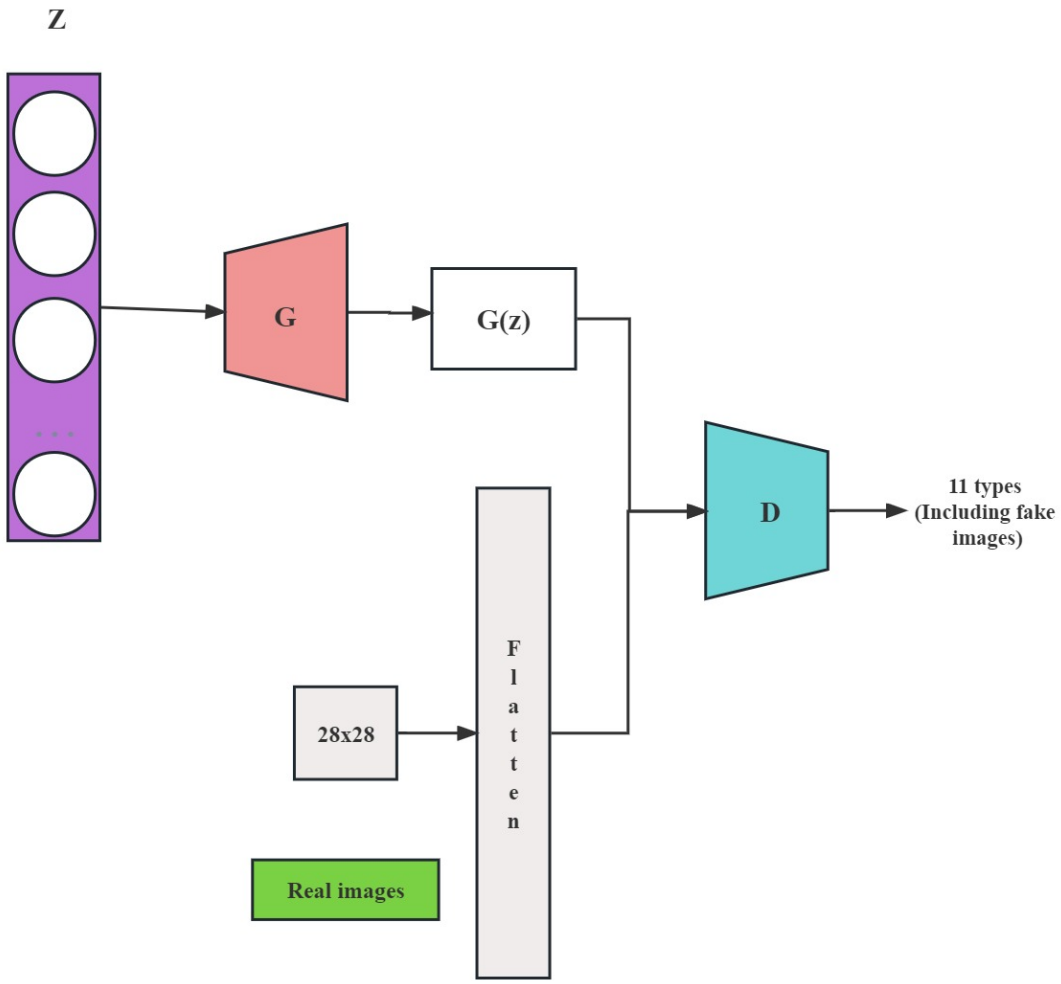


Figure 3: Structure for our training process. As principles of GAN, the gain of Adversary net is just the loss of the learner. And the learner will also be enhanced during training. The G refers to the adversary net (2),and the D refers to the learner net (1).



We just use this GAN to show the efficiency of our loss function following below manner:
as the definition introduced on 15:
Note that the calculation of m involves the discrimination of learner net and fake image generation by adversary net. After we calculate it, we can already do back propagation to

one particular net in an asynchronously way. Combining the simulated data and truth data, the GAN will train the learner with the assistence of the true data, where simulated data (derived in random process) served as a renforcement and supplement.
The overall training process can be illustrated using following algorithm:

Algorithm 1 Asynchronous progressive training

Require: f is in *Lipschitz* continuous
Ensure: m follow the definition in 15.
while $N \leq number\ of\ training\ iterations$ **do**
if $N \mod train\_learner\_cycle = 0$ **then**
$\nabla_{\theta_G} \max(0, (\alpha \sum_{i=1}^{n} (\max(0, -m)^2 - \beta \sum_{i=1}^{n} (k(X_{ijt}^{(1)} - X_{ij't}^{(1)}) + h(X_{i'jt}^{(2)} - X_{i'j't}^{(2)})))$
end if
if $N \mod train\_adversry\_cycle = 0$ **then**
$\nabla_{\theta_D} 2 \sum_{i=1}^{n} (max, -m)$
end if
$N \leftarrow N + 1$
end while

Where α and β are hyperparameters of the model, and $train\_learner\_cycle$ and $train\_learner\_cycle$ control the frequency when the learner and the adversary net are iterated.

5.2 Single-Variate Utility Functions

For the first part of experiments, We perform a simple simulation, where $X_{ijt}^{(2)} \in \mathbf{R}$ and f is a single-variate function.
Assumption: Formally, we assume that:

1. There are $T = 1$ geographic markets.
2. Each market has J products.
3. Customers choose one products among J products, which gives them the highest utility.

For $t = 1, 2, ..., T$and $j = 1, 2, ..., J$, we simulate $X_{ijt}^{(1)} \sim U(-1, 1)$, $X_{ijt}^{(2)} = \rho e_{jt} + U(-5, 5)$ where $e_{jt} \sim N(0, 0.5)$. We also let $\eta_{jt} = e_{jt}$ and $\epsilon_{ijt} \sim N(0, 3)$. Note, the extent of endogeneity is measured by the parameter ρ: it creates a correlation between

the unobserved product characteristic η_{jt} and the observed feature $X_{ijt}^{(2)}$. Further, in the simulation, we focus on the following structural functions of $h(x)$:

- **abs:** $h(x) = 2(||x| - 2| - 1.5)$
- **sin:** $h(x) = 2sin(x)$
- **log:** $h(x) = 2ln(|x|)$
- **step:** $h(x) = 2sgn(||x| - 5| - 2$, where $sgn(x) = 2 \times \mathbb{1}(x \geq 0) - 1$

The summarized results of these estimators can be found in Table A.

5.3 Experiments for Utility Function with Image Data

In the following part of experiments, we assume the utility component from the image is the digit written in that image captured by the function ϕ. In the simulation, we assume that consumer chooses the product that gives them the maximum utility. The task is to recover the function ϕ. To simulate the consumer choice data we assume $X_{ijt}^{(1)} \sim U(-1, 1); \epsilon_{ijt} \sim N(0, 3); \eta_{jt} \sim N(0, 0.5)$. We assume the utility a consumer i gets by purchasing a good j in market t is given by

$$u_{ijt} = X_{ijt}^{(1)} + \phi(Image_{ijt}) + \eta_{jt} + \epsilon_{ijt} \tag{19}$$

For both f and h, we utilized a pre-trained MLP with a simple structure consisting of three fully connected layers. Given that our experimental dataset is MNIST, employing excessively powerful models may not effectively demonstrate the efficacy of our approach.

Our experiments demonstrate that our approach enables the MLP to overcome overfitting on the MNIST dataset, resulting in higher accuracy. Additionally, we observe from different loss functions that the application of RDeepCI leads to reduced loss in the testing phase of the model during data evaluation. The result in B showcase the efficiency of our loss criterion (as introduced in 15) and training process (as introduced in 1).

6. Conclusion

In this study, we build upon the original work by (Singh Amandeep and Jiding, 2021) and present significant improvements in the form of RDeepCI. Our proposed estimator addresses the challenges posed by endogeneity and the inclusion of highly unstructured data. We demonstrate the consistency of our method under mild conditions. To assess its effectiveness, we conduct extensive numerical experiments, consistently showcasing the

superior performance of RDeepCI compared to standard approaches. Furthermore, we apply our method to real-world data, successfully addressing issues related to endogeneity and high-dimensionality.
However, our method does have certain limitations. Specifically, we do not provide any inference results for the estimated functional sets, which we believe presents an avenue for future research. Hence, our method does have certain limitations that warrant consideration. Additionally, our current approach to set construction for functional is more complicated than original one. Furthermore, we implement the potential for a more sophisticated approach by formulating the problem as a bi-level optimization problem. This avenue presents an exciting direction for future research and an opportunity to further extend the impact of our work.

Acknowledgement.

We would like to express our sincere gratitude to the authors of DeepCI (Singh Amandeep and Jiding, 2021) for their invaluable contribution and pioneering efforts in developing the initial framework. Their work laid the foundation upon which our project was built. We are truly indebted to their insights and ideas. Our team invested significant time and effort into extending and modifying the inequalities, ultimately leading to a successful implementation. We are also thankful to our collaborators for their support and valuable input throughout the project. Lastly, we would like to acknowledge the reviewers for their constructive feedback, which greatly improved the quality of our work.

References

Andrews, D. W., & Guggenberger, P. (2009). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory*, *25*(3), 669–709.

Andrews, D. W., & Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, *78*(1), 119–157.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Blundell, R., Chen, X., & Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, *75*(6), 1613–1669.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, *78*(2), 735–753.

Chintagunta, P. K., & Nair, H. S. (2011). Structural workshop paper—discrete-choice models of consumer demand in marketing. *Marketing Science*, *30*(6), 977–996.

Ciliberto, F., & Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, *77*(6), 1791–1828.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, *35*(1), 53–65.

Darolles, S., Fan, Y., Florens, J.-P., & Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, *79*(5), 1541–1565.

Hall, P., & Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables.

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. *International Conference on Machine Learning*, 1414–1423.

Ho, K., & Pakes, A. (2014). Hospital choices, hospital prices, and financial incentives to physicians. *American Economic Review*, *104*(12), 3841–3884.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Manski, C. F., & Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, *70*(2), 519–546.

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

Muandet, K., Mehrjou, A., Lee, S. K., & Raj, A. (2020). Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, *33*, 2710–2721.

Pakes, A., Porter, J., Ho, K., & Ishii, J. (2015). Moment inequalities and their application. *Econometrica*, *83*(1), 315–334.

Singh Amandeep, B. E., & Jiding, Z. Deep causal inequalities: Demand estimation in differentiated products markets. In: In *Causaluai 2021*. 2021. https://sites.google.com/uw.edu/causaluai2021/accepted-papers

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

A. Low Dimensional

Low-Dimensional Case					
Metric	J	function	PPHI	PolyPPHI	RDeepCI
RMSE	5	abs	1.09	0.85	0.27
RMSE	10	abs	1.10	0.85	0.28
RMSE	5	sin	1.06	0.62	0.45
RMSE	10	sin	1.06	0.62	0.26
RMSE	5	log	2.17	1.25	1.67
RMSE	10	log	2.17	1.25	1.73
RMSE	5	step	1.22	1.08	0.45
RMSE	10	step	1.22	1.08	0.40
MSE	5	abs	0.07	0.07	0.08
MSE	10	abs	0.08	0.08	0.08
MSE	5	sin	0.16	0.16	0.22
MSE	10	sin	0.10	0.10	0.07
MSE	5	log	2.84	2.84	2.81
MSE	10	log	2.91	2.91	2.97
MSE	5	step	0.17	0.17	0.20
MSE	10	step	0.19	0.19	0.16
MAE	5	abs	1.01	0.79	0.22
MAE	10	abs	1.01	0.79	0.23
MAE	5	sin	0.99	0.53	0.35
MAE	10	sin	0.99	0.53	0.21
MAE	5	log	1.96	1.02	1.30
MAE	10	log	1.96	1.02	1.43
MAE	5	step	1.21	1.03	0.35
MAE	10	step	1.21	1.03	0.29
MAPE	5	abs	146.26	117.68	138.39
MAPE	10	abs	146.26	117.68	168.45
MAPE	5	sin	247.39	262.83	992.97
MAPE	10	sin	247.39	262.83	267.78
MAPE	5	log	138.29	99.11	115.47
MAPE	10	log	138.29	99.11	217.15
MAPE	5	step	121.42	102.61	34.53
MAPE	10	step	121.42	102.61	28.70
bias	5	abs	-0.21	-0.07	0.02

bias	10	abs	-0.25	-0.12	-0.05
bias	5	sin	0.33	-0.18	-0.16
bias	10	sin	-0.23	-0.06	0.01
bias	5	log	-0.91	-0.65	-0.57
bias	10	log	-1.04	-0.81	-0.78
bias	5	step	-0.55	-0.40	-0.20
bias	10	step	-0.52	-0.37	-0.16

B. Image Dimensional

High-Data Case			
Metric	function	Pretained Net	RDeepCI
Cross-Entropy	abs	0.7466	0.7456
Cross-Entropy	sin	0.7468	0.7451
Cross-Entropy	log	0.7477	0.7460
Cross-Entropy	step	0.7423	0.7408
Focal	abs	0.4499	0.4488
Focal	sin	0.4502	0.4481
Focal	log	0.4512	0.4492
Focal	step	0.4449	0.4431
Test loss	abs	0.9572	0.9562
Test loss	sin	0.9574	0.9557
Test loss	log	0.9582	0.9566
Test loss	step	0.9529	0.9514
Accuracy	abs	97.85%	98.02%
Accuracy	sin	97.64%	97.94%
Accuracy	log	97.83%	98.10%
Accuracy	step	97.75%	98.08%