Kaji, Tetsuya; Manresa, Elena; Pouliot, Guillaume

Working Paper

An adversarial approach to structural estimation

cemmap working paper, No. CWP39/20

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Kaji, Tetsuya; Manresa, Elena; Pouliot, Guillaume (2020) : An adversarial approach to structural estimation, cemmap working paper, No. CWP39/20, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2020.3920

This Version is available at:
https://hdl.handle.net/10419/241914

cemmap
centre for microdata methods and practice

An adversarial approach to structural estimation

Tetsuya Kaji
Elena Manresa
Guillaume Pouliot

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper
CWP39/20

UKRI Economic and Social Research Council

AN ADVERSARIAL APPROACH TO STRUCTURAL ESTIMATION

Tetsuya Kaji[1], Elena Manresa[2], and Guillaume Pouliot[1]

[1]University of Chicago
[2]New York University

July 15, 2020

Abstract

We propose a new simulation-based estimation method, adversarial estimation, for structural models. The estimator is formulated as the solution to a minimax problem between a generator (which generates synthetic observations using the structural model) and a discriminator (which classifies if an observation is synthetic). The discriminator maximizes the accuracy of its classification while the generator minimizes it. We show that, with a sufficiently rich discriminator, the adversarial estimator attains parametric efficiency under correct specification and the parametric rate under misspecification. We advocate the use of a neural network as a discriminator that can exploit adaptivity properties and attain fast rates of convergence. We apply our method to the elderly's saving decision model and show that including gender and health profiles in the discriminator uncovers the bequest motive as an important source of saving across the wealth distribution, not only for the rich.

JEL Codes: C13, C45.

Keywords: structural estimation, generative adversarial networks, neural networks, simulated method of moments, indirect inference, efficient estimation.

We thank Mariacristina De Nardi and John Jones for sharing the data and codes for the empirical application and for very helpful discussion. We also thank Isaiah Andrews, Manuel Arellano, Stephane Bonhomme, Aureo De Paula, Costas Meghir, Chris Hansen, Koen Jochmans, Dennis Kristensen, Whitney Newey, Luigi Pistaferri, and Bernard Salanie, as well as numerous participants in conferences and venues for helpful discussion. Elsie Hoffet, Yijun Liu, Ignacio Ciggliutti, and Marcela Barrios provided superb research assistance. We gratefully acknowledge the support of the NSF by means of the Grant SES-1824304 and the Richard N. Rosett Faculty Fellowship and the Liew Family Faculty Fellowship at the University of Chicago Booth School of Business.

1 INTRODUCTION

Structural estimation is a useful tool to quantify economic mechanisms and learn about the effects of policies that are yet to be implemented. Structural models are naturally articulated as parametric models and, as such, may in principle be estimated using maximum likelihood (MLE). However, likelihood functions arising from economic models are sometimes too complex to evaluate or may not exist in closed form. Meanwhile, generating data from structural models is often feasible, even if it can be computationally intensive. This observation has spurred large literature on simulation-based estimation methods.

A prominent example of such methods is the simulated method of moments (SMM). If we have particular features of the data we want to reproduce, SMM is an attractive tool to naturally incorporate them. At the same time, a naive strategy to stack a large number of moments is known to yield poor finite sample properties (Altonji and Segal, 1996). This tradeoff is especially pronounced in models with rich heterogeneity, where the number of moments may grow exponentially with the number of covariates, leading to the curse of dimensionality. While it may be resolved if we can reduce the moments to a handful of informative ones, it is often the case that such a choice is not obvious.

This paper proposes a new simulation-based estimation method, *adversarial estimation*, that can be used regardless of whether we know which features to match. It is inspired by the *generative adversarial networks (GAN)*, a machine learning algorithm developed by Goodfellow et al. (2014) to generate realistic images. We adopt its adversarial framework to estimate the structural parameters that generate realistic economic data. While maintaining the flexibility of SMM, our method is demonstrated to work well under rich heterogeneity.

The generative adversarial estimation framework is a minimax game between two components—a *discriminator* and a *generator*—over classification accuracy:

$$\min_{\{generator\}} \max_{\{discriminator\}} classification\ accuracy.$$

The generator is an algorithm to simulate synthetic data; its objective is to find a data-generating algorithm that confuses the discriminator. The discriminator is a classification algorithm that distinguishes observed data from simulated data; it takes a data point as input and classifies if it comes from observed data or simulated data;

its objective is to maximize the accuracy of its classification.

In original GAN, both the discriminator and generator are given as neural networks (hence the name). In this paper, we take the generator to be (derived from) the structural model that we intend to estimate, and the discriminator to be an arbitrary classification algorithm (while our primary choice is still a neural network). For classification accuracy, we use the cross-entropy loss, following Goodfellow et al. (2014).[1] From a standpoint of econometrics, it can be seen that the generator is minimizing the distance between observed data and simulated data defined by the choice of the discriminator and inputs thereto.

Our method leverages not only GAN but also the growing literature on why neural networks excel. In the context of nonparametric regression, Bauer and Kohler (2019) show that a multilayer neural network circumvents the curse of dimensionality when the target function has a low-dimensional structure. Building on their approximation result, we show that the same holds true for the discriminator when the likelihood ratio has a low-dimensional structure.[2] Moreover, we propose a heuristic way to check low-dimensionality using *autoencoder*, another seminal machine learning algorithm.

Interestingly, our framework can be viewed as a bridge cast between SMM and MLE. When we use logistic regression as the discriminator, the resulting estimator is asymptotically equivalent to optimally weighted SMM (Example 2). When we use the oracle discriminator given by a likelihood ratio, the resulting estimator is equivalent to MLE (Example 3). What is interesting is the middle case, in which the oracle discriminator is not available but a sufficiently rich discriminator capable of approximating it is used (Example 4). We show that, under some conditions, the resulting estimator enjoys the best of both ends: (1) flexibility to choose moments if desired, (2) closed-form likelihood not required, (3) asymptotic efficiency as MLE.

Our theoretical development proceeds as follows. First, we establish the rate of convergence of a general discriminator (Theorem 1). Then, we apply this to the discriminator given by a neural network (Proposition 3). Next, we develop the parametric rate of convergence of the generator under possible global misspecification (Theorem 6). Finally, we deduce parametric efficiency of the generator under correct specification (Corollary 7). To the best of our knowledge, this is the first work to

[1] There are other losses considered in the literature (Nowozin et al., 2016; Arjovsky et al., 2017), which we do not cover.

[2] Low-dimensionality is a feature of some structural models, where a small number of factors drives variation of multiple outcomes (e.g., Cunha et al., 2010).

thoroughly characterize the statistical properties of a GAN-based algorithm. The generality of our theory allows unbounded random variables and discrete parametric models, unlike many neural network papers assuming bounded variables and simulation-based econometrics papers working with smooth models.

Many challenges in theory stem from the "$\log 0$" problem. On the one hand, logarithm is the benefactor that brings efficiency through the connection to the Jensen-Shannon divergence; on the other hand, it is the malefactor that causes troubles for its infamous divergence toward zero. We overcome this by sometimes borrowing insights from the nonparametric maximum likelihood literature while at other times establishing our own new results. A notable new result is Lemma 5, which bounds the Bernstein "norm" of an arbitrary log likelihood ratio by a possibly non-diverging multiple of a Hellinger distance, which substantially improves a previously known result (Ghosal et al., 2000, Lemma 8.7). This may be of independent interest in other areas such as nonparametric maximum likelihood or Bayes.

Along with our similitude with nonparametric maximum likelihood, we choose to measure the size of the discriminator by the bracketing entropy. To that end, we establish in Lemma 2 a bound on the bracketing number of a multilayer neural network with bounded weights and Lipschitz activation functions with respect to an arbitrary premetric, which is new in the literature and may be of independent interest to those who work on the statistics of neural networks.

Also, despite being a bit problem-specific, the proof of the lemma on convergence of the sample Jensen-Shannon divergence (Lemma 4) involves an algebraic decomposition that was not obvious to the authors in the beginning; it may be of interest to those who consider the problem of misspecification in similar divergence measures.

Using the adversarial estimation framework, we revisit investigation of the elderly's saving motive in the complex setting of De Nardi et al. (2010). Understanding different channels of saving motive is vital in the evaluation of social insurance such as Medicaid. Thus, we aim to disentangle three reasons to save: survival risk, medical expense risk, and bequest motive. The structural model is a dynamic one where agents face heterogeneous risk by gender, age, health status, and permanent income, and optimize their spending to maximize utility from consumption and bequeathing. We carry out adversarial estimation in two specifications: (1) the inputs representing similar identifying variation as De Nardi et al. (2010), (2) the inputs augmented with gender and health. We provide suggestive evidence in our data that unexpected

changes in the health status provide valuable variation to identify the bequest motive. We find that our method uncovers the bequest motive as an important source of saving for the elderly at all levels of the wealth distribution, not only for the rich.

This paper speaks to a wide range of topics in the literature. The first strand is the intersection of machine learning and economics. The use of neural networks in econometrics has a long history (Kuan and White, 1994; Chen and Shen, 1998; Chen and White, 1999; Chen, 2007). The advent of deep learning has triggered even further research on neural networks as nonparametric regression (Hartford et al., 2017; Farrell et al., 2019; Bauer and Kohler, 2019; Schmidt-Hieber, 2020). There is also growing literature on machine learning for causal inference (Chernozhukov et al., 2018; Mackey et al., 2018). Lewis and Syrgkanis (2018) and Bennett et al. (2019) use a non-generative adversarial framework to estimate conditional moment models. Athey et al. (2020) use GAN to create a generator of the Lalonde data. A connection between machine learning and indirect inference is explored in Forneron and Ng (2018).

This paper also relates to the literature on simulation-based estimation. SMM and indirect inference have been widely used in structural estimation (Gouriéroux and Monfort, 1997). There is also a strand of the literature on efficient simulation-based estimation. Fermanian and Salanié (2004) and Kristensen and Shin (2012) propose maximization of the likelihood that is nonparametrically estimated with a kernel. In a similar spirit, Nickl and Pötscher (2010) propose minimization of the Hellinger distance between the densities of actual and simulated data, estimated with a spline, when the data is one-dimensional. One of the major differences of this paper is that it bypasses estimation of the density; our nuisance parameter is the likelihood *ratio*, which suffers much less from issues related to the tail or the support. Gallant and Tauchen (1996) propose the generalized method of moments (GMM) using the score of an auxiliary model whose likelihood is available and show that it is efficient when the auxiliary model nests the structural model. This paper differs in not requiring a tractable auxiliary model that approximates the structural model.

Finally, this paper contributes to statistics. As much as statistical characterization of machine learning algorithms is an active area of research, it is also an important problem to characterize the statistical properties when the model is misspecified (Kleijn and van der Vaart, 2006, 2012; Jankowski, 2014). This paper adds to the list of such work by deriving the asymptotic distribution of the adversarial

estimator under global misspecification. As stated earlier, some intermediate results in the paper may be useful in various fields.

The rest is organized as follows. Section 2 defines the adversarial framework. Section 3 develops the asymptotic properties of the adversarial estimator. Section 4 discusses implementation of estimation and inference. Section 5 revisits investigation of the elderly's saving motive by De Nardi et al. (2010). The appendix contains the proofs. The online appendix contains a Monte Carlo exercise on a simplified Roy Model, an addendum on equivalence with SMM, and details on the empirical application.

2 ADVERSARIAL ESTIMATION FRAMEWORK

This section defines the adversarial estimation framework. It accommodates structural models with a finite number of parameters, possibly with covariates.

The estimation problem we consider is one for which likelihood evaluation is not feasible but simulation is. Hence, there are two sets of observations: the actual observations and the synthetic observations. We let $\{X_i\}_{i=1}^n$ represent the actual observations of size n drawn i.i.d. from a measure space $(\mathcal{X}, \mathcal{A}, P_0)$ and $\{X_i^\theta\}_{i=1}^m$ the synthetic observations of size m generated i.i.d. from $(\mathcal{X}, \mathcal{A}, P_\theta)$ where $\{P_\theta : \theta \in \Theta\}$ is a parametric model over $(\mathcal{X}, \mathcal{A})$. If there is $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$, the structural model is said to be *correctly specified*, while we allow for the possibility that this is not the case. Furthermore, we are concerned with the case where $\{X_i^\theta\}_{i=1}^m$ are generated through a set of common latent variables that do not depend on θ, that is, there exists a measure space $(\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, \tilde{P}_0)$ and i.i.d. observations therefrom $\{\tilde{X}_i\}_{i=1}^m$ such that $X_i^\theta = T_\theta(\tilde{X}_i)$ almost surely for a deterministic measurable function $T_\theta : \tilde{\mathcal{X}} \to \mathcal{X}$.[3] This implies that P_θ is a pushforward measure of $\tilde{P}_0$ under T_θ, that is, $P_\theta = \tilde{P}_0 \circ T_\theta^{-1}$.

This setup arises naturally in complex structural models with dynamic optimization, learning, or latent types that renders analytic characterization of the likelihood infeasible. We note that our framework does not cover structural models with a semiparametric component; such extension is left for future work.

Example 1 (Structural model)**.** Let $\{y_i, x_i\}_{i=1}^n$ be i.i.d. with $y_i \in \mathbb{R}^{d_y}$ and $x_i \in \mathbb{R}^{d_x}$. Consider a structural parametric conditional model where individual outcomes are functions of exogenous variables x_i, an error $\varepsilon_i \in \mathbb{R}^{d_\varepsilon}$ with a known distribution

[3]The latent variables are called *common random numbers* (Gouriéroux et al., 2010).

independent of x_i, and a finite-dimensional parameter $\theta \in \Theta \subset \mathbb{R}^K$; that is, $y_i^\theta = f(x_i, \varepsilon_i; \theta)$ for some function f. The object of interest is typically a function of the structural parameter θ such as the effect of a counterfactual policy.

It is often the case that the associated likelihood of a complex structural model is not available in closed form but simulation is feasible; in particular, we have access to an i.i.d. sample $\{(\varepsilon_i, x_i)\}_{i=1}^m$ of size m, where in conditional models $\{x_i\}_{i=1}^m$ is typically sampled from the empirical distribution of $\{x_i\}_{i=1}^n$, and for any value of θ we can map it into $\{(y_i^\theta, x_i)\}_{i=1}^m$.

Let $X_i = G(y_i, x_i) \in \mathbb{R}^d$ be a set of d functions of (y_i, x_i) representing *the features of the data the researcher chooses to use in estimation.* Some examples of X_i are a subvector of (y_i, x_i), transformations (like logarithms, growth rates, or interactions), or simply the full vector (y_i, x_i). The simulated counterpart, $X_i^\theta = G(y_i^\theta, x_i)$, is the same transformation now as a function of y_i^θ and x_i. □

If we choose θ such that P_θ is very different from P_0, it would be easy to distinguish X_i^θ from X_i. Conversely, if P_θ is close to P_0, distinction would be harder. The idea behind our method is to pick a classification algorithm, possibly state-of-the-art machine learning, and search for the value of θ for which the algorithm can classify the least.

Classification is defined formally as a function $D : \mathcal{X} \to [0, 1]$ such that for given X, $D(X)$ represents the likeliness of X being an actual observation in the scale of a unity; $D(X) = 1$ means that X is classified as "actual" with certainty; $D(X) = 0$ that X is classified as "synthetic". Let $\mathcal{D}$ be the class of classification functions realizable in the algorithm, e.g., the class of appropriate neural networks.

The *adversarial estimator* is defined by the following minimax problem:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \max_{D \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \log D(X_i) + \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(X_i^\theta)). \tag{1}$$

Since D is a function between 0 and 1, both $\log D$ and $\log(1 - D)$ are nonpositive. If X_i and X_i^θ are very different from each other (which is the case when P_θ is far from P_0), the discriminator may be able to find D that assigns 1 on the support of X_i and 0 on the support of X_i^θ, in which case the inner maximization attains the value of zero. Meanwhile, however close X_i^θ is to X_i, the discriminator can at least pick $D \equiv 1/2$,[4] in which case the maximized value is at least $2\log(1/2)$. In general, therefore, the

[4]This is of course provided that a constant function 1/2 is in $\mathcal{D}$, which is usually the case.

inner maximization will give a number between $2\log(1/2)$ and 0, and the closer it is to $2\log(1/2)$, the less able the discriminator is to classify the observations.

Note that the population counterpart of the problem is

$$\min_{\theta\in\Theta}\max_{D\in\mathcal{D}} \mathbb{E}_{X_i\sim P_0}[\log D(X_i)] + \mathbb{E}_{X_i^\theta\sim P_\theta}[\log(1-D(X_i^\theta))].$$

If we do not have a restriction on $\mathcal{D}$ (so that any function $D:\mathcal{X}\to[0,1]$ is allowed), the optimum classification function for the inner maximization is known to be

$$D_\theta(X) := \frac{p_0(X)}{p_0(X)+p_\theta(X)},$$

where p_0 and p_θ are the densities of P_0 and P_θ with respect to some common dominating measure (Goodfellow et al., 2014, Proposition 1). Note here that the objective function with this choice of D is equal to the Jensen-Shannon divergence between P_0 and P_θ. If the model is correctly specified, then θ_0 is the unique solution to the outer minimization (Goodfellow et al., 2014, Theorem 1). In turn, when the model is not correctly specified, we set our target parameter—denoted as well by θ_0—to be the pseudo-parameter that minimizes the Jensen-Shannon divergence.[5]

We now look at three examples of $\mathcal{D}$.

Example 2 (Logistic discriminator)**.** Let $\Lambda(t) = (1+e^{-t})^{-1}$ and $\mathcal{D}$ be the class of logistic discriminators $D(X) = \Lambda(\lambda' X)$ for $\lambda\in\mathbb{R}^d$. The objective function can be interpreted as the log-likelihood of a logistic regression model where the actual observation is associated with outcome 1 and the synthetic with 0. Here, we give a rough intution that the adversarial estimator matches moment $\mathbb{E}[X_i]$.

The first-order condition (FOC) of the inner maximization is

$$\frac{1}{n}\sum_{i=1}^n (1-\Lambda(\lambda' X_i))X_i - \frac{1}{m}\sum_{i=1}^m \Lambda(\lambda' X_i^\theta)X_i^\theta = 0.$$

Thus, the discriminator searches for λ that matches the weighted averages of X_i and X_i^θ. If there exists θ_0 for which $\mathbb{E}[X_i] = \mathbb{E}[X_i^{\theta_0}]$, then $\lambda_0(\theta_0) = 0$ would a solution, since then $\mathbb{E}[(1-\Lambda(0))X_i] = \mathbb{E}[X_i]/2 = \mathbb{E}[X_i^{\theta_0}]/2 = \mathbb{E}[\Lambda(0)X_i^{\theta_0}]$. As a matter of fact, by concavity of the objective function with respect to λ, it is the only solution. Recalling then that the unique outer minimum is attained when the inner maximizer

[5] This is analogous to MLE estimating a pseudo-parameter that minimizes the Kullback-Leibler divergence under misspecification (Huber, 1967; White, 1982; Patilea, 2001).

is $D \equiv 1/2$ (i.e., $\lambda = 0$), we find that $\hat{\theta}$ solves $\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{m}\sum_{i=1}^{m} X_i^{\hat{\theta}} + o_p(1)$. Thus, $\hat{\theta}$ matches the means of X_i and $X_i^{\hat{\theta}}$. In Appendix S.3, we prove asymptotic equivalence of this $\hat{\theta}$ and the optimally weighted SMM with moment $\mathbb{E}[X_i]$. □

Example 3 (Oracle discriminator)**.** Let D be the oracle discriminator D_θ. Then, the estimator boils down to the minimizer of the sample Jensen-Shannon divergence

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_0(X_i)}{p_0(X_i) + p_\theta(X_i)} + \frac{1}{m}\sum_{i=1}^{m} \log \frac{p_\theta(X_i^\theta)}{p_0(X_i^\theta) + p_\theta(X_i^\theta)}.$$

Taking the FOC reveals that the minimizer matches the scores of the actual and synthetic observations. In particular, the associated estimator is efficient under correct specification as $n/m \to 0$.[6] □

Example 4 (Nonparametric discriminator and neural network)**.** In general, we do not know the oracle D_θ in closed form, but we may consider a sieve $\mathcal{D}_n$ of classes of functions that expands as the sample size increases (Chen, 2007). If we choose a sieve of neural networks, D can be written in the following form. Denote the hidden-layer activation function by $\sigma : \mathbb{R} \to \mathbb{R}$ and the output activation function by $\Lambda : \mathbb{R} \to \mathbb{R}$. Let L be the number of hidden and output layers. Let $w_{\ell ij}$ be the weight for the ith node in the $(\ell+1)$th layer on the jth node in the ℓth layer; for example, the input to the second node in the first layer is $w_{021}x_1 + \cdots + w_{02U}x_U$, where $X = (x_1, \ldots, x_U)$ is the input to the network. Let $w_{\ell i} = (w_{\ell i1}, \ldots, w_{\ell iU})'$ be the column vector of weights for the ith node in the $(\ell+1)$th layer. Let $w_\ell = (w_{\ell 1}, \ldots, w_{\ell U})$ be the matrix with columns $w_{\ell i}$; note that for $\ell = L$, w_L is just a column vector as there is only one output. Let w be the vector of all parameters. Then, the discriminator is given by[7]

$$D(X; w) = \Lambda(w_L'\sigma(w_{L-1}'\sigma(\cdots w_1'\sigma(w_0'X)))),$$

where $\sigma(v)$ for a vector v is elementwise application. There is enormous literature on why (deep) neural networks do well (Yarotsky, 2017; Bach, 2017; Mhaskar and Poggio, 2020). Among them, we exploit Bauer and Kohler (2019) in Proposition 3. □

[6] Moreover, estimation based on matching scores can have better properties than estimation based on equating the score to 0. In the dynamic fixed effect panel model, Gouriéroux et al. (2010) show that the resulting estimator is unbiased, while MLE suffers from the incidental parameter problem.

[7] If we include a constant input and a constant node (also known as the "bias" term), it is assumed to be already incorporated in X and w.

3 STATISTICAL PROPERTIES

To help simplify exposition, we denote the empirical measure corresponding to X_i by $\mathbb{P}_n$, to X_i^θ by $\mathbb{P}_m^\theta$, and to $\tilde{X}_i$ by $\tilde{\mathbb{P}}_m$; note that we also have $\mathbb{P}_m^\theta = \tilde{\mathbb{P}}_m \circ T_\theta^{-1}$. Let μ be a measure that dominates P_0 and $\{P_\theta\}$ and denote their densities by p_0 and $\{p_\theta\}$. We usually omit $d\mu$, for example, $\int f p_0 = \int f p_0 d\mu = \int f dP_0$. We employ the operator notation for expectation, e.g., $P_0 \log D = \mathbb{E}_{X_i \sim P_0}[\log D(X_i)]$ and $\mathbb{P}_m^\theta \log(1-D) = \frac{1}{m}\sum_{i=1}^m \log(1 - D(X_i^\theta)) = \tilde{\mathbb{P}}_m \log(1-D) \circ T_\theta$. As a shorthand, we denote the population and sample objective functions by

$$M_\theta(D) := P_0 \log D + P_\theta \log(1-D), \qquad \mathbb{M}_{n,m}^\theta(D) := \mathbb{P}_n \log D + \mathbb{P}_m^\theta \log(1-D).$$

The sample inner maximizer given θ is denoted by $\hat{D}_{n,m}^\theta$ and the outer minimizer by $\hat{\theta}_{n,m}$. The distance of discriminators is measured by a Hellinger-type distance

$$d_\theta(D_1, D_2) := \sqrt{h_\theta(D_1, D_2)^2 + h_\theta(1-D_1, 1-D_2)^2}$$

where $h_\theta(D_1, D_2) := \sqrt{(P_0 + P_\theta)(\sqrt{D_1} - \sqrt{D_2})^2}$. The distance of θ is measured by the Hellinger distance on probability distributions, $h(p,q) := \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$. We use the shorthand $h(\theta_1, \theta_2)$ for $h(p_{\theta_1}, p_{\theta_2})$. We also occasionally use the distance[8]

$$\tilde{h}(\theta_1, \theta_2) := \left[\tilde{P}_0 \left(\sqrt{\frac{p_0}{p_{\theta_1}} \circ T_{\theta_1}} - \sqrt{\frac{p_0}{p_{\theta_2}} \circ T_{\theta_2}} \right)^2 \right]^{1/2}.$$

The size of the sieve is measured by the bracketing entropy.

Definition (Bracketing number and bracketing entropy integral)**.** The *ε-bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, d)$ of a set $\mathcal{F}$ with respect to a premetric d is the minimal number of ε-brackets in d needed to cover $\mathcal{F}$.[9] The *δ-bracketing entropy integral* of $\mathcal{F}$ with respect to d is $J_{[]}(\delta, \mathcal{F}, d) := \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, d)} d\varepsilon$.

[8] Note that $h(\theta_1, \theta_2)$ is roughly equal to $[P_0(\sqrt{p_{\theta_1}/p_0} - \sqrt{p_{\theta_2}/p_0})^2]^{1/2}$. Therefore, h and $\tilde{h}$ are the Hellinger distances measured by $X \sim P_0$ and $\tilde{X} \sim \tilde{P}_0$, respectively, so to speak. A similar Hellinger-like distance is considered in Patilea (2001).

[9] A *premetric* on $\mathcal{F}$ is a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that satisfies $d(f,f) = 0$ and $d(f,g) = d(g,f) \geq 0$ for every $f, g \in \mathcal{F}$. It is also called "pseudosemimetric".

3.1 Assumptions

On the Sieve

Let $\mathcal{D}^{\theta}_{n,\delta} := \{D \in \mathcal{D}_n : d_\theta(D, D_\theta) \leq \delta\}$. The following requires that the sieve does not grow too fast.

Assumption 1 (Entropy of sieve)**.** There exists $\alpha < 2$ such that $J_{[]}(\delta, \mathcal{D}^{\theta}_{n,\delta}, d_\theta)/\delta^\alpha$ is decreasing in δ uniformly in $\theta \in \Theta$. There exists $\delta_n = o(n^{-1/4})$ such that $J_{[]}(\delta_n, \mathcal{D}^{\theta}_{n,\delta_n}, d_\theta) \lesssim \delta_n^2\sqrt{n}$ uniformly in $\theta \in \Theta$.

Next is a refinement of the "bounded likelihood ratio" condition used in nonparametric maximum likelihood.[10] It is often trivial if we assume a compact support for X_i, which is standard in the neural network literature.

Assumption 2 (Support compatibility)**.** Let $P(X|A)$ be $P(X\mathbb{1}\{A\})/P(A)$ if $P(A) > 0$ and 0 otherwise. There exist $\delta_n = o(n^{-1/4})$ and M such that uniformly in $\theta \in \Theta$,

$$\sup_{D \in \mathcal{D}^{\theta}_{n,\delta_n}} P_0\left(\frac{D_\theta}{D} \,\middle|\, \frac{D_\theta}{D} \geq \frac{25}{16}\right) < M, \quad \sup_{D \in \mathcal{D}^{\theta}_{n,\delta_n}} P_\theta\left(\frac{1-D_\theta}{1-D} \,\middle|\, \frac{1-D_\theta}{1-D} \geq \frac{25}{16}\right) < M.$$

Also, the brackets $\{\ell \leq D \leq u\}$ in Assumption 1 can be taken so that $(P_0 + P_\theta)(\frac{D_\theta}{\ell}(\sqrt{u} - \sqrt{\ell})^2)$ and $(P_0 + P_\theta)(\frac{1-D_\theta}{1-u}(\sqrt{1-\ell} - \sqrt{1-u})^2)$ are $O(d_\theta(u, \ell)^2)$.

The following is a sufficient condition for a particular family of neural network discriminators to satisfy Assumption 1 when $d^* < 2p$. It also accommodates the low-dimensional structure of Bauer and Kohler (2019).[11]

Assumption 3 (Neural network)**.** Let P_0 and P_θ have subexponential tails and finite first moments uniformly in $\theta \in \Theta$.[12] Let $\log(p_0/p_\theta)$ satisfy the assumptions for m in Bauer and Kohler (2019, Theorem 3) uniformly in $\theta \in \Theta$; in particular, $\log(p_0/p_\theta)$ satisfies a (p, C)-smooth generalized hierarchical interaction model of order d^* and finite level l with K components for $p = q + s$, $q \in \mathbb{N}_0$, and $s \in (0, 1]$; all partial derivatives of order q or less of functions g_k, $f_{j,k}$ of $\log(p_0/p_\theta)$ are bounded; all g_k are Lipschitz with a positive constant. Let $\mathcal{D}_n := \{\Lambda(f) : f \in \mathcal{H}^{(l)}\}$, $\Lambda(x) := 1/(1+e^{-x})$, be a sieve of neural network discriminators that satisfy the assumption of Lemma 2.

[10] E.g., van der Vaart and Wellner (1996, Theorem 3.4.4) and Ghosal et al. (2000, Lemma 8.7).

[11] The low-dimensional structure in Bauer and Kohler (2019) is related to the the target function satisfying a generalized hierarchical interaction representation. See Appendix S.2.4 for the definition.

[12] We say that P on $\mathbb{R}^d$ has *subexponential tails* if $\log P(\|X\|_\infty > a) \lesssim -a$ for large a.

Finally, let $\mathcal{H}^{(l)}$ satisfy the assumptions for the neural network in Bauer and Kohler (2019, Theorem 3); in particular, $\mathcal{H}^{(l)}$ is defined as Bauer and Kohler (2019, (6)) with K, d, d^* as in the structure of $\log(p_0/p_\theta)$; the activation function is q-admissible;

$$M_* = \binom{d^*+q}{d^*}(q+1)\left(\left[\frac{(\log\delta_n)^{2(2q+3)}}{\delta_n}\right]^{\frac{1}{p}}+1\right)^{d^*},\ \alpha = \left[\frac{(\log\delta_n)^{2(2q+3)}}{\delta_n}\right]^{\frac{d^*+p(2q+3)+1}{p}}\frac{\log n}{\delta_n^2}$$

for $\delta_n = [(\log n)^{\frac{p+2d^*(2q+3)}{p}}/n]^{\frac{p}{2p+d^*}}$.

On the Estimation Procedure

The following allows us to establish results at rates in terms of n.

Assumption 4 (Growing synthetic sample size)**.** n/m converges.

The following makes sure that the trained discriminator converges to the true discriminator at a rate fast enough to yield a meaningful estimator for θ_0.

Assumption 5 (Approximately maximizing discriminator)**.** The trained discriminator $\hat{D}^\theta_{n,m} \in \mathcal{D}_n$ satisfies $\mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m}) \geq \mathbb{M}^\theta_{n,m}(D_\theta) - o_P(n^{-1/4})$ uniformly over $\theta \in \Theta$.

The following ensures that the derivative of the sample objective function converges to that of the population. This is a standard assumption in M-estimation that involves nuisance parameters (Klein and Spady, 1993; Gouriéroux and Monfort, 1997; Fermanian and Salanié, 2004; Nickl and Pötscher, 2010) to obtain a regular estimator for θ_0 (Newey, 1994). For this, it is important in practice to fix the structural shocks that generate synthetic data as well as random seeds in any stochastic optimization algorithm involved.

Assumption 6 (Approximately minimizing generator and orthogonality)**.** There exists open $G \subset \Theta$ that contains θ_0 such that the estimator $\hat{\theta}_{n,m}$ satisfies

$$\mathbb{M}^{\hat{\theta}_{n,m}}_{n,m}(\hat{D}^{\hat{\theta}_{n,m}}_{n,m}) \leq \inf_{\theta\in G}\mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m}) + o_P(n^{-1}),$$
$$\inf_{\theta\in G}\left[\mathbb{M}^{\hat{\theta}_{n,m}}_{n,m}(\hat{D}^{\hat{\theta}_{n,m}}_{n,m}) - \mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m})\right] - \left[\mathbb{M}^{\hat{\theta}_{n,m}}_{n,m}(D_{\hat{\theta}_{n,m}}) - \mathbb{M}^\theta_{n,m}(D_\theta)\right] \geq o_P(n^{-1}).$$

On the Structural Model

Assumption 7 (Identification)**.** For every open $G \subset \Theta$ that contains θ_0, we have $\inf_{\theta\notin G} h(\theta,\theta_0) > 0$ and $\inf_{\theta\notin G} M_\theta(D_\theta) > M_{\theta_0}(D_{\theta_0})$.

The following assumes that the entropy of the structural model is low enough to admit a $\sqrt{n}$-estimator of θ_0.

Assumption 8 (Hellinger bracketing of generative model)**.** Let $\mathcal{P}_\delta := \{p_\theta : \theta \in \Theta, h(\theta_0, \theta) \leq \delta\}$ and $\tilde{\mathcal{P}}_\delta := \{(p_0/p_\theta) \circ T_\theta : \theta \in \Theta, \tilde{h}(\theta_0, \theta) \leq \delta\}$. There exists $r < \infty$ such that $N_{[]}(\varepsilon, \mathcal{P}_\delta, h) \lesssim (\delta/\varepsilon)^r$ and $N_{[]}(\varepsilon, \tilde{\mathcal{P}}_\delta, \tilde{h}) \lesssim (\delta/\varepsilon)^r$ for $0 < \varepsilon \leq \delta$. $\tilde{h}(\theta_0, \theta) = O(h(\theta_0, \theta))$ as $\theta \to \theta_0$.

The following assumes a type of twice differentiability that is weaker than the pointwise. Notably, it can be satisfied by densities with jumps and kinks, which appear in censored models, auctions, search models, and corporate finance (Chernozhukov and Hong, 2004; Strebulaev and Whited, 2011). It builds on Le Cam's differentiability in quadratic mean (Pollard, 1997; van der Vaart, 1998, Chapter 7) and adds local uniformity and twice differentiability. Local uniformity is required as our method involves measuring the distance with both actual and synthetic samples. Twice differentiability is needed to accommodate misspecification. The map $\dot{\ell}_{\theta_0}$ is the *score function* for θ_0, and the matrix I_{θ_0} the *Fisher information matrix* for θ_0.

Assumption 9 (Uniform and twice differentiability in quadratic mean)**.** The parameter space Θ is (a subset of) a Euclidean space $\mathbb{R}^k$. The structural model $\{P_\theta : \theta \in \Theta\}$ is *(locally) uniformly differentiable in quadratic mean* at θ_0, that is, there exists a $k \times 1$ vector of measurable functions $\dot{\ell}_{\theta_0} : \mathcal{X} \to \mathbb{R}^k$ such that for $h, g \in \mathbb{R}^k$ and $g \to 0$,

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0+g}} - \frac{1}{2}(h-g)' \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0+g}} \right]^2 = o(\|h-g\|^2).$$

It is also *twice differentiable in quadratic mean* at θ_0, that is, there exists a $k \times k$ matrix of measurable functions $\ddot{\ell}_{\theta_0} : \mathcal{X} \to \mathbb{R}^{k \times k}$ such that for $h \in \mathbb{R}^k$ and $h \to 0$,

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2} h' \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} - \frac{1}{4} h' \ddot{\ell}_{\theta_0} h \sqrt{p_{\theta_0}} - \frac{1}{8} h' \dot{\ell}_{\theta_0} \dot{\ell}'_{\theta_0} h \sqrt{p_{\theta_0}} \right]^2 = o(\|h\|^4)$$

and $I_{\theta_0} := P_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}'_{\theta_0} = -P_{\theta_0} \ddot{\ell}_{\theta_0}$. The matrix $\tilde{I}_{\theta_0} := 2P_{\theta_0}(D_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}'_{\theta_0} + (\ddot{\ell}_{\theta_0} + \dot{\ell}_{\theta_0} \dot{\ell}'_{\theta_0}) \log(1 - D_{\theta_0}))$ is positive definite.

Remark. The matrix $\tilde{I}_{\theta_0}$ is the curvature of the outer minimization.

Remark. Under correct specification, the annoying term $(\mathbb{P}_m^\theta - \mathbb{P}_m^{\theta_0}) \log(1 - D_{\theta_0})$ in Lemma 4 goes away, making twice differentiability unnecessary.

We impose very mild smoothness on the simulated data transformation compared to, e.g., Nickl and Pötscher (2010, Assumptions P1–2, R) or Gouriéroux and Monfort (1997, Chapter 2). Importantly, we do not exclude cases where T_θ is discontinuous. Such situations arise frequently in economics (Frazier et al., 2019) while many existing econometric theories rule them out.[13]

Assumption 10 (Smooth synthetic data generation)**.** For every compact $K \subset \Theta$,

$$\sqrt{\frac{n}{m}} \sup_{h \in K} \left\| \sqrt{m}(\tilde{\mathbb{P}}_m - \tilde{P}_0) D_{\theta_0} (\dot{\ell}_{\theta_0} \circ T_{\theta_0 + h/\sqrt{n}} - \dot{\ell}_{\theta_0} \circ T_{\theta_0}) \right\| = o_P^*(1).$$

For the rate of convergence, we need that P_0 is "close enough" to P_{θ_0} in the sense that the Hellinger convergence of P_θ to P_{θ_0} takes place on the support of P_0.

Assumption 11 (Smooth synthetic model and overlapping support with P_0)**.** There exists open $G \subset \Theta$ containing θ_0 in which $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \gtrsim h(\theta, \theta_0)^2$. For every compact $K \subset \Theta$,

$$\sqrt{\frac{n}{m}} \sup_{h \in K} \left| \sqrt{m} \frac{(\mathbb{P}_m^{\theta_0 + h/\sqrt{n}} - \mathbb{P}_m^{\theta_0}) - (P_{\theta_0 + h/\sqrt{n}} - P_{\theta_0})}{1/\sqrt{n}} \log(1 - D_{\theta_0}) \right| = o_P^*\left(1 + \frac{n}{m}\right).$$

Also, $h(\theta, \theta_0)^2 = O(\int D_{\theta_0} (\sqrt{p_{\theta_0}} - \sqrt{p_\theta})^2)$ as $\theta \to \theta_0$.

Remark. The first condition of Assumption 11 is implied by positive definiteness of $\tilde{I}_{\theta_0}$ in Assumption 9.

The following assumption is required for efficiency.

Assumption 12 (Correct specification)**.** The synthetic model $\{P_\theta : \theta \in \Theta\}$ is correctly specified, that is, $P_{\theta_0} = P_0$ and $D_{\theta_0} \equiv 1/2$.

Remark. Assumption 12 implies Assumption 11.

3.2 Theorems

Theorem 1 (Rate of convergence of discriminator)**.** *Under Assumptions 1, 4, and 5,* $d_\theta(\hat{D}_{n,m}^\theta, D_\theta) = o_P^*(n^{-1/4})$ *uniformly in* $\theta \in \Theta$.

Theorem 2 (Rate of convergence of objective function)**.** *Under Assumptions 1, 2, 4, and 5,* $\mathbb{M}_{n,m}^\theta(\hat{D}_{n,m}^\theta) - \mathbb{M}_{n,m}^\theta(D_\theta) = o_P(n^{-1/2})$ *uniformly in* $\theta \in \Theta$.

[13] For example, limited dependent variable models satisfiy Assumption 10 under Assumption 4.

The following provides the rate of convergence for a particular neural network. The structure of the network and the rate of convergence depend on smoothness and the underlying dimension of the likelihood ratio, not on the dimension of X_i.

Proposition 3 (Rate of convergence of neural network discriminator)**.** *Under Assumptions 3 to 5,* $d_\theta(\hat{D}^\theta_{n,m}, D_\theta) = O^*_P(\delta_n)$.

Consistency can be proved with different, conceptually weaker assumptions.

Theorem 4 (Consistency of generator)**.** *Suppose that for every open* $G \subset \Theta$ *that contains* θ_0, $\inf_{\theta \notin G} M_\theta(D_\theta) > M_{\theta_0}(D_{\theta_0})$, *that* $\mathcal{M}^1 := \{\log D_\theta : \theta \in \Theta\}$ *and* $\mathcal{M}^2 := \{\log(1 - D_\theta) \circ T_\theta : \theta \in \Theta\}$ *are* P_0*- and* $\tilde{P}_0$*-Glivenko-Cantelli respectively, and that the estimator* $\hat{\theta}_{n,m}$ *satisfies* $\mathbb{M}^{\hat{\theta}_{n,m}}_{n,m}(\hat{D}^{\hat{\theta}_{n,m}}_{n,m}) \leq \inf_{\theta \in \Theta} \mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m}) + o^*_P(1)$. *Then, under the conclusion of Theorem 2′ with* $\delta_n \to 0$, $h(\hat{\theta}_{n,m}, \theta_0) \to 0$ *in outer probability.*

Theorem 5 (Rate of convergence of generator)**.** *Under Assumptions 4, 6 to 8, and 11,* $h(\hat{\theta}_{n,m}, \theta_0) = O^*_P(n^{-1/2})$.

Theorem 6 (Asymptotic distribution of generator)**.** *Under the conclusion of Theorem 5 and Assumptions 4, 6, 7, and 9 to 11,*

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{n,m} - \theta_0) &= 2\tilde{I}^{-1}_{\theta_0}\sqrt{n}\Big[\mathbb{P}_n(1 - D_{\theta_0})\dot{\ell}_{\theta_0} - \mathbb{P}^{\theta_0}_m D_{\theta_0}\dot{\ell}_{\theta_0}\Big] + o^*_P(1) \\ &\rightsquigarrow N\bigg(0,\ \tilde{I}^{-1}_{\theta_0}\bigg[\Big(P_{\theta_0} + \lim_{n\to\infty}\frac{n}{m}P_0\Big)4D_{\theta_0}(1 - D_{\theta_0})\dot{\ell}_{\theta_0}\dot{\ell}'_{\theta_0}\bigg]\tilde{I}^{-1}_{\theta_0}\bigg).\end{aligned}$$

Corollary 7 (Efficiency of generator)**.** *Under the conclusion of Theorem 6 and Assumption 12,*

$$\sqrt{n}(\hat{\theta}_{n,m} - \theta_0) \rightsquigarrow N\bigg(0, \Big[1 + \lim_{n\to\infty}\frac{n}{m}\Big]I^{-1}_{\theta_0}\bigg).$$

Remark. If $n/m \to 0$, $\hat{\theta}_{n,m}$ attains parametric efficiency.

3.3 What If $\mathcal{D}$ Is Not Rich Enough?

Our theory assumes that $\mathcal{D}$ is a sieve that eventually is capable of representing D_θ. In finite samples, however, we do not know how well $\mathcal{D}$ approximates D_θ. Therefore, it is interesting to know what happens when $\mathcal{D}$ is not a sieve but a fixed class of functions. Although the complete treatment of this case is beyond our scope, we examine what happens to the population problem as we enrich $\mathcal{D}$, e.g., by gradually adding nodes and layers to the neural network.

For simplicity, we maintain Assumptions 2 and 12 and assume that $\mathcal{D}$ contains a constant function 1/2. Let $\tilde{D}_\theta$ be the population maximizer of $M_\theta(D)$ in $\mathcal{D}$. Since $M_\theta(D) - M_\theta(D_\theta) = -2d_\theta(D, D_\theta)^2 + o(d_\theta(D, D_\theta)^2)$ by Theorem 2', $\tilde{D}_\theta$ is equivalent to a minimizer of $d_\theta(D, D_\theta)^2$ in $\mathcal{D}$ up to $o(d_\theta(D, D_\theta)^2)$. Under Assumption 12, $\tilde{D}_{\theta_0} = D_{\theta_0} \equiv 1/2$ and $M_{\theta_0}(1/2) = M_\theta(1/2)$. By Theorem 2',

$$\begin{aligned} M_{\theta_0}(\tilde{D}_{\theta_0}) - M_\theta(\tilde{D}_\theta) &= M_\theta(D_{\theta_0}) - M_\theta(D_\theta) + M_\theta(D_\theta) - M_\theta(\tilde{D}_\theta) \\ &= -2d_\theta(D_{\theta_0}, D_\theta)^2 + 2d_\theta(\tilde{D}_\theta, D_\theta)^2 + o(d_\theta(D_{\theta_0}, D_\theta)^2) + o(d_\theta(\tilde{D}_\theta, D_\theta)^2). \end{aligned}$$

Note that by Lemma 7,

$$\begin{aligned} d_\theta(D_{\theta_0}, D_\theta)^2 &= \int \left(\sqrt{\tfrac{p_0+p_\theta}{2}} - \sqrt{p_0}\right)^2 + \int \left(\sqrt{\tfrac{p_0+p_\theta}{2}} - \sqrt{p_\theta}\right)^2 \\ &= \tfrac{1}{2}\int \tfrac{p_0}{p_0+p_0}(\sqrt{p_0} - \sqrt{p_\theta})^2 + \tfrac{1}{2}\int \tfrac{p_\theta}{p_\theta+p_\theta}(\sqrt{p_0} - \sqrt{p_\theta})^2 + o(h(p_0, p_\theta)^2) \\ &= \tfrac{1}{2}h(p_0, p_\theta)^2 + o(h(p_0, p_\theta)^2). \end{aligned}$$

Thus, we obtain

$$M_{\theta_0}(\tilde{D}_{\theta_0}) - M_\theta(\tilde{D}_\theta) = -h(p_0, p_\theta)^2 + 2d_\theta(\tilde{D}_\theta, D_\theta)^2 + o(h(p_0, p_\theta)^2).$$

If $\mathcal{D}$ contains D_θ, then the second term is zero and the Hellinger curvature allows us to estimate θ efficiently; if $\mathcal{D}$ is a singleton set that contains only 1/2, the first and second terms cancel and the objective function becomes completely flat, rendering estimation of θ impossible. Therefore, the second term represents the loss in efficiency due to the limited capacity of $\mathcal{D}$. For the regular logit case, we know that $\mathcal{D}$ is already rich enough that the curvature admits $\sqrt{n}$-estimation. Then, as we enrich $\mathcal{D}$, it becomes more and more capable of minimizing $d_\theta(\tilde{D}_\theta, D_\theta)^2$, getting closer and closer to efficiency. Of course, such enrichment should not be too fast to avoid overfitting, the conditions of which are characterized above.

4 PRACTICAL ASPECTS

4.1 Choice of Inputs and Discriminators

The method requires the choice of inputs X_i and the choice of the discriminator $\mathcal{D}$. A natural choice of X_i is the entire set of observables, $X_i = (y_i, x_i)$. While our method is intended so that we need not worry about selecting or creating moments,

in the event that we want to emphasize specific aspects of the data, we may still do so by dropping a part of the observables or by transforming them. For example, although our theory allows for discontinuous T_θ, we may still want to adopt the fix of Bruins et al. (2018) to accomodate gradient-based optimization methods. At any rate, the choice of inputs must ensure that the parameters of the structural model are identified.

The choice of the discriminator is more nuanced in that there is no natural, obvious choice.[14] However, if a generative model is not computationally demanding, we may test several discriminators on their abilities to recover the generative parameters. In particular, pick an arbitrary θ as the "true" value and generate data; treat them as the observed data and run adversarial estimation with several choices of $\mathcal{D}$; then, pick one that performs the best. (Indeed, this can also be used to try out different choices of inputs.) If we are also worried about severe misspecification, we may also test using the actual data; split the data into two and make sure that the discriminator cannot separate them too well.

In applications where generating synthetic data is very costly (as in our empirical application), we suggest choosing the discriminator based on cross validation as follows. Fix θ at some value; split the actual data into two, say samples 1 and 2; use sample 1 and synthetic data to estimate D for different choices of $\mathcal{D}$; use sample 2 and new synthetic data to evaluate the classification accuracy of each $\mathcal{D}$; pick the one with the highest accuracy. For the value of θ, we may use estimates from a previous study if available, or try a few different values to check for robustness. See Section 5.4 for more on what we did in our empirical application.

We note that the analysis of the estimator taking into account the selection of inputs and the discriminator is left for future work.

4.2 Autoencoder to Explore the Underlying Dimension

It is helpful to fit an autoencoder on X to get a sense of its underlying dimensionality. Proposition 3 shows that the convergence rate of the neural network discriminator depends on the underlying dimension d^* —rather than the dimension—of X. The bottleneck of the autoencoder (the middle layer with the smallest number of nodes) is indicative of the underlying dimension. See Appendix S.2.4 for intuition and evidence of reduced dimensionality of X_2.

[14] The network structure in Assumption 3 depends on unknown constants such as d^* and α.

4.3 Estimation Procedure

We consider an iterative algorithm that solves the optimization problem in (1).

Algorithm (Estimation)**.**

i. Initialize $\theta = \theta^{(0)}$. Fix a set of random shocks $\{\tilde{X}_i\}_{i=1}^m$ and any random seed if stochastic optimization is used.

ii. For given $\theta = \theta^{(s)}$, generate $\{X_i^{\theta^{(s)}}\}_{i=1}^m$ using $\{\tilde{X}_i\}_{i=1}^m$.

iii. Reset the random seed and train $\hat{D}_{n,m}^{\theta^{(s)}}$ with $\{X_i\}_{i=1}^n$ and $\{X_i^{\theta^{(s)}}\}_{i=1}^m$.

iv. Compute the gradient $\Delta(\theta^{(s)})$ of the objective function with respect to θ.

v. Set $\theta^{(s+1)} = \theta^{(s)} - \xi\Delta(\theta^{(s)})$ where $\xi > 0$ is a learning rate.

vi. Repeat (ii)–(v) until $\Delta(\theta) \approx 0$.

To train the neural network discriminator, we make use of off-the-shelf routines in the R Keras package. They come with implementations of various techniques such as back-propagation, automated differentiation, and stochastic gradient descent.

The algorithm may get stuck in a local minimum. It is advised to use several different initial values to explore a wide space. If it is computationally intensive, we can also start at the value of alternative estimators or previously known estimates. See Appendix S.2.1 for further discussion on the estimation algorithm and details on implementation.

4.4 Inference

The asymptotic variance formula given in Theorem 6 is challenging to estimate since we do not have the closed-form likelihood.[15] We advocate the use of bootstrap as the crux of the theory is that the estimation error of $\hat{D}_{n,m}^{\theta}$ can be ignored in the asymptotics of $\hat{\theta}$. When standard bootstrap is computationally burdensome, we can use the bootstrap proposed by Honoré and Hu (2017), as we do so in Section 5.

Algorithm (Bootstrap)**.**

[15] There is a relation between D_θ and the score and Hessian, $\dot{\ell}_\theta = \frac{1}{D_\theta}\frac{\partial \log(1-D_\theta)}{\partial \theta} = -\frac{1}{1-D_\theta}\frac{\partial \log D_\theta}{\partial \theta}$ and $\ddot{\ell}_\theta + \dot{\ell}_\theta \dot{\ell}_\theta' = \frac{1}{1-D_\theta}[\frac{\partial \log D_\theta}{\partial \theta}\frac{\partial \log D_\theta}{\partial \theta'} - \frac{\partial^2 \log D_\theta}{\partial \theta \partial \theta'}]$, so it is possible to construct the sample counterpart of the variance in Theorem 6, though we do not pursue the proof of its convergence in this paper.

i. Let $\{X_i^*\}_{i=1}^n$ and $\{\tilde{X}_i^{\theta*}\}_{i=1}^m$ be the bootstrap samples of actual and synthetic observations of sizes n and m, drawn randomly with replacement.

ii. Solve (1) with $\{X_i^*\}_{i=1}^n$ and $\{\tilde{X}_i^{\theta*}\}_{i=1}^m$ to obtain a bootstrap estimator $\hat{\theta}_{n,m}^{*(1)}$.

iii. Repeat (i)–(ii) for S times to obtain S bootstrap estimators $\{\hat{\theta}_{n,m}^{*(1)}, \ldots, \hat{\theta}_{n,m}^{*(S)}\}$.

iv. Use the distribution of $\{\hat{\theta}_{n,m}^{*(s)}\}_{s=1}^S$ to approximate the distribution of $\hat{\theta}_{n,m}$.

5 EMPIRICAL APPLICATION: "WHY DO THE ELDERLY SAVE?"

Using the adversarial framework, we examine the elderly's saving, following De Nardi et al. (2010) (henceforth DFJ). The elderly save for various reasons—uncertainty on survival, bequest motive, or ever-rising medical expenses as they age. Different motives for saving yield different implications on policy evaluation such as Medicaid and Medicare. Hence, it is an important and active area of research.

The risk the elderly face is highly heterogeneous, depending on their gender, age, health status, and permanent income. This implies potentially large heterogeneity in the saving motive across individuals; not accounting for this can bias the estimates of utility. For example, the rich live several years more than the poor on average. Failure to reflect this difference can make the rich look thriftier than they are. On the other hand, existing estimation methods such as SMM may suffer from severe lack of precision when various heterogeneity is introduced. This motivates adversarial estimation with a flexible discriminator that parses information in an adaptive and parsimonious way. Indeed, our adversarial estimates, using the same model and the same data as in DFJ, will see considerable gains in precision.

5.1 Agent's Problem

We focus on the behavior of single, retired individuals of age 70 and older. In each period, a surviving single retired agent receives utility $u(c)$ from consumption c and, if they die in that period, additional utility $\phi(e)$ from leaving estate e, where

$$u(c) := \frac{c^{1-\nu}}{1-\nu}, \qquad \phi(e) := \vartheta \frac{(e+k)^{1-\nu}}{1-\nu},$$

and ν is the relative risk aversion and ϑ and k are the intensity and curvature of the bequest motive. Each individual is associated with gender g and permanent income

I, and carries six state variables: age t, asset a_t, nonasset income y_t, health status h_t, medical expense shock ζ_t, and survival s_t. Health and survival are binary, where $h_t = 1$ means they are healthy at age t, and $s_t = 1$ they survive to the next period.

They face three channels of uncertainty: health, survival, and medical expenses. Heath and survival evolve as Markov chains. We denote

$$\pi_H(g, h_t, I, t) := \Pr(h_{t+1} = 1 \mid g, h_t, I, t), \quad \pi_S(g, h_t, I, t) := \Pr(s_{t+1} = 1 \mid g, h_t, I, t).$$

The medical expenses they incur are given by

$$\log m_t = m(g, h_t, I, t) + \sigma(g, h_t, I, t) \times \psi_t,$$

where m and σ are deterministic functions, $\psi_t = \zeta_t + \xi_t$, $\xi_t \sim N(0, \sigma_\xi^2)$, $\zeta_t = \rho\zeta_{t-1} + \epsilon_t$, and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. The nonasset income evolves deterministically as $y_t = y(g, I, t)$. The asset evolves as

$$a_{t+1} = a_t + y_n(ra_t + y_t, \tau) + b_t - m_t - c_t,$$

where $b_t \geq 0$ is the *government transfer*, r the *risk-free pretax rate of return*, $y_n(\cdot, \tau)$ the *posttax income*, and τ the *tax structure*. The agent faces a borrowing constraint $a_t \geq 0$ while social insurance guarantees minimum consumption $c_t \geq \underline{c}$; government transfer b_t is positive only when both constraints cannot be satisfied without it.

The timing in each period is given as follows. Heath h_t and medical expenses m_t realize; then the individual chooses consumption c_t; then survival s_t realizes; if $s_t = 0$, they leave the remaining assets as bequest; if $s_t = 1$, move on to the next period.

Denoting the *cash-on-hand* by $x_t := c_t + a_{t+1}$, the agent's Bellman equation is

$$V_t(x, g, h, I, \zeta) = \max_{c, x'} u(c, h) + \beta[s\mathbb{E}_t V_{t+1}(x', g, h', I, \zeta') + (1 - s)\phi(e)]$$

subject to $x' = (x - c) + y_n(r(x - c) + y', \tau) + b' - m'$, $e = (x - c) - \max\{0, \tilde{\tau}(x - c - \tilde{x})\}$, and $x \geq c \geq \underline{c}$. The first constraint is the budget constraint; the second the bequest (taxed at rate $\tilde{\tau}$ with deduction $\tilde{x}$); the last the borrowing and consumption constraints.

We also look at two transformations: the *marginal propensity to consume at the moment of death* MPC $:= (1 + r)/(1 + r + [\beta\vartheta(1 + r)]^{1/\nu})$ and the *implied asset floor* $\underline{a} := k/[\beta\vartheta(1 + r)]^{1/\nu}$ above which individuals get utility from bequeathing.[16]

[16] The *marginal propensity to bequeath (MPB)* is defined by $1 -$ MPC.

5.2 Data

We use the same data as DFJ, taken from *Assets and Health Dynamics Among the Oldest Old (AHEAD)*. The sample consists of non-institutionalized individuals of age 70 and older in 1994. It contains 8,222 individuals in 6,047 households (3,872 singles and 2,175 couples). The survey took place biyearly from 1994 to 2006. We focus on 3,259 single retired individuals, 592 of which are men and 2,667 women.[17] Of those, 884 were alive in 2006. We drop the first survey in 1994 for reliability, following DFJ.

The survey collects information on age t, financial wealth a_t, nonasset income y_t, medical expenses m_t, and health status h_t. Financial wealth includes real estate, autos, several other liquid assets, retirement accounts, etc. Nonasset income includes social security benefits, veteran's benefits, and other benefits. Medical expenses are total out-of-pocket spending; the average yearly expenses are \$3,700 with standard deviation \$13,400. The permanent income is not observed, but we use as a proxy the ranking of individual average income over time. The health status is a binary variable indicating whether the individual perceives themself as healthy.

5.3 Identifying Role of Health Status

The health status is a variable that was not used in the moments of DFJ; we argue that this gives additional variation to identify the bequest motive (Kopczuk, 2007).

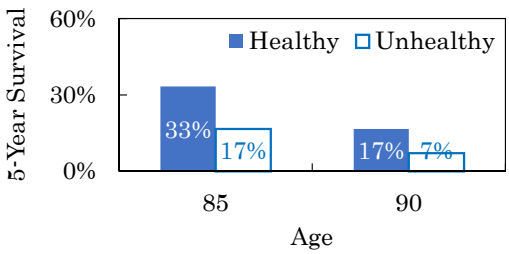
Disentangling the bequest motive from medical expenditure risk is a challenging task. As the bequest is a luxury good, we expect that its identifying power comes from wealthy individuals. Meanwhile, wealthy individuals are also ones with the longest life expectancy, being motivated to save for medical expenses.

Indeed, DFJ document that the medical expenditure for the rich skyrockets after age 95, reaching \$15,000 by age 100. However, if the health condition diminishes their life expectancy, those with shorter horizons would face much less incentive to save for the coming medical expenses while as much incentive to save for bequests.

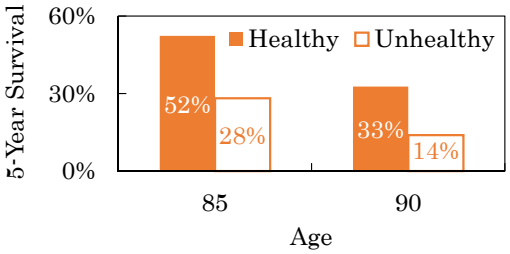
We find some evidence of this in our dataset. Figures 1a and 1b are the proportions of individuals who survive for the next five years at ages 85 and 90, conditional on gender and health. We see that the health status, along with gender, is a strong predictor of life expectancy in years when the medical expenditure soars.

Heterogeneity in the survival materializes as a difference in the savings. Figures 1c

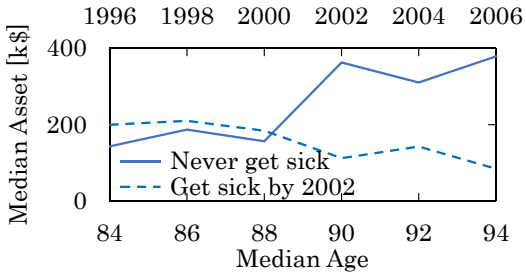
[17] Single individuals are those who were neither married nor cohabiting at any point in the analysis.



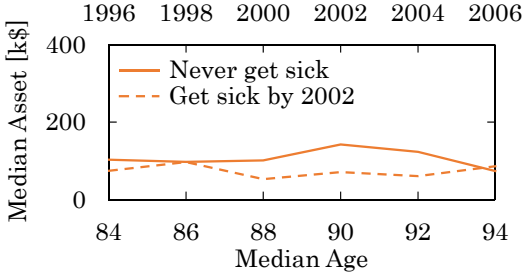
(a) Men's five-year survival rates.



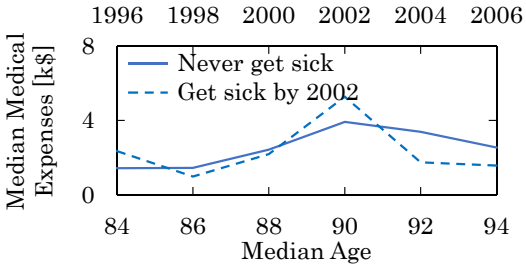
(b) Women's five-year survival rates.



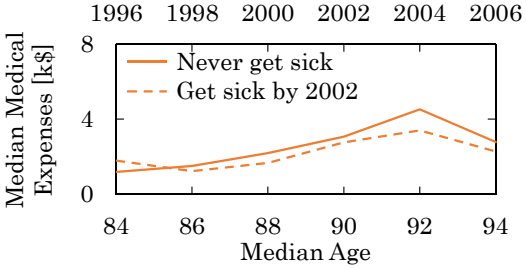
(c) Men's asset.



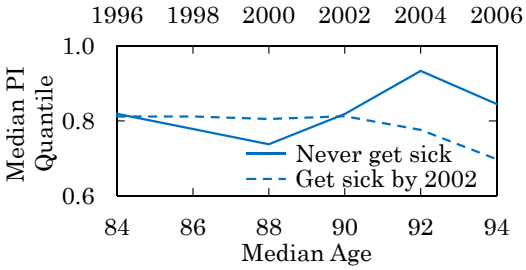
(d) Women's asset.



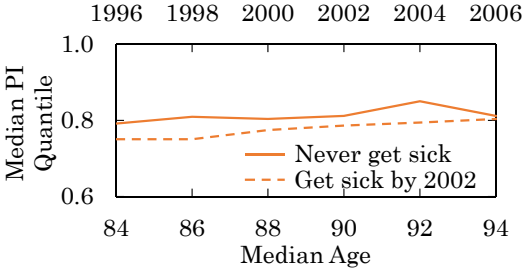
(e) Men's medical expenses.



(f) Women's medical expenses.



(g) Men's permanent income.



(h) Women's permanent income.

Figure 1: Profiles by gender and health. Figures 1c to 1h are for 4–5th PIqs in Cohort 3. Solid lines are for those who stay healthy for the duration of their observation; dashed lines for those who are healthy in 1996 and become unhealthy by 2002.

and 1d give the trajectories of the median assets for the 4th and 5th PI quintiles in Cohort 3. The solid lines are those who were healthy throughout the survey periods

and the dashed lines are those who were healthy in 1996 but reported unhealthy in 1998, 2000, or 2002. We see that men who were exposed to the health shock (hence the survival shock) dig into their savings much more than healthy men. With higher survival rates, women exhibit the trend to a much lesser degree.

Such difference in the asset profiles seems to be driven neither by the difference in medical expenses nor by survival selection among the rich. Figures 1e and 1f show the median medical expenses during the same periods; we observe similar trajectories across gender and health. Figures 1g and 1h show the median PI quantiles of the survivors; if there is attrition of rich or poor individuals that affects the median assets, we expect to see a change in the median PI quantiles. However, they do not differ much by at least age 90 while bifurcation of the asset profiles begins at age 90.

These findings are suggestive that the difference in the asset profiles is attributable to the change in the saving behaviors. The health status changes the exposure to the medical expenditure risk through the survival probability, which then induces changes in the saving behavior by shifting the balance between the bequest motive and medical expenditure risk.

5.4 Estimation

Following DFJ, we carry out estimation in two steps: (1) estimate π_H, π_S, m, σ, ρ_m, σ_ξ, σ_ϵ (in fact, we borrow numbers from DFJ), (2) estimate ν, MPC, and k using our adversarial approach. The parameters r, τ, $\tilde{\tau}$, and $\tilde{x}$ are fixed as in the original paper, and $\beta = 0.971$. For $\underline{c}$, we fix it at \$4,500 to reflect annual social security payments.[18] After the second step, we can also recover ϑ and $\underline{a}$.

We consider two different sets of inputs to the discriminator. The first set consists of the log age of an individual in 1996, permanent income (the aforementioned proxy), the profile (full history) of asset holdings, and the profile of survival indicators,[19]

$$X_1 := (1, \log t_{1996}, I, a_{t_{1996}}, \ldots, a_{t_{2006}}, s_{t_{1998}}, \ldots, s_{t_{2006}}) \in \mathbb{R}^{14}.$$

This is intended to capture similar identifying variation as DFJ. The second set is

[18] In their preferred specification DFJ estimate β and c_{floor}, in addition to ν, MPC and k. Instead, we fix β and c_{floor} to a reasonable value according to the literature. Sensitivity analysis shows that changing $\underline{c}$ mostly affects the risk aversion parameter.

[19] All individuals are alive in 1996, so we drop $s_{t_{1996}}$.

augmented with gender and the profile of health status,

$$X_2 := (X_1, g, h_{t_{1996}}, \ldots, h_{t_{2006}}) \in \mathbb{R}^{21},$$

aiming to capture more variation for the bequest motive as explained in Section 5.3. The results on the autoencoders for X_2 are presented in Appendix S.2.4.

We use cross validation to choose the discriminator (Section 4.1). We focus on feed-forward neural networks with sigmoid activation functions with at most two hidden layers. We fix θ at a preliminary estimate; split the actual data into sample 1 (80%) and sample 2 (20%); estimate D with sample 1, varying the numbers of nodes and layers; evaluate their classification accuracy with sample 2;[20] pick the network configuration with the highest accuracy. The selected neural network discriminator consists of two hidden layers, the first with 20 nodes and the second 10 nodes.

We compare our estimates with SMM in DFJ. They use 150 moments consisting of median assets of groups divided by the cohort and permanent income quintile in each calendar year. The cohort is defined on a four-year window; Cohort 1 are those who were 72–6 years old in 1996; Cohort 2 were 77–81; Cohort 3 were 82–6; Cohort 4 were 87–91; Cohort 5 were 92 and older. Details are in DFJ. We note that accounting for health and gender is infeasible in SMM since it yields too many moments, while it is effortless in our framework.

5.5 Results

Table 1 gives the parameter estimates from DFJ and our adversarial method with specifications X_1 and X_2. Parenthesized numbers are the standard errors; we use Honoré and Hu (2017) to compute them for the adversarial estimates. The first row is the SMM estimates in DFJ. The second and third rows come from the adversarial estimation; the second uses X_1 (14 variables) and the third X_2 (21 variables).

A major difference between our estimates and DFJ's is the curvature of the utility of bequests k. Our estimate is an order of magnitude smaller, which has an important implication: while DFJ conclude only the super rich would obtain utility from bequeathing, our estimate suggests bequeathing matters across the entire permanent income distribution. A related number is the implied asset floor $\underline{a}$. We obtain estimates of \$1,320 and \$4,243, which are on the lower side of the estimates known

[20] We use the classification accuracy provided by Keras's ADAM, which is based on thresholding.

Table 1: Estimates of the structural parameters. The choice of inputs to the discriminator X_1 is intended to capture similar identifying variation as DFJ. The inputs X_2 contain additional variation in gender and health, which is our preferred specification. Standard errors for the adversarial estimates are obtained by the poor (wo)man's bootstrap.

	β	$\underline{c}$ [\$]	ν	ϑ	k [k\$]	MPC	$\underline{a}$ [\$]	Loss
DFJ, Table 3	0.97	2,665	3.84	2,360	273	0.12	36,215	−0.67
	(0.05)	(353)	(0.55)	(8,122)	(446)			
Adversarial X_1	0.97	4,500	6.14	4,865	16.89	0.20	4,243	−0.67
			(.009)	(9.002)	(.030)	(.017)	(19.73)	
Adversarial X_2	0.97	4,500	5.99	192,676	10.02	0.12	1,320	−0.78
			(.005)	(8,112)	(.015)	(.014)	(3.66)	

in the literature. However, they correspond to the 22nd and 24th percentiles of the distribution of assets one period before deaths (see Section 5.6) in our sample, respectively. We interpret these numbers as our method providing a sensible fit of the data. In contrast, DFJ's implied asset floor is \$36,215, which corresponds to the 40th percentile.

Overall, the intensity of the bequest motive is minor in DFJ and X_1 but non-negligible in X_2. While k is low for both X_1 and X_2, MPC is almost twice as large in X_1 compared to X_2. Consequently, individuals care about bequests less than their own consumption according to X_1.

DFJ and adversarial also differ in risk aversion ν. A large value of risk aversion rationalizes the observed saving patterns when the consumption floor $\underline{c}$ is fixed at \$4,500, a reasonable value in the literature.[21]

In line with our theory, adversarial estimation provides substantial gains in precision relative to DFJ. The decrease in standard errors reflects that the data is sufficiently powerful to conclude the importance of the bequest motive, especially when exploiting additional variation in gender and health.

The last column reports the cross-entropy loss of each set of parameter estimates. To make a fair comparison, we take each set of estimates and solve the inner maximization of (1) using X_2 as the input. The loss does not improve with X_1 relative to DFJ but does so substantially with X_2, which is consistent with our observation that gender and health provide useful variation for identifying the bequest motive. This

[21] DFJ's risk aversion estimate increases from 3.84 to 6.04 in an alternative specification where $\underline{c}$ is fixed at \$5,000. However, according to their criterion, the fit of the model decreases substantially.

makes X_2 our preferred specification.

5.6 Fit and Counterfactual Simulations

Similarly as DFJ, we look at the assets one period before deaths to compare the fit and counterfactuals. Individuals who passed away during the survey periods are divided into five groups of permanent income quintiles (PIqs). We take the assets in the last survey when they were alive and sum these across individuals in each group.

Table 2 shows the assets one period before deaths for the actual data and simulation. Adversarial X_2 baseline and DFJ baseline rows are the simulations of the models with parameters equal to the estimates of our preferred specification and of DFJ. Our estimates fit the assets for low PIqs well but overestimates high PIqs, while DFJ show the opposite pattern.[22] In Appendix S.2.5, we provide additional evidence of the good fit of the data.

Next, we perform two counterfactual simulations to measure the elderly's saving motive in terms of (i) bequest and (ii) medical expenditure risk. We simulate the model with the same parameters except that we kill either the bequest incentive, $\phi \equiv 0$, or the medical expenditure risk, $\sigma \equiv 0$. The "(% difference)" rows give the difference of the baseline and counterfactual relative to the baseline.

The contribution of the bequest motive to the savings differs substantially between our estimates and DFJ. In our estimates, the lack of the bequest motive decreases the savings by 13.7% to 19.2%, while DFJ estimates suggest at most 2.1% decrease. This is largely due to the difference in the estimates of the curvature k. According to our estimates, the bequest motive is an important and substantial source of savings for both the poor and the rich. This finding is consistent with Lockwood (2018) who uses additional data on annuity takeup to identify the bequest motive.

The contribution of the medical expenditure risk looks much more in line for the two models. The amount of savings to prepare for uncertain medical expenses is substantial in both predictions. This is because rich individuals live long and hence are at high risk of large medical expenses. Poor individuals do not survive long enough to face it and are more likely to be covered by social insurance programs.

[22] Trimming the observations above the top 1% of mean assets decreases the discrepancy between X_2 and the actual data significantly. Results are available upon request. In addition, the gap in the fit between the poor and the rich might be attributed to the rich doing inter vivos transfers more often than the poor, biasing the assets of the rich downwards toward the end of their lives (McGarry, 1999).

Table 2: Fit of the savings and counterfactual simulations without bequest motive and medical expense risk. "No bequest" rows are the simulations of the model with $\vartheta = 0$ (so $\phi \equiv 0$). "No medical risk" rows are the simulations of the model with $\sigma \equiv 0$ (so $\log m_t = m$). Each number is a cross-sectional sum of assets of individuals one period before their death given in the units of k$, a proxy for their intended bequest. Percentages are relative to the corresponding baselines.

	Permanent income quintile				
	1st	2nd	3rd	4th	5th
Actual data	18,191	25,266	42,006	50,495	85,814
Adversarial X_2 baseline	20,441	26,366	51,339	62,662	110,385
No bequest	17,644	21,587	42,586	50,631	95,212
(% difference)	(13.7%)	(18.1%)	(17.1%)	(19.2%)	(13.7%)
No medical risk	18,890	23,252	43,789	49,385	90,204
(% difference)	(7.6%)	(11.8%)	(14.7%)	(21.2%)	(18.3%)
DFJ baseline	16,527	19,672	38,157	42,737	83,814
No bequest	16,342	19,605	37,387	42,425	83,563
(% difference)	(1.1%)	(0.3%)	(2.1%)	(0.7%)	(0.5%)
No medical risk	16,440	19,242	36,157	38,053	76,080
(% difference)	(0.5%)	(2.2%)	(5.4%)	(11.0%)	(9.4%)

To summarize, our adversarial estimates reveal with precision that the bequest motive contributes in similar magnitudes to the slow decrease in the elderly's savings across PIqs. The uncertainty in medical expenses contribute less for poor individuals.

APPENDIX

A PROOFS

Let $m_q^p := \log \frac{p+q}{2q}$. To derive asymptotic properties of the discriminator, it is helpful to think in terms of the pseudo-objective functions[23]

$$\tilde{M}_\theta(D) := P_0 m_{D_\theta}^D + P_\theta m_{1-D_\theta}^{1-D}, \qquad \tilde{\mathbb{M}}_{n,m}^\theta(D) := \mathbb{P}_n m_{D_\theta}^D + \tilde{\mathbb{P}}_m m_{1-D_\theta}^{1-D} \circ T_\theta,$$

since concavity of the logarithm implies

$$\tilde{\mathbb{M}}_{n,m}^\theta(\hat{D}_{n,m}^\theta) - \tilde{\mathbb{M}}_{n,m}^\theta(D_\theta) \geq \frac{1}{2}[\mathbb{M}_{n,m}^\theta(\hat{D}_{n,m}^\theta) - \mathbb{M}_{n,m}^\theta(D_\theta)] \geq -o_P(n^{-1/2}).$$

[23] See, e.g., van der Vaart and Wellner (1996, Section 3.4.1) and van der Vaart (1998, Section 5.5).

Occasionally, we use the Bernstein "norm" $\|f\|_{P,B} := \sqrt{2P(e^{|f|} - 1 - |f|)}$ that induces a premetric without the triangle inequality (van der Vaart and Wellner, 1996, p. 324).

A.1 Discriminators

Let $\mathcal{M}^{\theta,1}_{n,\delta} := \{m^D_{D_\theta} : D \in \mathcal{D}^\theta_{n,\delta}\}$ and $\mathcal{M}^{\theta,2}_{n,\delta} := \{m^{1-D}_{1-D_\theta} : D \in \mathcal{D}^\theta_{n,\delta}\}$.

Lemma 1 (Maximal inequality for pseudo-cross-entropy discriminator)**.** *For every $D \in \mathcal{D}$, $\tilde{M}_\theta(D) - \tilde{M}_\theta(D_\theta) \le -d_\theta(D, D_\theta)^2/(1+\sqrt{2})^2$. For every $\delta > 0$,*

$$\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta}} \sqrt{n}\Big|(\tilde{\mathbb{M}}^\theta_{n,m} - \tilde{M}_\theta)(D) - (\tilde{\mathbb{M}}^\theta_{n,m} - \tilde{M}_\theta)(D_\theta)\Big| \lesssim J_{[]}(\delta, \mathcal{D}^\theta_{n,\delta}, d_\theta)\left[1 + \sqrt{\frac{n}{m}} + \left(1 + \frac{n}{m}\right)\frac{J_{[]}(\delta, \mathcal{D}^\theta_{n,\delta}, d_\theta)}{\delta^2\sqrt{n}}\right].$$

Proof. Since $\log x \le 2(\sqrt{x} - 1)$ for every $x > 0$,

$$P_0 \log \tfrac{D}{D_\theta} \le 2P_0\Big(\sqrt{\tfrac{D}{D_\theta}} - 1\Big) = \left[2P_0\frac{\sqrt{D(p_0+p_\theta)}}{\sqrt{p_0}} - \int D(p_0+p_\theta) - \int p_0\right] + (P_0 + P_\theta)(D - D_\theta) = -h_\theta(D, D_\theta)^2 + (P_0 + P_\theta)(D - D_\theta).$$

Similarly, $P_\theta \log \frac{1-D}{1-D_\theta} \le -h_\theta(1-D, 1-D_\theta)^2 - (P_0 + P_\theta)(D - D_\theta)$. Replacing D and $1 - D$ with $(D + D_\theta)/2$ and $(1 - D + 1 - D_\theta)/2$ and summing them up yield

$$P_0 m^D_{D_\theta} + P_\theta m^{1-D}_{1-D_\theta} \le -h_\theta\Big(\tfrac{D+D_\theta}{2}, D_\theta\Big)^2 - h_\theta\Big(\tfrac{1-D+1-D_\theta}{2}, 1 - D_\theta\Big)^2.$$

Since $\sqrt{2}h_\theta(\frac{p+q}{2}, q) \le h_\theta(p, q) \le (1+\sqrt{2})h_\theta(\frac{p+q}{2}, q)$ (van der Vaart and Wellner, 1996, Problem 3.4.4), we obtain the first inequality. For the second inequality, observe that

$$\sqrt{n}\Big[(\tilde{\mathbb{M}}^\theta_{n,m} - \tilde{M}_\theta)(D) - (\tilde{\mathbb{M}}^\theta_{n,m} - \tilde{M}_\theta)(D_\theta)\Big] = \sqrt{n}(\mathbb{P}_n - P_0)m^D_{D_\theta} + \sqrt{n}(\mathbb{P}^\theta_m - P_\theta)m^{1-D}_{1-D_\theta}.$$

Therefore, it suffices to separately bound

$$\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta}} \Big|\sqrt{n}(\mathbb{P}_n - P_0)m^D_{D_\theta}\Big| \quad \text{and} \quad \sqrt{\tfrac{n}{m}}\,\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta}} \Big|\sqrt{m}(\mathbb{P}^\theta_m - P_\theta)m^{1-D}_{1-D_\theta}\Big|.$$

Since $m^D_{D_\theta}, m^{1-D}_{1-D_\theta} \ge \log(1/2)$ and $e^{|x|} - 1 - |x| \le 4(e^{x/2} - 1)^2$ for every $x \ge \log(1/2)$,

$$\Big\|m^D_{D_\theta}\Big\|^2_{P_0,B} \le 8P_0\Big(e^{m^D_{D_\theta}/2} - 1\Big)^2 \le 8h_\theta\Big(\tfrac{D+D_\theta}{2}, D_\theta\Big)^2 \le 4h_\theta(D, D_\theta)^2,$$
$$\Big\|m^{1-D}_{1-D_\theta}\Big\|^2_{P_\theta,B} \le 4h_\theta(1-D, 1-D_\theta)^2.$$

By van der Vaart and Wellner (1996, Lemma 3.4.3), the first supremum is bounded by $J_{[]}(2\delta, \mathcal{M}^{\theta,1}_{n,\delta}, \|\cdot\|_{P_0,B})[1 + J_{[]}(2\delta, \mathcal{M}^{\theta,1}_{n,\delta}, \|\cdot\|_{P_0,B})/(4\delta^2\sqrt{n})]$. Let $[\ell, u]$ be an ε-bracket in $\mathcal{D}$ with respect to d_θ. Since $u - \ell \geq 0$ and $e^{|x|} - 1 - |x| \leq 2(e^{x/2} - 1)^2$ for $x \geq 0$,

$$\left\|m^u_{D_\theta} - m^\ell_{D_\theta}\right\|^2_{P_0,B} \leq 4\int\left(\sqrt{\tfrac{u+D_\theta}{\ell+D_\theta}} - 1\right)^2 p_0 \leq 4\int\left(\sqrt{u+D_\theta} - \sqrt{\ell+D_\theta}\right)^2(p_0+p_\theta)$$
$$\leq 4h_\theta(u,\ell)^2 \leq 4\varepsilon^2.$$

Thus, $[m^\ell_{D_\theta}, m^u_{D_\theta}]$ makes a 2ε-bracket in $\mathcal{M}^{\theta,1}$ with respect to $\|\cdot\|_{P_0,B}$, so $J_{[]}(2\delta, \mathcal{M}^{\theta,1}_{n,\delta}, \|\cdot\|_{P_0,B}) \leq 2J_{[]}(\delta, \mathcal{D}^\theta_{n,\delta}, d_\theta)$. Analogous argument for the second supremum yields the second inequality. ■

Now, Theorems 1 and 2 follow immediately from the following general versions.

Theorem 1' (Rate of convergence of discriminator)**.** *Suppose Assumption 4 holds and* $\mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m}) \geq \mathbb{M}^\theta_{n,m}(D_\theta) - O_P(\delta_n^2)$ *for a nonnegative sequence* δ_n*. If we have* $J_{[]}(\delta_n, \mathcal{D}^\theta_{n,\delta_n}, d_\theta) \lesssim \delta_n^2\sqrt{n}$ *and there exists* $\alpha < 2$ *such that* $J_{[]}(\delta, \mathcal{D}^\theta_{n,\delta}, d_\theta)/\delta^\alpha$ *is decreasing in* δ*, then* $d_\theta(\hat{D}^\theta_{n,m}, D_\theta) = O^*_P(\delta_n)$*.*

Proof. As noted at the beginning of the section, the condition of the theorem implies $\tilde{\mathbb{M}}^\theta_{n,m}(\hat{D}^\theta_{n,m}) \geq \tilde{\mathbb{M}}^\theta_{n,m}(D_\theta) - O_P(\delta_n^2)$. Then, the theorem follows from van der Vaart and Wellner (1996, Theorem 3.4.1) applied with Lemma 1. ■

Theorem 2' (Rate of convergence of objective function)**.** *Under Assumption 2,* $M_\theta(D) - M_\theta(D_\theta) = -2d_\theta(D, D_\theta)^2 + o(d_\theta(D, D_\theta)^2)$*. Under the assumptions of Theorem 1' and Assumption 2,* $\mathbb{M}^\theta_{n,m}(\hat{D}^\theta_{n,m}) - \mathbb{M}^\theta_{n,m}(D_\theta) = O^*_P(\delta_n^2)$*.*

Proof. Note that for every $D \in \mathcal{D}$,

$$\mathbb{M}^\theta_{n,m}(D) - \mathbb{M}^\theta_{n,m}(D_\theta) = M_\theta(D) - M_\theta(D_\theta) + (\mathbb{P}_n - P_0)\log\tfrac{D}{D_\theta} + (\mathbb{P}^\theta_m - P_\theta)\log\tfrac{1-D}{1-D_\theta}.$$

Let $W_1 := \sqrt{\frac{D}{D_\theta}} - 1$, $W_2 := \sqrt{\frac{1-D}{1-D_\theta}} - 1$, and $\delta := d_\theta(D, D_\theta)$. By Taylor's theorem, $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{2}x^2R(x)$ where $R(x) = O(x)$ as $x \to 0$. Therefore,

$$M_\theta(D) - M_\theta(D_\theta) = P_0\log\tfrac{D}{D_\theta} + P_\theta\log\tfrac{1-D}{1-D_\theta} = 2P_0\log(1+W_1) + 2P_\theta\log(1+W_2)$$
$$= 2P_0W_1 - P_0W_1^2 + P_0W_1^2R(W_1) + 2P_\theta W_2 - P_\theta W_2^2 + P_\theta W_2^2R(W_2).$$

Note that $P_0W_1^2 = P_0(\sqrt{D/D_\theta} - 1)^2 = h_\theta(D, D_\theta)^2$ and $P_\theta W_2^2 = h_\theta(1-D, 1-D_\theta)^2$.

Since $W_j^2 \geq 0$, this implies that $W_1(X_i)^2 = O_P(\delta^2)$ and $W_2(X_i^\theta)^2 = O_P(\delta^2)$. Moreover,

$$\begin{aligned} 2P_0W_1 &= \left[2P_0\tfrac{\sqrt{D(p_0+p_\theta)}}{\sqrt{p_0}} - \int D(p_0+p_\theta) - \int p_0\right] + (P_0+P_\theta)(D-D_\theta) \\ &= -h_\theta(D,D_\theta)^2 + (P_0+P_\theta)(D-D_\theta), \\ 2P_\theta W_2 &= -h_\theta(1-D,1-D_\theta)^2 - (P_0+P_\theta)(D-D_\theta). \end{aligned}$$

Thus, $2P_0W_1 + 2P_\theta W_2 = -d_\theta(D,D_\theta)^2$ and $W_1(X_i)$ and $W_2(X_i^\theta)$ are $o_P(1)$ since $|D - D_\theta| \leq 2|\sqrt{D} - \sqrt{D_\theta}|$. Also, $R(W_1(X_i))$ and $R(W_2(X_i^\theta))$ are $o_P(1)$. For $1/5 \leq c < 1$,

$$\begin{aligned} |P_0W_1^2R(W_1)| &\leq P_0W_1^2|R(W_1)|\mathbb{1}\{W_1 \leq -c\} + P_0W_1^2|R(W_1)|\mathbb{1}\{W_1 > -c\} \\ &\leq P_0(-R(W_1)\mathbb{1}\{W_1 \leq -c\}) + P_0W_1^2|R(-c) \vee R(W_1)|. \end{aligned}$$

Since $R(x) < 1$, the second term is $o(\delta^2)$ for every c by the dominated convergence theorem. By the diagonal argument, there exists a sequence $c \to 1$ for given $D \to D_\theta$ such that the second term remains $o(\delta^2)$. Since $0 < -R(x) < -2\log(1+x)$ for $x \leq -\frac{1}{5}$,

$$\begin{aligned} P_0(-R(W_1)\mathbb{1}\{W_1 \leq -c\}) \leq P_0(\log\tfrac{D_\theta}{D}\mathbb{1}\{W_1 \leq -c\}) &= P_0(\tfrac{D}{D_\theta}\log\tfrac{D_\theta}{D}\cdot\tfrac{D_\theta}{D}\mathbb{1}\{W_1 \leq -c\}) \\ &\leq \sup_{x \geq (1-c)^{-2}} |\tfrac{1}{x}\log x| \cdot P_0(\tfrac{D_\theta}{D}\mathbb{1}\{W_1 \leq -c\}). \end{aligned}$$

The first term is $o(1)$ as $c \to 1$. The second term is bounded by $P_0(\frac{D_\theta}{D}\mathbb{1}\{W_1 \leq -\frac{1}{5}\}) = P_0(W_1 \leq -\frac{1}{5})P_0(\frac{D_\theta}{D} \mid \frac{D_\theta}{D} \geq \frac{25}{16}) \leq P_0(W_1 \leq -\frac{1}{5})M$ by Assumption 2. By Markov's inequality, $P_0(W_1 \leq -\frac{1}{5}) \leq 25P_0W_1^2 = O(\delta^2)$. Thus, we have shown $|P_0W_1^2R(W_1)| = o(\delta^2)$. Similarly, $|P_\theta W_2^2R(W_2)| = o(\delta^2)$. Then, the first claim follows.

Now, we bound the suprema of the two random terms

$$\mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} \left|\sqrt{n}(\mathbb{P}_n - P_0)\log\tfrac{D}{D_\theta}\right| \quad \text{and} \quad \mathbb{E}^* \sup_{D \in \mathcal{D}^\theta_{n,\delta_n}} \left|\sqrt{m}(\mathbb{P}^\theta_m - P_\theta)\log\tfrac{1-D}{1-D_\theta}\right|.$$

Under Assumption 2, it follows from (the remark after) Lemma 5 that for $D \in \mathcal{D}^\theta_{n,\delta_n}$,

$$\left\|\tfrac{1}{2}\log\tfrac{D}{D_\theta}\right\|^2_{P_0,B} \leq 2(1+M)h_\theta(D,D_\theta)^2, \ \left\|\tfrac{1}{2}\log\tfrac{1-D}{1-D_\theta}\right\|^2_{P_\theta,B} \leq 2(1+M)h_\theta(1-D,1-D_\theta)^2.$$

Assumption 2 also implies that an ε-bracket in $\mathcal{M}^{\theta,1}$ induces

$$\left\|\log\tfrac{u}{D_\theta} - \log\tfrac{\ell}{D_\theta}\right\|^2_{P_0,B} \leq 4P_0\Big(\sqrt{\tfrac{u}{\ell}} - 1\Big)^2 = 4(P_0+P_\theta)\tfrac{D_\theta}{\ell}(\sqrt{u} - \sqrt{\ell})^2 \leq Cd_\theta(u,\ell)^2,$$

$$\left\|\log\tfrac{1-\ell}{1-D_\theta}-\log\tfrac{1-u}{1-D_\theta}\right\|_{P_\theta,B}^2\le 4(P_0+P_\theta)\tfrac{1-D_\theta}{1-u}(\sqrt{1-\ell}-\sqrt{1-u})^2\le Cd_\theta(u,\ell)^2,$$

for some $C>0$. By similar arguments as in the proof of Theorem 1', the two suprema are of orders $\sqrt{n}\delta_n^2$ and $\sqrt{m}\delta_n^2$. [24] With Assumption 4 follows the theorem. ■

A.2 Neural Network Discriminators

We establish a bound on the bracketing number of a (possibly sparse) neural network with bounded weights and Lipschitz activation functions.

Lemma 2 (Bracketing number of neural network with bounded weights)**.** *Let $\mathcal{F}$ be a class of neural networks defined in Example 4. Denote the total number of nonzero weights by S and the maximum number of nonzero weights in each node (except for the first layer taking inputs) by $\tilde{U}$.*[25] *Assume that σ and Λ are Lipschitz with constant 1 and $\|w\|_\infty\le C$ for some C. Assume innocuously that $\tilde{U}C\ge 2$ and let $\sigma_0:=|\sigma(0)|$. Define $F:\mathbb{R}^d\to\mathbb{R}$ by $F(x):=\sigma_0+\|x\|_\infty$. Then, for any premetric $d_{\mathcal{F}}$ and $\|f\|_{d_{\mathcal{F}}}:=\sup_{g\in\mathcal{F}}d_{\mathcal{F}}(g-f/2,g+f/2)$,*

$$N_{[]}(\|\varepsilon F\|_{d_{\mathcal{F}}},\mathcal{F},d_{\mathcal{F}})\le\left\lceil\frac{2(L+1)(\tilde{U}C)^{L+1}d}{\varepsilon}\right\rceil^S.$$

For a fully connected network, $\tilde{U}=U$ and $S=(LU+1)U+(d-U)U$. For a hierarchical network in Bauer and Kohler (2019), $S=O(\tilde{U}^{(L+4)/3}d)$.

Proof. Recall from Example 4 that $f(x;w)=\Lambda(w_L'\sigma(w_{L-1}'\sigma(\cdots w_1'\sigma(w_0'x))))$. We can bound the outputs of the ℓth layer by

$$\begin{aligned}\|\sigma(w_{\ell-1}'\sigma(\cdots))\|_\infty&\le\sigma_0+\|w_{\ell-1}'\sigma(\cdots)\|_\infty\le\sigma_0+\tilde{U}C\|\sigma(\cdots)\|_\infty\\&\le[1+\tilde{U}C+\cdots+(\tilde{U}C)^{\ell-1}]\sigma_0+\tilde{U}^{\ell-1}C^\ell d\|x\|_\infty\\&\le\tilde{U}^{\ell-1}C^\ell(\tilde{U}\sigma_0+d\|x\|_\infty)\le(\tilde{U}C)^\ell d(\sigma_0+\|x\|_\infty),\end{aligned}$$

where the fourth inequality holds for $\tilde{U}C\ge 2$. For two sets of weights, w and $\tilde{w}$,

$$\begin{aligned}|f(x;w)-f(x;\tilde{w})|\le\tilde{U}\|w_L-\tilde{w}_L\|_\infty(\|\sigma(w_{L-1}'\sigma(\cdots))\|_\infty\vee\|\sigma(\tilde{w}_{L-1}'\sigma(\cdots))\|_\infty)\\+\tilde{U}C\|\sigma(w_{L-1}'\sigma(\cdots))-\sigma(\tilde{w}_{L-1}'\sigma(\cdots))\|_\infty\end{aligned}$$

[24] We can write $\|\frac{1}{2}\log\frac{D}{D_\theta}\|_{P_0,B}^2\le[2(1+M)\vee C]h_\theta(D,D_\theta)^2$ and $\|\log\frac{u}{D_\theta}-\log\frac{\ell}{D_\theta}\|_{P_0,B}^2\le[2(1+M)\vee C]d_\theta(u,\ell)^2$ to apply the same argument as in Theorem 1'.

[25] The number of nonzero elements in each row of each matrix w_ℓ, $\ell\ge 1$, is bounded by $\tilde{U}$.

$$
\begin{aligned}
&\leq \tilde{U}^{L+1} C^L d \|w_L - \tilde{w}_L\|_\infty (\sigma_0 + \|x\|_\infty) + \cdots \\
&+ \tilde{U}^{L+1} C^L d \|w_1 - \tilde{w}_1\|_\infty (\sigma_0 + \|x\|_\infty) + \tilde{U}^L C^L d \|w_0 - \tilde{w}_0\|_\infty \|x\|_\infty \\
&\leq (L+1) \tilde{U}^{L+1} C^L d \|w - \tilde{w}\|_\infty (\sigma_0 + \|x\|_\infty).
\end{aligned}
$$

Let $A := (L+1)\tilde{U}^{L+1} C^L d$. Partitioning the weight space $[-C, C]^S$ into cubes of length $2\varepsilon/A$ creates $\lceil CA/\varepsilon \rceil^S$ cubes. Hence, $N(\varepsilon, [-C, C]^S, \|\cdot\|_\infty) \leq \lceil CA/\varepsilon \rceil^S$. The bound follows by van der Vaart and Wellner (1996, Theorem 2.7.11), observing that the proof thereof works for a premetric with modification of $2\varepsilon\|F\|$ to $\|2\varepsilon F\|_{d_\mathcal{F}}$.

For a fully connected network, the number of all weights is dU (weights for the first layer) plus $(L-1)U^2$ (weights for the remaining hidden layers) plus U (weights in the output layer), summing to $(LU+1)U + (d-U)U$.[26] For a network $\mathcal{H}^{(0)}$ in Bauer and Kohler (2019) (in their notation), the number of all weights is $A^{(0)} := d(4d^* M_*) + 4d^* M_* + M_* = 4(1+d)d^* M_* + M_*$. For $\mathcal{H}^{(1)}$, $A^{(1)} := A^{(0)} K + K(4d^* M_*) + 4d^* M_* + M_* = A^{(0)} K + 4(1+K)d^* M_* + M_*$. For $\mathcal{H}^{(l)}$, $A^{(l)} := A^{(l-1)} K + 4(1+K)d^* M_* + M_* = A^{(0)} K^l + \sum_{j=0}^{l-1} K^j [4(1+K)d^* M_* + M_*] = 4d^* M_* [(1+d)K^l + \frac{1-K^l}{1-K}(1+K)] + M_* \frac{1-K^{l+1}}{1-K} = O(dd^* M_* K^l)$. Then use $L = 2 + 3l$ and $\tilde{U} = M_* \vee (4d^*) \vee K$. ■

Remark. Lemma 2 assumes a Lipschitz property for the activation and output functions, which accommodates ReLU, softplus, and sigmoid.

Remark. If a premetric d satisfies the property that $\ell \leq f \leq u$ implies $d(\ell, f) \leq d(\ell, u)$, then the ε-covering number of $\mathcal{F}$ with respect to d is bounded by $N_{[]}(\varepsilon, \mathcal{F}, d)$. Another popular way to bound the covering number is by the dimension of $\mathcal{F}$ (van der Vaart and Wellner, 1996, Chapter 2.6; Anthony and Bartlett, 1999, Chapter 12). However, dimension bounds for neural networks often come with strong functional-form assumptions on the activation function (Bartlett and Maass, 2003; Bartlett et al., 2019). Our approach does not require that at the cost of bounded weights.

Proof of Proposition 3. We use Lemma 2 to bound the bracketing number in Theorem 1'. Since D is nonnegative, we can extend d_θ to accommodate arbitrary functions f_1 and f_2 by $d_\theta(f_1, f_2) := d_\theta(0 \vee f_1, 0 \vee f_2)$. In the notation of Lemma 2,

$$
\begin{aligned}
\|\varepsilon^2 F\|_{d_\theta}^2 &= \sup_{D \in \mathcal{D}} d_\theta(D - \varepsilon^2 F/2, D + \varepsilon^2 F/2)^2 \leq h_\theta(0, \varepsilon^2 F)^2 + h_\theta(0, \varepsilon^2 F)^2 \\
&= 2\varepsilon^2 (P_0 + P_\theta) F = 2\varepsilon^2 [2\sigma_0 + (P_0 + P_\theta)\|X\|_\infty] =: B\varepsilon^2.
\end{aligned}
$$

[26] If the network has a bias term, the actual variable weights are slightly fewer, but it does not change the order.

Since P_0 and P_θ have uniformly bounded first moments, $B < \infty$. Therefore,

$$\log N_{[]}(\varepsilon, \mathcal{D}_n, d_\theta) \le \log N_{[]}\Big(\Big\|\tfrac{\varepsilon^2}{B}F\Big\|_{d_\theta}, \mathcal{D}_n, d_\theta\Big) \le S \log\Big\lceil \tfrac{2B(L+1)(\tilde{U}C)^{L+1}d}{\varepsilon^2}\Big\rceil.$$

The same bound holds for $\log N_{[]}(\varepsilon, 1-\mathcal{D}_n, h_\theta)$. Observe that for $0 < \delta \le e^a$,

$$\int_0^\delta \sqrt{a - \log\varepsilon}\, d\varepsilon = \tfrac{\sqrt{\pi}}{2} e^a \operatorname{erfc}\Big(\sqrt{a-\log\delta}\Big) + \delta\sqrt{a-\log\delta} \lesssim \delta\sqrt{a-\log\delta}.$$

Therefore,

$$\begin{aligned} J_{[]}(\delta, \mathcal{D}_n, h_\theta) &\lesssim \int_0^\delta \sqrt{1 + S[\log(2B(L+1)(\tilde{U}C)^{L+1}d) - 2\log\varepsilon]}\, d\varepsilon \\ &\lesssim \delta\sqrt{1 + S[\log(2B(L+1)(\tilde{U}C)^{L+1}d) - 2\log\delta]} \lesssim \delta\sqrt{SL\log(\tilde{U}C) - S\log\delta}. \end{aligned}$$

Again, $J_{[]}(\delta, 1-\mathcal{D}_n, h_\theta)$ is likewise bounded. By Theorem 1' and Assumption 4, this gives rise to the rate

$$\delta_n = O\Big(\sqrt{\tfrac{SL\log(\tilde{U}C) + S\log n}{n}}\Big). \tag{2}$$

To attain this, the sieve must be rich enough so that $\inf_{D\in\mathcal{D}_n} d_\theta(D, D_\theta) \lesssim \delta_n$.

Since $\mathcal{D}_n = \Lambda(\mathcal{H}^{(l)})$, we use Bauer and Kohler (2019, Theorem 3) to derive the network configuration that attains this rate. For that, we need to choose "N, η_n, a_n, M_n" in their notation. First, we set $N = q$ and $\eta_n = \delta_n^2$. By subexponentiality, we have $\log P_0(\|X\|_\infty > a) + \log P_\theta(\|X\|_\infty > a) \lesssim -a$ for large a. Therefore, we want $a_n \gg -2\log\delta_n$ so that $(P_0 + P_\theta)(\|X\|_\infty > a_n) \lesssim \delta_n^2$.[27] We can do this by setting $a_n = (-\log\delta_n)^2$. Finally, we want to choose M_n so that $a_n^{N+q+3} M_n^{-p} \sim \delta_n$; set $M_n = (\log\delta_n)^{2(N+q+3)/p}/\delta_n^{1/p}$. Let $A \subset [-a_n, a_n]^d$ be the set for which $(P_0+P_\theta)(A) \le c\eta_n$ in Bauer and Kohler (2019, Theorem 3). Then,

$$\begin{aligned} h_\theta(D, D_\theta)^2 &\le \Big(\int_{\|x\|_\infty > a_n} + \int_A + \int_{\{\|x\|_\infty \le a_n\}\setminus A}\Big)(\sqrt{D} - \sqrt{D_\theta})^2(p_0 + p_\theta) \\ &\le (P_0+P_\theta)(\|X\|_\infty > a_n) + (P_0+P_\theta)(A) + \int_{\{\|x\|_\infty \le a_n\}\setminus A} (\sqrt{D} - \sqrt{D_\theta})^2(p_0+p_\theta). \end{aligned}$$

The first two terms are bounded by $\delta_n^2 + c\delta_n^2$. For $D = \Lambda(f)$,

$$\int_{\{\|x\|_\infty \le a_n\}\setminus A} (\sqrt{D} - \sqrt{D_\theta})^2(p_0+p_\theta) = \int_{\{\|x\|_\infty \le a_n\}\setminus A} \Big(\sqrt{\Lambda(f)} - \sqrt{\Lambda(\Lambda^{-1}\circ D_\theta)}\Big)^2 (p_0+p_\theta)$$

[27] If we set $a_n \sim -2\log\delta_n$, then we can only say $(P_0+P_\theta)(\|X\|_\infty > a_n) \lesssim \delta_n^c$ for some c.

$$\leq \tfrac{2}{27}\left\|f - \Lambda^{-1} \circ D_\theta\right\|^2_{\infty,\{\|x\|_\infty \leq a_n\}\setminus A} = \tfrac{2}{27}\left\|f - \log \tfrac{p_0}{p_\theta}\right\|^2_{\infty,\{\|x\|_\infty \leq a_n\}\setminus A},$$

since $\sqrt{\Lambda(\cdot)}$ is Lipschitz with constant $1/(3\sqrt{3})$. We may likewise bound $h_\theta(1-D, 1-D_\theta)^2$. By Bauer and Kohler (2019, Theorem 3), $\inf_{f\in\mathcal{H}^{(l)}} \|f - \log \frac{p_0}{p_\theta}\|_{\infty,\{\|x\|_\infty \leq a_n\}\setminus A} \lesssim \delta_n$. Thus, we obtain $\inf_{D\in\mathcal{D}_n} d_\theta(D, D_\theta) \lesssim \delta_n$.

Meanwhile, substituting $S = O(dd^*M_*K^l) \sim M_*$, $\tilde{U} = M_* \vee (4d^*) \vee K \sim M_*$, $C = \alpha$, and $L = 2 + 3l = O(1)$ into (2) yields $\delta_n^2 \sim M_* \frac{\log(M_*\alpha)+\log n}{n}$. Here,

$$M_* = \binom{d^*+N}{d^*}(N+1)(M_n+1)^{d^*} \sim M_n^{d^*} = \frac{(\log \delta_n)^{2d^*(N+q+3)/p}}{\delta_n^{d^*/p}},$$
$$\alpha = \frac{M_n^{d^*+p(2N+3)+1}}{\eta_n}\log n = \frac{(\log \delta_n)^{2(N+q+3)[d^*+p(2N+3)+1]/p}}{\delta_n^{2+[d^*+p(2N+3)+1]/p}}\log n.$$

Thus, $\delta_n \sim [(\log n)^{\frac{p+2d^*(N+q+3)}{p}}/n]^{\frac{p}{2p+d^*}}$. The result follows by substituting $N = q$. ■

A.3 Generators

Proof of Theorem 4. For simplicity, we omit the subscripts n, m. Note that

$$\mathbb{M}^{\hat{\theta}}(D_{\hat{\theta}}) - \inf_{\theta\in\Theta}\mathbb{M}^\theta(D_\theta) \leq \left[\mathbb{M}^{\hat{\theta}}(\hat{D}^{\hat{\theta}}) - \inf_{\theta\in\Theta}\mathbb{M}^\theta(\hat{D}^\theta)\right] + \left[\mathbb{M}^{\hat{\theta}}(D_{\hat{\theta}}) - \mathbb{M}^{\hat{\theta}}(\hat{D}^{\hat{\theta}})\right] + \sup_{\theta\in\Theta}\left[\mathbb{M}^\theta(\hat{D}^\theta) - \mathbb{M}^\theta(D_\theta)\right].$$

The first difference is less than $o_P^*(1)$ by assumption; the other two are $o_P^*(1)$ by Theorem 2'. Therefore, $\mathbb{M}^{\hat{\theta}}(D_{\hat{\theta}}) \leq \inf_{\theta\in\Theta}\mathbb{M}^\theta(D_\theta) + o_P^*(1)$.

By the assumption of Glivenko-Cantelli, $\|\mathbb{P}_n - P_0\|_{\mathcal{M}^1} \to 0$ and $\|\tilde{\mathbb{P}}_m - \tilde{P}_0\|_{\mathcal{M}^2} \to 0$ in outer probability as $n, m \to \infty$. By van der Vaart and Wellner (1996, Corollary 3.2.3 (i)), it follows that $\hat{\theta}_{n,m} \to \theta_0$ in outer probability. ■

The next theorem is a generalization of Theorem 5 on the rate of convergence of $\hat{\theta}_{n,m}$. The parametric rate can be achieved if P_{θ_0} is "close enough" to P_0.

Theorem 5' (Rate of convergence of generator)**.** *Suppose*

$$\mathbb{M}_{n,m}^{\hat{\theta}_{n,m}}(\hat{D}_{n,m}^{\hat{\theta}_{n,m}}) \leq \mathbb{M}_{n,m}^{\theta_0}(\hat{D}_{n,m}^{\theta_0}) + O_P^*(\kappa_n^2),$$
$$\left[\mathbb{M}_{n,m}^{\hat{\theta}_{n,m}}(\hat{D}_{n,m}^{\hat{\theta}_{n,m}}) - \mathbb{M}_{n,m}^{\theta_0}(\hat{D}_{n,m}^{\theta_0})\right] - \left[\mathbb{M}_{n,m}^{\hat{\theta}_{n,m}}(D_{\hat{\theta}_{n,m}}) - \mathbb{M}_{n,m}^{\theta_0}(D_{\theta_0})\right] = O_P^*(\kappa_n^2)$$

for a nonnegative sequence κ_n. Then, under Assumptions 4, 7, 8, and 11, $h(\hat{\theta}_{n,m}, \theta_0) \vee \tilde{h}(\hat{\theta}_{n,m}, \theta_0) = O_P^*(\kappa_n \vee n^{-1/2})$.

Proof. The displayed condition implies $\mathbb{M}^{\hat{\theta}}(D_{\hat{\theta}}) \le \mathbb{M}^{\theta_0}(D_{\theta_0}) + O_P^*(\kappa_n^2)$, so we apply van der Vaart and Wellner (1996, Theorem 3.2.5) to $\mathbb{M}^{\theta}(D_\theta)$. By Assumptions 7 and 11, $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \gtrsim h(\theta, \theta_0)^2 \wedge c$ for some $c > 0$ globally in $\theta \in \Theta$.

Next, we show the convergence of the sample objective function. Note that

$$(\mathbb{M}^{\theta_0} - M_{\theta_0})(D_{\theta_0}) - (\mathbb{M}^{\theta} - M_\theta)(D_\theta) = (\mathbb{P}_n - P_0) \log \tfrac{D_{\theta_0}}{D_\theta} + (\tilde{\mathbb{P}}_m - \tilde{P}_0) \log \tfrac{(1-D_{\theta_0}) \circ T_{\theta_0}}{(1-D_\theta) \circ T_\theta}.$$

By Lemma 6, $\| \log \frac{D_{\theta_0}}{D_\theta} \|_{P_0,B}^2 \le 4h(\theta, \theta_0)^2$ and $\| \log \frac{(1-D_{\theta_0}) \circ T_{\theta_0}}{(1-D_\theta) \circ T_\theta} \|_{\tilde{P}_0,B}^2 \le 4\tilde{h}(\theta, \theta_0)^2$. For $\delta > 0$, define $\mathcal{M}_\delta^1 := \{\log \frac{D_{\theta_0}}{D_\theta} : h(\theta, \theta_0) \le \delta\}$ and $\mathcal{M}_\delta^2 := \{\log \frac{(1-D_{\theta_0}) \circ T_{\theta_0}}{(1-D_\theta) \circ T_\theta} : \tilde{h}(\theta, \theta_0) \le \delta\}$. By van der Vaart and Wellner (1996, Lemma 3.4.3),

$$\mathbb{E}^* \sup_{h(\theta,\theta_0)<\delta} \left| \sqrt{n}(\mathbb{P}_n - P_0) \log \tfrac{D_{\theta_0}}{D_\theta} \right| \lesssim J_{[]}(2\delta, \mathcal{M}_\delta^1, \| \cdot \|_{P_0,B}) \left[1 + \tfrac{J_{[]}(2\delta, \mathcal{M}_\delta^1, \| \cdot \|_{P_0,B})}{4\delta^2 \sqrt{n}} \right].$$

Let $[\ell, u]$ be an ε-bracket in $\{p_\theta\}$ with respect to h. Since $u - \ell \ge 0$ and $e^{|x|} - 1 - |x| \le 2(e^{x/2} - 1)^2$ for every $x \ge 0$,

$$\begin{aligned} \left\| \log \tfrac{p_0 + u}{p_0 + p_{\theta_0}} - \log \tfrac{p_0 + \ell}{p_0 + p_{\theta_0}} \right\|_{P_0,B}^2 &\le 4 \int \left(\sqrt{\tfrac{p_0 + u}{p_0 + \ell}} - 1 \right)^2 p_0 \\ &\le 4 \int (\sqrt{p_0 + u} - \sqrt{p_0 + \ell})^2 \le 4h(u, \ell)^2 \le 4\varepsilon^2. \end{aligned}$$

Thus, $[\log \frac{p_0 + \ell}{p_0 + p_{\theta_0}}, \log \frac{p_0 + u}{p_0 + p_{\theta_0}}]$ makes a 2ε-bracket in $\mathcal{M}^1$. Hence, $N_{[]}(2\varepsilon, \mathcal{M}_\delta^1, \| \cdot \|_{P_0,B}) \le N_{[]}(\varepsilon, \mathcal{P}_\delta, h) \lesssim (\delta/\varepsilon)^r$ by Assumption 8. This induces $J_{[]}(2\delta, \mathcal{M}_\delta^1, \| \cdot \|_{P_0,B}) \lesssim \delta$. Therefore,

$$\mathbb{E}^* \sup_{h(\theta,\theta_0)<\delta} \left| \sqrt{n}(\mathbb{P}_n - P_0) \log \tfrac{D_{\theta_0}}{D_\theta} \right| \lesssim \delta + \tfrac{1}{\sqrt{n}}.$$

Similarly, $\mathbb{E}^* \sup_{\tilde{h}(\theta,\theta_0)<\delta} |\sqrt{m}(\tilde{\mathbb{P}}_m - \tilde{P}_0) \log \frac{1-D_{\theta_0}}{1-D_\theta}| \lesssim \delta + \frac{1}{\sqrt{m}}$. Then, the result follows by van der Vaart and Wellner (1996, Theorem 3.2.5). ∎

Lemma 3. *Under Assumption 9, for every $h \in \mathbb{R}^k$ and $h \to 0$,*

$$\begin{aligned} \int \left[\sqrt{\frac{p_{\theta_0} + p_{\theta_0 + h}}{2}} - \sqrt{p_{\theta_0}} - \frac{1}{4} h' \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right]^2 &= o(\|h\|^2), \\ \int \left[\sqrt{\frac{p_{\theta_0} + p_{\theta_0 + h}}{2}} - \sqrt{p_{\theta_0 + h}} + \frac{1}{4} h' \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0 + h}} \right]^2 &= o(\|h\|^2). \end{aligned}$$

Proof. Denote $p := p_{\theta_0}$ and $p_h := p_{\theta_0+h}$. For the first statement, it suffices to show

$$\int \left[\left(\sqrt{\tfrac{p+p_h}{2}} - \sqrt{p}\right) - \left(\tfrac{\sqrt{p_h}}{2} - \tfrac{\sqrt{p}}{2}\right)\right]^2 = \int \left(\sqrt{\tfrac{p+p_h}{2}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2}\right)^2 = o(\|h\|^2).$$

For every $\varepsilon > 0$, there exists $M > 1$ such that[28]

$$\int \left(\sqrt{\tfrac{p+p_h}{2}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2}\right)^2 \le \varepsilon + \int_{p_h/p \le M} \left(\sqrt{\tfrac{p+p_h}{2}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2}\right)^2.$$

By Taylor's theorem and concavity of the square root,

$$0 \le \sqrt{\tfrac{p+p_h}{2}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2} \le \sqrt{p} + \tfrac{p_h - p}{4\sqrt{p}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2} = \tfrac{1}{4}(\sqrt{p_h} - \sqrt{p})\left(\sqrt{\tfrac{p_h}{p}} - 1\right).$$

Thus, one obtains

$$\int_{p_h/p \le M} \left(\sqrt{\tfrac{p+p_h}{2}} - \tfrac{\sqrt{p_h}+\sqrt{p}}{2}\right)^2 \le \tfrac{1}{16} \int_{p_h/p \le M} (\sqrt{p_h} - \sqrt{p})^2 \left(\sqrt{\tfrac{p_h}{p}} - 1\right)^2.$$

For $p_h/p \le M$, $(\sqrt{p_h/p} - 1)^2$ is bounded by M, so the RHS is bounded by $\frac{1}{16} h' I_\theta h M = O(\|h\|^2 M)$. Moreover, $(\sqrt{p_h/p} - 1)^2$ converges to zero almost everywhere as p_h converges to p in DQM; therefore, by the dominated convergence theorem, the RHS is $o(\|h\|^2 M)$. By the diagonal argument, the original integral is $o(\|h\|^2)$.

For the second statement, we have shown $\int [(\sqrt{\frac{p+p_h}{2}} - \sqrt{p_h}) - (\frac{\sqrt{p}}{2} - \frac{\sqrt{p_h}}{2})]^2 = o(\|h\|^2)$, which, with Assumption 9, implies $\int [\sqrt{\frac{p+p_h}{2}} - \sqrt{p_h} - \frac{1}{4}(-h)'\dot{\ell}_{\theta_0}\sqrt{p_h}]^2 = o(\|h\|^2)$. This completes the proof. ■

The following lemma states local convergence of the objective function.

Lemma 4 (Asymptotic distribution of objective function)**.** *Under Assumptions 4 and 9, for every compact $K \subset \Theta$, uniformly in $h \in K$,*

$$\begin{aligned} & n\left[\mathbb{M}_{n,m}^{\theta_0+h/\sqrt{n}}(D_{\theta_0+h/\sqrt{n}}) - \mathbb{M}_{n,m}^{\theta_0}(D_{\theta_0})\right] \\ & \quad = -\sqrt{n}\mathbb{P}_n h'\dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^{\theta_0+h/\sqrt{n}}) D_{\theta_0+h/\sqrt{n}} h'\dot{\ell}_{\theta_0} \\ & \qquad + \sqrt{n}\frac{(\mathbb{P}_m^{\theta_0+h/\sqrt{n}} - P_{\theta_0+h/\sqrt{n}}) - (\mathbb{P}_m^{\theta_0} - P_{\theta_0})}{1/\sqrt{n}} \log(1 - D_{\theta_0}) + \frac{h'\tilde{I}_{\theta_0} h}{4} + o_P(1). \end{aligned}$$

With Assumptions 10 and 11, this reduces to

$$-\sqrt{n}\mathbb{P}_n h'\dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^{\theta_0}) D_{\theta_0} h'\dot{\ell}_{\theta_0} + \frac{h'\tilde{I}_{\theta_0} h}{4} + o_P(1).$$

[28] This M applies uniformly over every small h.

Proof. Let $\theta := \theta_0 + h/\sqrt{n}$, $W := \sqrt{D_\theta/D_{\theta_0}} - 1$, $\tilde{W} := \sqrt{p_{\theta_0}/p_\theta} - 1$. Observe that

$$n[\mathbb{M}^\theta(D_\theta) - \mathbb{M}^{\theta_0}(D_{\theta_0})] = n(\mathbb{P}_n + \mathbb{P}_m^\theta)\log\tfrac{D_\theta}{D_{\theta_0}} - n\mathbb{P}_m^\theta \log\tfrac{p_{\theta_0}}{p_\theta} + n(\mathbb{P}_m^\theta - \mathbb{P}_m^{\theta_0})\log(1 - D_{\theta_0}).$$

We examine each term separately. By Assumption 9,

$$\begin{aligned} n(P_\theta - P_{\theta_0})\log(1 - D_{\theta_0}) &= n\int(\sqrt{p_\theta} + \sqrt{p_{\theta_0}})(\sqrt{p_\theta} - \sqrt{p_{\theta_0}})\log(1 - D_{\theta_0}) \\ &= \int\Big(\sqrt{n}h'\dot{\ell}_{\theta_0} + \tfrac{h'\ddot{\ell}_{\theta_0}h}{2} + \tfrac{h'\dot{\ell}_{\theta_0}\dot{\ell}'_{\theta_0}h}{2}\Big)p_{\theta_0}\log(1 - D_{\theta_0}) + o(1). \end{aligned}$$

The first term is zero since $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \geq 0$ and $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) = 2\int D_{\theta_0}(\sqrt{p_\theta} - \sqrt{p_{\theta_0}})^2 + o(h(\theta,\theta_0)^2) + (P_\theta - P_{\theta_0})\log(1 - D_{\theta_0})$.[29] Therefore, $n(P_\theta - P_{\theta_0})\log(1 - D_{\theta_0}) = \frac{1}{2}P_{\theta_0}(h'\ddot{\ell}_{\theta_0}h + h'\dot{\ell}_{\theta_0}\dot{\ell}'_{\theta_0}h)\log(1 - D_{\theta_0}) + o(1)$. If Assumption 11 holds, then $n[(\mathbb{P}_m^\theta - \mathbb{P}_m^{\theta_0}) - (P_\theta - P_{\theta_0})]\log(1 - D_{\theta_0}) = o_P(1 + n/m)$.

Using $\log x = 2(\sqrt{x} - 1) - (\sqrt{x} - 1)^2 + (\sqrt{x} - 1)^2 R(\sqrt{x} - 1)$ for $R(x) = O(x)$,

$$n(\mathbb{P}_n + \mathbb{P}_m^\theta)\log\tfrac{D_\theta}{D_{\theta_0}} = 2n(\mathbb{P}_n + \mathbb{P}_m^\theta)W - n(\mathbb{P}_n + \mathbb{P}_m^\theta)W^2 + n(\mathbb{P}_n + \mathbb{P}_m^\theta)W^2 R(W_n).$$

Let $\breve{I}_{\theta_0} := 2P_{\theta_0}D_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}'_{\theta_0}$. Observe that

$$(P_0 + P_\theta)\Big(\sqrt{n}W + \tfrac{h'\dot{\ell}_{\theta_0}}{2}(1 - D_\theta)\Big)^2 = n\int\Big[\sqrt{p_0 + p_{\theta_0}} - \sqrt{p_0 + p_\theta} + \tfrac{h'\dot{\ell}_{\theta_0}}{2\sqrt{n}}\sqrt{(1 - D_\theta)p_\theta}\Big]^2,$$

which is $o(\|h\|^2/n)$ by Lemma 7 and Assumption 9. Thus, the RHS converges to zero uniformly over every compact $K \subset \Theta$. We draw two observations: (i) the mean and variance of $(\sqrt{n}W + (1 - D_\theta)h'\dot{\ell}_{\theta_0}/2)(X_i)$, $X_i \sim (P_0 + P_{\theta_n})/2$, converge to zero and so does the variance of $\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(\sqrt{n}W + (1 - D_\theta)h'\dot{\ell}_{\theta_0}/2)$ under Assumption 4;[30] (ii) $(P_0 + P_\theta)|nW^2 - (1 - D_\theta)^2(h'\dot{\ell}_{\theta_0}/2)^2| \to 0$, so $n(\mathbb{P}_n + \mathbb{P}_m^\theta)W^2 = (\mathbb{P}_n + \mathbb{P}_m^\theta)(1 - D_\theta)^2(h'\dot{\ell}_{\theta_0}/2)^2 + o_P(1) \to h'I_{\theta_0}h/4 - h'\breve{I}_{\theta_0}h/8$. Next,

$$\begin{aligned} n(P_0 + P_\theta)W &= -\tfrac{n}{2}h(p_0 + p_{\theta_0}, p_0 + p_\theta)^2 \longrightarrow -\tfrac{h'I_{\theta_0}h}{8} + \tfrac{h'\breve{I}_{\theta_0}h}{16}, \\ \sqrt{n}(P_0 + P_\theta)(1 - D_\theta)\tfrac{h'\dot{\ell}_{\theta_0}}{2} &= \sqrt{n}P_\theta\tfrac{h'\dot{\ell}_{\theta_0}}{2} = \sqrt{n}(P_\theta - P_{\theta_0})\tfrac{h'\dot{\ell}_{\theta_0}}{2} \to \tfrac{h'I_{\theta_0}h}{2}. \end{aligned}$$

This implies that the mean of $\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(\sqrt{n}W + (1 - D_\theta)h'\dot{\ell}_{\theta_0}/2)$ converges to

[29] The term $P_{\theta_0}h'\dot{\ell}_{\theta_0}\log(1 - D_{\theta_0})$ is the only term that is linear in $h = h(\theta,\theta_0)$, so if it is not zero, then $M_\theta(D_\theta) - M_{\theta_0}(D_{\theta_0}) \geq 0$ is violated.

[30] This does not imply that the mean of $\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(\sqrt{n}W + (1 - D_\theta)h'\dot{\ell}_{\theta_0}/2)$ converges to zero.

$3h'I_{\theta_0}h/8 + h'\breve{I}_{\theta_0}h/16$. Combining with (i), we find

$$n(\mathbb{P}_n + \mathbb{P}_m^\theta)W = -\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(1 - D_\theta)\tfrac{h'\dot{\ell}_{\theta_0}}{2} + \tfrac{3h'I_{\theta_0}h}{8} + \tfrac{h'\breve{I}_{\theta_0}h}{16} + o_P(1).$$

The remainer term $n(\mathbb{P}_n + \mathbb{P}_m^\theta)W^2R(W_n)$ vanishes by the same logic as van der Vaart (1998, Theorem 7.2).

Next, observe that $n\mathbb{P}_m^\theta \log \frac{p_{\theta_0}}{p_\theta} = 2n\mathbb{P}_m^\theta \tilde{W} - n\mathbb{P}_m^\theta \tilde{W}^2 + n\mathbb{P}_m^\theta \tilde{W}^2 R(\tilde{W})$ and

$$P_\theta\Big(\sqrt{n}\tilde{W} + \tfrac{h'\dot{\ell}_{\theta_0}}{2}\Big)^2 = n\int\Big[\sqrt{p_{\theta_0}} - \sqrt{p_\theta} + \tfrac{h'\dot{\ell}_\theta}{2\sqrt{n}}\sqrt{p_\theta}\Big]^2 = o\big(\tfrac{\|h\|^2}{n}\big).$$

Again, (i) the mean and variance of $(\sqrt{n}\tilde{W} + h'\dot{\ell}_{\theta_0}/2)(X_i)$, $X_i \sim P_\theta$, converge to zero and so does the variance of $\sqrt{n}\mathbb{P}_m^\theta(\sqrt{n}\tilde{W} + h'\dot{\ell}_{\theta_0}/2)$ under Assumption 4; (ii) $P_\theta|n\tilde{W}^2 - (h'\dot{\ell}_{\theta_0}/2)^2| \to 0$, so $n\mathbb{P}_m^\theta\tilde{W}^2 \to P_\theta(h'\dot{\ell}_{\theta_0}/2)^2 \to h'I_{\theta_0}h/4$. Next, $nP_\theta\tilde{W} = -nh(\theta,\theta_0)^2/2 \to -h'I_{\theta_0}h/8$ and $\sqrt{n}P_\theta h'\dot{\ell}_{\theta_0}/2 \to h'I_{\theta_0}h/2$. This implies that the mean of $\sqrt{n}\mathbb{P}_m^\theta(\sqrt{n}\tilde{W} + h'\dot{\ell}_{\theta_0}/2)$ converges to $3h'I_{\theta_0}h/8$. Thus, we find

$$n\mathbb{P}_m^\theta\tilde{W} = -\sqrt{n}\mathbb{P}_m^\theta\tfrac{h'\dot{\ell}_{\theta_0}}{2} + \tfrac{3h'I_{\theta_0}h}{8} + o_P(1).$$

Again, we may ignore the remainer term $n\mathbb{P}_m^\theta\tilde{W}^2R(\tilde{W})$. Altogether,

$$\begin{aligned} n[\mathbb{M}^\theta(D_\theta) - \mathbb{M}^{\theta_0}(D_{\theta_0})] = -\sqrt{n}\mathbb{P}_n h'\dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)D_\theta h'\dot{\ell}_{\theta_0} + \tfrac{h'\tilde{I}_{\theta_0}h}{4} \\ + n[(\mathbb{P}_m^\theta - \mathbb{P}_m^{\theta_0}) - (P_\theta - P_{\theta_0})]\log(1 - D_{\theta_0}) + o_P(1). \end{aligned}$$

For the second claim, it remains to show that with Assumption 10,

$$\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)D_\theta h'\dot{\ell}_{\theta_0} - \sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^{\theta_0})D_{\theta_0}h'\dot{\ell}_{\theta_0} = o_P(1).$$

Note that $(P_0 + P_\theta)D_\theta h'\dot{\ell}_{\theta_0} - (P_0 + P_{\theta_0})D_{\theta_0}h'\dot{\ell}_{\theta_0} = 0$. Write

$$\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(D_\theta - D_{\theta_0})h'\dot{\ell}_{\theta_0} + \sqrt{n}(\mathbb{P}_m^\theta - \mathbb{P}_m^{\theta_0})D_{\theta_0}h'\dot{\ell}_{\theta_0}.$$

Since $p/(p+x)$ is convex in $x \geq 0$ for $p > 0$, $D_{\theta_0}\frac{p_{\theta_0} - p_\theta}{p_0 + p_{\theta_0}} \leq D_\theta - D_{\theta_0} \leq D_\theta\frac{p_{\theta_0} - p_\theta}{p_0 + p_\theta}$ by Taylor's theorem. Therefore, by Assumption 9,

$$\begin{aligned} -(\mathbb{P}_n + \mathbb{P}_m^\theta)D_{\theta_0}(1 - D_{\theta_0})(h'\dot{\ell}_{\theta_0})^2 + o_P(1) \leq \sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(D_\theta - D_{\theta_0})h'\dot{\ell}_{\theta_0} \\ \leq -(\mathbb{P}_n + \mathbb{P}_m^\theta)D_\theta(1 - D_\theta)(h'\dot{\ell}_{\theta_0})^2 + o_P(1). \end{aligned}$$

Thus, $\sqrt{n}(\mathbb{P}_n + \mathbb{P}_m^\theta)(D_\theta - D_{\theta_0})h'\dot{\ell}_{\theta_0}$ converges to $-P_{\theta_0}D_{\theta_0}(h'\dot{\ell}_{\theta_0})^2 = -h'\breve{I}_{\theta_0}h/2$ in

probability. The second term converges to $h'\breve{I}_{\theta_0}h/2$ under Assumption 10. ■

Proof of Theorem 6. By Theorem 5 and Assumption 7, $\hat{\theta}$ is consistent and $\sqrt{n}(\hat{\theta} - \theta_0)$ is uniformly tight. Assumption 6 implies $\mathbb{M}^{\hat{\theta}}(D_{\hat{\theta}}) \leq \inf_{\theta \in O} \mathbb{M}^{\theta}(D_\theta) + o_P^*(n^{-1})$. Let $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P_0)$, $\mathbb{G}_m^{\theta_0} := \sqrt{m}(\mathbb{P}_m^{\theta_0} - P_{\theta_0})$, and $\mathbb{G}_{n,m}^{\theta_0} f := \mathbb{G}_n(1 - D_{\theta_0})f - \sqrt{n/m}\mathbb{G}_m^{\theta_0} D_{\theta_0} f$. With Assumptions 4 and 9 to 11, Lemma 4 implies that uniformly in $h \in K$ compact,

$$n\Big[\mathbb{M}^{\theta_0 + h/\sqrt{n}}(D_{\theta_0 + h/\sqrt{n}}) - \mathbb{M}^{\theta_0}(D_{\theta_0})\Big] = -h'\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0} + \tfrac{h'\tilde{I}_{\theta_0}h}{4} + o_P\Big(1 + \tfrac{n}{m}\Big).$$

In particular, this holds for both $\hat{h} := \sqrt{n}(\hat{\theta} - \theta_0)$ and $\breve{h} := 2\tilde{I}_{\theta_0}^{-1}\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0}$, so

$$\begin{aligned}
n\Big[\mathbb{M}^{\theta_0 + \hat{h}/\sqrt{n}}(D_{\theta_0 + \hat{h}/\sqrt{n}}) - \mathbb{M}^{\theta_0}(D_{\theta_0})\Big] &= -\hat{h}'\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0} + \tfrac{1}{4}\hat{h}'\tilde{I}_{\theta_0}\hat{h} + o_P^*\Big(1 + \tfrac{n}{m}\Big), \\
n\Big[\mathbb{M}^{\theta_0 + \breve{h}/\sqrt{n}}(D_{\theta_0 + \breve{h}/\sqrt{n}}) - \mathbb{M}^{\theta_0}(D_{\theta_0})\Big] &= -\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0}'\tilde{I}_{\theta_0}^{-1}\mathbb{G}_{n,m}\dot{\ell}_{\theta_0} + o_P\Big(1 + \tfrac{n}{m}\Big).
\end{aligned}$$

Since $\hat{h}$ minimizes $\mathbb{M}^{\theta}(D_\theta)$ up to $o_P^*(1/n)$, the LHS of the first equation is larger than that of the second up to $o_P^*(1)$. Subtracting the two,

$$\tfrac{1}{4}\Big(\hat{h} - 2\tilde{I}_{\theta_0}^{-1}\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0}\Big)'\tilde{I}_{\theta_0}\Big(\hat{h} - 2\tilde{I}_{\theta_0}^{-1}\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0}\Big) + o_P^*\Big(1 + \tfrac{n}{m}\Big) \leq 0.$$

Since $\tilde{I}_{\theta_0}$ is assumed positive definite, $\hat{h} - 2\tilde{I}_{\theta_0}^{-1}\mathbb{G}_{n,m}^{\theta_0}\dot{\ell}_{\theta_0} = o_P^*(\sqrt{1 + n/m})$, proving the first expression. Since $\mathbb{P}_n$ and $\mathbb{P}_m^{\theta_0}$ are independent, the asymptotic variance is

$$\begin{aligned}
&\tilde{I}_{\theta_0}^{-1} 4\Big[P_0(1 - D_{\theta_0})^2\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}' + \Big(\lim_{n\to\infty}\tfrac{n}{m}\Big)P_{\theta_0}D_{\theta_0}^2\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}'\Big]\tilde{I}_{\theta_0}^{-1} \\
&\qquad = \tilde{I}_{\theta_0}^{-1} 4\Big[P_{\theta_0}D_{\theta_0}(1 - D_{\theta_0})\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}' + \Big(\lim_{n\to\infty}\tfrac{n}{m}\Big)P_0 D_{\theta_0}(1 - D_{\theta_0})\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}'\Big]\tilde{I}_{\theta_0}^{-1}.
\end{aligned}$$

■

A.4 Supporting Lemmas

The next lemma allows us to bound the Bernstein "norm" of an arbitrary log likelihood ratio by the Hellinger distance without having to assume a bounded likelihood ratio. This is a major improvement from Ghosal et al. (2000, Lemma 8.7) in that the multiple of the Hellinger need not diverge as $h(p, p_0) \to 0$.

Lemma 5 (Bernstein "norm" of log likelihood ratio)**.** *For any pair of probability*

measures P and P_0 such that $P_0(p_0/p) < \infty$,

$$\begin{aligned}\left\|\frac{1}{2}\log\frac{p}{p_0}\right\|_{P_0,B}^2 &\le h(p,p_0)^2\Big[2+\inf_{c\ge 1} cP_0\Big(\frac{p_0}{p}\;\Big|\;\frac{p_0}{p}\ge\Big[1+\frac{1}{2c}\Big]^2\Big)\Big]\\ &\le 2h(p,p_0)^2\Big[1+P_0\Big(\frac{p_0}{p}\;\Big|\;\frac{p_0}{p}\ge\frac{25}{16}\Big)\Big],\end{aligned}$$

where $P_0(p_0/p \mid p_0/p \ge a) = 0$ if $P_0(p_0/p \ge a) = 0$.

Proof. Using $e^{|x|}-1-|x| \le (e^x-1)^2$ for $x \ge -\frac{1}{2}$ and $e^{|x|}-1-|x| < e^x - \frac{3}{2}$ for $x > \frac{1}{2}$,

$$\left\|\log\sqrt{\tfrac{p}{p_0}}\right\|_{P_0,B}^2 \le 2P_0\Big(\sqrt{\tfrac{p}{p_0}}-1\Big)^2\mathbb{1}\Big\{\tfrac{p}{p_0}\ge\tfrac{1}{e}\Big\} + 2P_0\Big(\sqrt{\tfrac{p_0}{p}}-\tfrac{3}{2}\Big)\mathbb{1}\Big\{\tfrac{p_0}{p}>e\Big\}.$$

The first term is bounded by $2h(p,p_0)^2$. For every $c \ge 1$,

$$\begin{aligned}P_0\Big(\sqrt{\tfrac{p_0}{p}}-\tfrac{3}{2}\Big)\mathbb{1}\Big\{\tfrac{p_0}{p}>e\Big\} &\le P_0\Big(\sqrt{\tfrac{p_0}{p}}-1-\tfrac{1}{2c}\Big)\mathbb{1}\Big\{\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big\}\\ &= P_0\Big(\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)-\tfrac{1}{2c}\Big].\end{aligned}$$

Using $x - \frac{1}{2c} \le \frac{c}{2}x^2$ for every x,

$$\begin{aligned}P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)-\tfrac{1}{2c} &\le \tfrac{c}{2}\Big[P_0\Big(\sqrt{\tfrac{p_0}{p}}-1\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\Big]^2\\ &\le \tfrac{c}{2}P_0\Big(\tfrac{p_0}{p}\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)P_0\Big(\Big[1-\sqrt{\tfrac{p}{p_0}}\Big]^2\;\Big|\;\sqrt{\tfrac{p_0}{p}}\ge 1+\tfrac{1}{2c}\Big)\end{aligned}$$

by the Cauchy-Schwarz inequality. Then the first inequality follows. For the second, let $c = 2$. ∎

Remark. Since the Bernstein "norm" dominates L_2-norm, we have $P_0(\frac{1}{2}\log\frac{p_0}{p})^2 \le \|\frac{1}{2}\log\frac{p_0}{p}\|_{P_0,B}^2$, which may be better than Ghosal et al. (2000, Lemma 8.6).

Remark. Similarly, we have

$$\begin{aligned}\left\|\frac{1}{2}\log\frac{D}{D_\theta}\right\|_{P_0,B}^2 &\le 2h_\theta(D,D_\theta)^2\left[1+P_0\left(\frac{D_\theta}{D}\;\middle|\;\frac{D_\theta}{D}\ge\frac{25}{16}\right)\right],\\ \left\|\frac{1}{2}\log\frac{1-D}{1-D_\theta}\right\|_{P_\theta,B}^2 &\le 2h_\theta(1-D,1-D_\theta)^2\left[1+P_\theta\left(\frac{1-D_\theta}{1-D}\;\middle|\;\frac{1-D_\theta}{1-D}\ge\frac{25}{16}\right)\right].\end{aligned}$$

Lemma 6 (Bernstein "norm" of log discriminator ratio)**.** *For every $\theta_1, \theta_2 \in \Theta$,*

$$\left\|\log\frac{D_{\theta_1}}{D_{\theta_2}}\right\|_{P_0,B}^2 \le 8h(\theta_1,\theta_2)^2, \qquad \left\|\log\frac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}\right\|_{\tilde{P}_0,B}^2 \le 8\tilde{h}(\theta_1,\theta_2)^2.$$

Proof. Since $e^{|x|} - 1 - |x| \le 2(e^{x/2} - 1)^2$ for $x \ge 0$,

$$\begin{aligned}\left\|\log \tfrac{D_{\theta_1}}{D_{\theta_2}}\right\|_{P_0,B}^2 &\le 4P_0\Big(\sqrt{\tfrac{D_{\theta_1}}{D_{\theta_2}}} - 1\Big)^2 \mathbb{1}\{D_{\theta_1} \ge D_{\theta_2}\} + 4P_0\Big(\sqrt{\tfrac{D_{\theta_2}}{D_{\theta_1}}} - 1\Big)^2 \mathbb{1}\{D_{\theta_1} < D_{\theta_2}\} \\ &\le 4P_0\Big(\sqrt{\tfrac{p_0+p_{\theta_2}}{p_0+p_{\theta_1}}} - 1\Big)^2 + 4P_0\Big(\sqrt{\tfrac{p_0+p_{\theta_1}}{p_0+p_{\theta_2}}} - 1\Big)^2 \\ &\le 8\int (\sqrt{p_0 + p_{\theta_1}} - \sqrt{p_0 + p_{\theta_2}})^2 \le 8\int (\sqrt{p_{\theta_1}} - \sqrt{p_{\theta_2}})^2 \le 8h(\theta_1, \theta_2)^2.\end{aligned}$$

Similarly,

$$\left\|\log \tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}\right\|_{\tilde{P}_0,B}^2 \le 4\tilde{P}_0\Big(\sqrt{\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}} - 1\Big)^2 + 4\tilde{P}_0\Big(\sqrt{\tfrac{(1-D_{\theta_2})\circ T_{\theta_2}}{(1-D_{\theta_1})\circ T_{\theta_1}}} - 1\Big)^2 \le 8\tilde{h}(\theta_1, \theta_2)^2$$

since

$$\begin{aligned}\tilde{P}_0\Big(\sqrt{\tfrac{(1-D_{\theta_1})\circ T_{\theta_1}}{(1-D_{\theta_2})\circ T_{\theta_2}}} - 1\Big)^2 &\le \tilde{P}_0\big(\tfrac{1}{\sqrt{(1-D_{\theta_2})\circ T_{\theta_2}}} - \tfrac{1}{\sqrt{(1-D_{\theta_1})\circ T_{\theta_1}}}\big)^2 \\ &\le \tilde{P}_0\Big(\sqrt{\tfrac{p_0}{p_{\theta_2}} \circ T_{\theta_2}} - \sqrt{\tfrac{p_0}{p_{\theta_1}} \circ T_{\theta_1}}\Big)^2 = \tilde{h}(\theta_1, \theta_2)^2.\end{aligned}$$

■

Lemma 7 (Hellinger distance of sums of densities)**.** *For arbitrary densities p, p_0, p_1,*

$$h(p + p_0, p + p_1)^2 = \int \frac{p_0}{p + p_0}(\sqrt{p_0} - \sqrt{p_1})^2 + o(h(p_0, p_1)^2),$$

where $p_0/(p + p_0) = 1$ if $p = p_0 = 0$.

Proof. Since $\sqrt{p + x^2}$ is convex in x, by Taylor's theorem,

$$\sqrt{p + p_0} \ge \sqrt{p + p_1} + \sqrt{\tfrac{p_1}{p+p_1}}(\sqrt{p_0} - \sqrt{p_1}),$$

where $p_1/(p + p_1)$ is defined to be 1 if $p = p_1 = 0$. If $p_0 \ge p_1$, therefore,

$$0 \le \sqrt{\tfrac{p_1}{p+p_1}}(\sqrt{p_0} - \sqrt{p_1}) \le \sqrt{p + p_0} - \sqrt{p + p_1} \le \sqrt{\tfrac{p_0}{p+p_0}}(\sqrt{p_0} - \sqrt{p_1}).$$

Thus, we get the following lower and upper bounds

$$\int \big[\tfrac{p_0}{p+p_0} \wedge \tfrac{p_1}{p+p_1}\big](\sqrt{p_0} - \sqrt{p_1})^2 \le h(p + p_0, p + p_1)^2 \le \int \big[\tfrac{p_0}{p+p_0} \vee \tfrac{p_1}{p+p_1}\big](\sqrt{p_0} - \sqrt{p_1})^2.$$

By the dominated convergence theorem follows the claim. ■

REFERENCES

Altonji, J. G. and L. M. Segal (1996): "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business & Economic Statistics*, 14, 353–366.

Anthony, M. and P. L. Bartlett (1999): *Neural Network Learning: Theoretical Foundations*, New York: Cambridge University Press.

Arjovsky, M., S. Chintala, and L. Bottou (2017): "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, International Convention Centre, Sydney, Australia: PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 214–223.

Athey, S., G. Imbens, J. Metzger, and E. Munro (2020): "Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations," ArXiv:1909.02210.

Bach, F. (2017): "Breaking the Curse of Dimensionality with Convex Neural Networks," *Journal of Machine Learning Research*, 18, 629–681.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019): "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks," *Journal of Machine Learning Research*, 20, 1–17.

Bartlett, P. L. and W. Maass (2003): "Vapnik-Chervonenkis Dimension of Neural Nets," in *The Handbook of Brain Theory and Neural Networks*, ed. by M. A. Arbib, Cambridge: MIT Press, 1188–1192, second ed.

Bauer, B. and M. Kohler (2019): "On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression," *Annals of Statistics*, 47, 2261–2285.

Bennett, A., N. Kallus, and T. Schnabel (2019): "Deep Generalized Method of Moments for Instrumental Variable Analysis," in *Advances in Neural Information Processing Systems 32*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Curran Associates, Inc., 3564–3574.

Bruins, M., J. A. Duffy, M. P. Keane, and A. A. Smith, Jr. (2018): "Generalized Indirect Inference for Discrete Choice Models," *Journal of Econometrics*, 205, 177–203.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, vol. 6B, chap. 76, 5549–5632.

Chen, X. and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

Chen, X. and H. White (1999): "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21, C1–C68.

Chernozhukov, V. and H. Hong (2004): "Likelihood Estimation and Inference in a Class of Nonregular Econometric Models," *Econometrica*, 72, 1445–1480.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

De Nardi, M., E. French, and J. B. Jones (2010): "Why Do the Elderly Save? The Role of Medical Expenses," *Journal of Political Economy*, 118, 39–75.

Farrell, M. H., T. Liang, and S. Misra (2019): "Deep Neural Networks for Estimation and Inference," ArXiv:1809.09953.

Fermanian, J.-D. and B. Salanié (2004): "A Nonparametric Simulated Maximum Likelihood Estimation Method," *Econometric Theory*, 20, 701–734.

Forneron, J.-J. and S. Ng (2018): "The ABC of Simulation Estimation with Auxiliary Statistics," *Journal of Econometrics*, 205, 112–139.

Frazier, D. T., T. Oka, and D. Zhu (2019): "Indirect Inference with a Non-Smooth Criterion Function," *Journal of Econometrics*, 212, 623–645.

Gallant, A. R. and G. Tauchen (1996): "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000): "Convergence Rates of Posterior Distributions," *Annals of Statistics*, 28, 500–531.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014): "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2672–2680.

Gouriéroux, C. and A. Monfort (1997): *Simulation-Based Econometric Methods*, Oxford; New York: Oxford University Press.

Gouriéroux, C., P. C. Phillips, and J. Yu (2010): "Indirect Inference for Dynamic Panel Models," *Journal of Econometrics*, 157, 68–77.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017): "Deep IV: A Flexible Approach for Counterfactual Prediction," in *Proceedings of the 34th International Conference on Machine Learning*, ed. by D. Precup and Y. W. Teh, International Convention Centre, Sydney, Australia: PMLR, vol. 70 of *Proceedings of Machine Learning Research*, 1414–1423.

Honoré, B. E. and L. Hu (2017): "Poor (Wo)man's Bootstrap," *Econometrica*, 85, 1277–1301.

Huber, P. J. (1967): "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L. M. Le Cam and J. Neyman, Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press, vol. 1, 221–233.

Jankowski, H. (2014): "Convergence of Linear Functionals of the Grenander Estimator under Misspecification," *Annals of Statistics*, 42, 625–653.

Kleijn, B. J. K. and A. W. van der Vaart (2006): "Misspecification in Infinite-Dimensional Bayesian Statistics," *Annals of Statistics*, 34, 837–877.

——— (2012): "The Bernstein-Von-Mises Theorem under Misspecification," *Electronic Journal of Statistics*, 6, 354–381.

Klein, R. W. and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Kopczuk, W. (2007): "Bequest and Tax Planning: Evidence from Estate Tax Returns," *Quarterly Journal of Economics*, 122, 1801–1854.

Kristensen, D. and Y. Shin (2012): "Estimation of Dynamic Models with Nonparametric Simulated Maximum Likelihood," *Journal of Econometrics*, 167, 76–94.

Kuan, C.-M. and H. White (1994): "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews*, 13, 1–91.

Lewis, G. and V. Syrgkanis (2018): "Adversarial Generalized Method of Moments," ArXiv:1803.07164.

Lockwood, L. M. (2018): "Incidental Bequests and the Choice to Self-Insure Late-Life Risks," *American Economic Review*, 108, 2513–50.

Mackey, L., V. Syrgkanis, and I. Zadik (2018): "Orthogonal Machine Learning: Power and Limitations," in *Proceedings of the 35th International Conference on Machine Learning*, ed. by J. Dy and A. Krause, Stockholmsmässan, Stockholm Sweden: PMLR, vol. 80 of *Proceedings of Machine Learning Research*, 3375–3383.

McGarry, K. (1999): "Inter Vivos Transfers and Intended Bequests," *Journal of Public Economics*, 73, 321–351.

Mhaskar, H. N. and T. Poggio (2020): "Function Approximation by Deep Networks," *Communications on Pure & Applied Analysis*, 19, 4085–4095.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Nickl, R. and B. M. Pötscher (2010): "Efficient Simulation-Based Minimum Distance Estimation and Indirect Inference," *Mathematical Methods of Statistics*, 19, 327–364.

Nowozin, S., B. Cseke, and R. Tomioka (2016): "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Curran Associates, Inc., 271–279.

Patilea, V. (2001): "Convex Models, MLE and Misspecification," *Annals of Statistics*, 29, 94–123.

Pollard, D. (1997): "Another Look at Differentiability in Quadratic Mean," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, New York: Springer, chap. 19, 305–314.

Schmidt-Hieber, J. (2020): "Nonparametric Regression using Deep Neural Networks with ReLU Activation Function," *Annals of Statistics*, forthcoming.

Strebulaev, I. A. and T. M. Whited (2011): "Dynamic Models and Structural Estimation in Corporate Finance," *Foundations and Trends® in Finance*, 6, 1–163.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Yarotsky, D. (2017): "Error Bounds for Approximations with Deep ReLU Networks," *Neural Networks*, 94, 103–114.

AN ADVERSARIAL APPROACH TO STRUCTURAL ESTIMATION
Online Appendix

Tetsuya Kaji[1], Elena Manresa[2], and Guillaume Pouliot[1]

[1]University of Chicago
[2]New York University

July 15, 2020

S.1 MONTE CARLO EXERCISE OF A ROY MODEL

We conduct simulation of a Roy model with two sectors and two periods. The Roy model encompasses two essential features of economic environments: comparative advantage and selection. It is often estimated with indirect inference as the likelihood is hard to characterize.

S.1.1 Design

We implement a simplified version of the Roy model with no covariates. There are two sectors in which individuals work for wages. The wage in period 1 is determined by

$$\log w_{i1} = \mu_{d(i1)} + \varepsilon_{id(i1)1},$$

where $d(i1) \in \{1, 2\}$ is the sector chosen by individual i in period $t = 1$, μ_1 and μ_2 are sector-specific mean wage, and $\varepsilon_{id(i1)1}$ is an individual and sector-specific shock distributed normally. The wage in period 2 is determined by

$$\log w_{i2} = \mu_{d(i2)} + \gamma_{d(i2)} \mathbb{1}\{d_{i1} = d(i2)\} + \varepsilon_{id(i2)2},$$

where d_{i2} is the sector chosen by i at $t = 2$, $\gamma_{d(i2)}$ is the returns to experience if i chooses the same sector, and ε_{i12} and ε_{i22} are the shock, possibly correlated with the previous shock.

In this model individuals make different choices because they have different comparative advantages in one sector versus the other. There are four different sources of heterogeneity: two idiosyncratic shocks in period 1 for two sectors and two idiosyncratic shocks in period 2 for two sectors.

Individuals choose location d_{i1} to maximize the present value of current and future wages. In period 1, an individual works in sector 1 if the following inequality holds

$$w_{i11} + \beta \mathbb{E}[w_{i2} \mid d_{i1} = 1] > w_{i2} + \beta \mathbb{E}[w_{i21} \mid d_{i1} = 2],$$

where β is a discount factor and $w_{i2} = \max\{w_{i12}, w_{i22}\}$, and w_{i1d} is the potential wage in period 1 and location d. Expectations are taken with respect to the idiosyncratic shock $(\varepsilon_{i12}, \varepsilon_{i22})$. Since ε_{i11} and ε_{i21} are normally distributed, the expectations have closed forms.

In period 2, an individual, conditional on their choice of sector in period 1, observes ε_{i12} and ε_{i22} and choose the sector based on the maximum wage.

Thus, the sector choice and wage for each period can be written as a function of the structural parameters $\theta = (\mu_1, \mu_2, \gamma_1, \gamma_2, \sigma_1, \sigma_2, \rho_t, \rho_s, \beta)$, where ρ_t is the correlation between period 1 and period 2 in both locations, and ρ_s is the correlation between locations.

As actual observations, we generate data for $n = 1{,}000$ individuals with the true parameter $\theta_0 = (1.8, 2, 0.5, 0, 1, 1, 0.9, 0.9, 0)$.

S.1.2 Estimation

We consider adversarial estimation using 1-hidden layer neural networks of increasing number of neurons (from 2 to 100). We follow the two-step iterative algorithm described in Section 4.3. More specifically: we initialize θ at some value and generate a fixed set of shocks. We pick $m = n$. After training the neural network, we hold fix the estimated weights and calculate the gradient of (1) for small changes of θ. Then, we update θ in the direction of the gradient and generate corresponding synthetic data using the same shocks.[1]

The NN have been specified using a sigmoid link function of in all its layers. In addition, we incorporate dropout of 10% of the nodes during training, and allow for early stopping. The NN are trained with the R keras package. In particular, using stochastic gradient descent and backpropagation. We fix the randomness of the stochastic gradient descent across iterations of the estimation algorithm.

We set $X_i = (w_{i1}, d_{i1}, w_{i2}, d_{i2})$, i.e. the vector of all outcomes. For each replication

[1] While gradient-based methods are not justified in this context, given the discrete nature of some of the outcomes, we did not encounter numerical problems following this strategy.

we use 5 different initial conditions. We define the estimate as the one that minimizes the loss across the 5 minimizations.

S.1.3 Results

Figure 2 contains 8 panels with the mean estimation of each parameter, across 1,000 Monte Carlo simulations. The x axis represents the number of nodes of the hidden layer, and in parenthesis the total number of parameters in the NN, from 2 to 100. The green line denotes the true value of the parameter. The different shades of grey indicate different quantiles of the Monte Carlo distribution.

For all sizes of the NN the estimator is essentially unbiased. However, for smaller NN the variability around the mean can be large. The variability decreases as the size of the NN grows, up until the point where the size of the NN is around 10. This exercise provides evidence that, in line with our theory, a more flexible discriminator delivers estimators with smaller variance. We attribute this finding to the ability of the NN to better approximate the infeasible discriminator, D^{θ}, which attains the Cramer Rao bound.

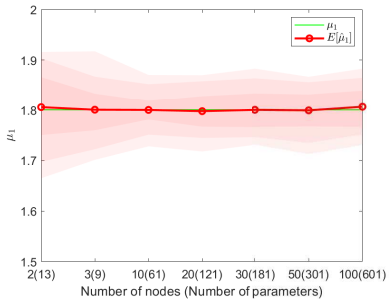
Worth noting is also the fact that for larger NN there seems to be limited increase in variance. This is likely due to the ability of the training algorithms to incorporate regularization through different strategies.

S.2 ADDITIONAL NOTES ON THE EMPIRICAL APPLICATION

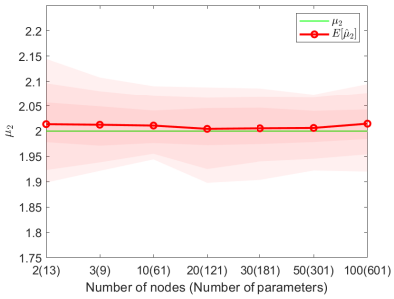
S.2.1 Details on Estimation Algorithm

Estimation of GAN in its original formulation (i.e. for training a generative model of images) is notoriously challenging (e.g., see Arjovsky and Bottou, 2017). Two main issues have been raised in the literature: (i) "mode-seeking behavior" of the discriminator due to imbalances between synthetic and actual sample sizes, and (ii) "flat or vanishing gradient" of the objective function in terms of the parameters of the generative model when synthetic and actual samples are easily distinguishable by the discriminator.

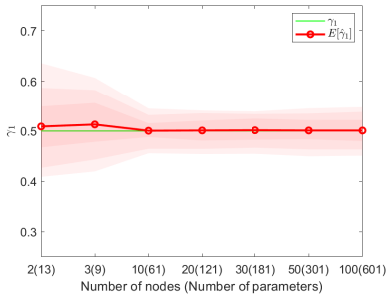
Imbalances in the sample size of synthetic versus actual data arise naturally in our context. Indeed, in order to reduce inflation of the variance of structural parameter estimates it is useful to choose $m >> n$. When this is the case, there is a risk that a good discriminator is one where it always predicts "synthetic", regardless of the input.



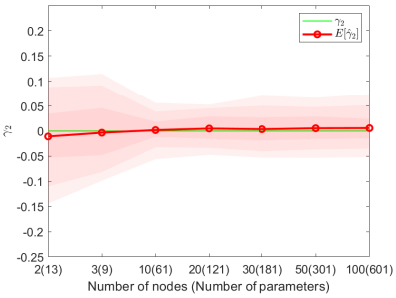
(a) μ_1



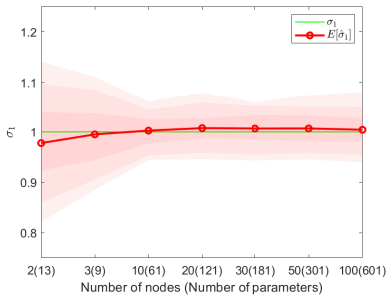
(b) μ_2



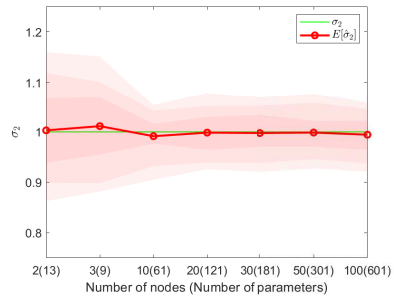
(c) γ_1



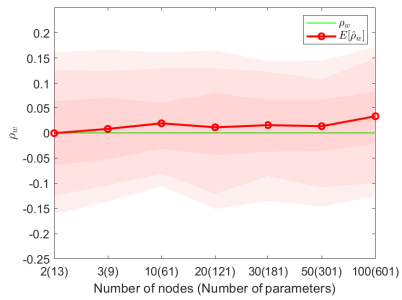
(d) γ_2



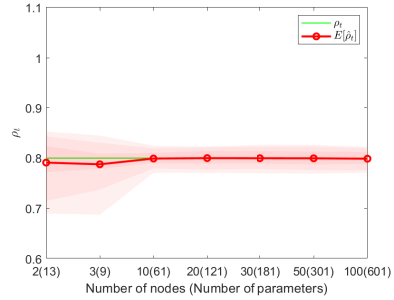
(e) σ_1



(f) σ_2



(g) ρ_w



(h) ρ_t

Figure 2: Results different NN

However, this is not a useful discriminator in our endeavor. We follow the literature recommendation in Machine Learning and mitigate this problem by performing data augmentation on the actual samples. In particular, we use a naive bootstrap strategy to resample with replacement histories of assets of individuals until both samples are even.

As per the flat gradient, we argue that this problem is not nearly as pervasive when the generative model is a typical structural economic model (provided the discriminator is parsimonious enough and is not overfitting). Indeed, Arjovsky and Bottou (2017) show that the problem of flat gradients is closely related to problems of overlapping support in typical generative models of images (see Lemma 1 and Theorem 2.1. in their paper), where the set of realizable images are measure zero in the space of all possible images. Typical economic models are very different from image generative models: (i) they tend to be embedded in low-dimensional spaces (the space of the endogenous outcomes), and (ii) they tend to be parametrized by low-dimensional vectors, where searching for configurations that provide overlapping support might be computationally feasible. Nonetheless, we could still encounter this problem, especially when outcomes are discrete.

In the context of our empirical application, outcomes are continuous and overlapping support is not a first order problem. Nonetheless, gradients of the structural parameters tend to be close to 0 when the conditional distribution of the outcomes generated by the model and the actual data are far apart, hence making naive gradient descent a very slow strategy. We implement two speeding strategies that have recently become popular in the context of training neural networks: NAG (Nesterov Acceleredated Gradient), an accelerated gradient descent method featuring momentum (Nesterov, 1983), and RPROP, an adaptive learning rate algorithm (Riedmiller and Braun, 1993).

Finally, we now give details on our choice of tuning parameters of the algorithm for training the discriminator. Recall we choose $\mathcal{D}$ the set of feedforward neural network with 2-hidden layers with 20 and 10 neurons, respectively, with sigmoid activation functions in both layers. We rely on state of the art estimation algorithms in the R Keras package for training the discriminator. In particular, we use the default ADAM optimization algorithm, which incorporates stochastic gradient descent, and backpropagation for fast computation of gradients. For implementation of stochastic gradient descent, we select a small batch size of 120 samples per gradient calculation,

and a large number of epochs (2000). As opposed to other implementations of GAN, we train the discriminator "to completion", and we fix the seed of the stochastic gradient to preserve non-randomness of the criterion as a function of structural parameters. We find this strategy to be the one that delivers the most reliable estimates, albeit at the cost of being computationally intensive. In order to avoid overfitting in the discriminator we make use of callback options that track the evolution of out of sample accuracy measures over epochs.

In S.2.3 below, we provide evidence that our estimation algorithm can successfully recover the true parameters in a Monte Carlo exercise tailored to the empirical application.

S.2.2 Details on Implementation of Poor (Wo)man's Bootstrap

We implement a "fast" bootstrap alternative proposed in Honoré and Hu (2017). Our estimates are based on 50 replications. For each replication we conduct 9 different univariate optimization problems.

S.2.3 Monte Carlo Excercise

In order to provide confidence on the results of the empirical application we conduct a simulation exercise in a design that mimics the DFJ model.

We simulate asset profiles conditional on the real distribution of health, PI, gender, etc. for $N = 2{,}688$ individuals according to the DFJ model and the following values of the structural parameters: $\beta = 0.971$, $\nu = 5.5$, $\underline{c} = 4{,}750$, MPC $= 0.23$, and $k = 13{,}797$. We then implement the adversarial estimation procedure as if this data is real against 250 independent sets of synthetic data.

For each set of shocks, we use 5 different starting values chosen randomly in a large neighborhood around the true values. For each set of initial values, we have 250 estimates with 250 synthetic data to calculate the mean and standard deviation of the estimator. The results can be found in Table 3.

The results reveal that the estimator is able to recover the true parameters with substantial precision. In particular, the lower quantile of the estimates for MPC at death is well separated from 0. This exercise gives us confidence on the ability of the method, as well as the optimization algorithm, to recover the true parameters of the model.

Table 3: MC tailored to the empirical application

	truth	mean	sd	[95% CI]	
ν	5.50	5.46	0.37	4.76	6.20
MPC	0.23	0.25	0.06	0.16	0.39
k [k$]	13.80	13.32	4.40	6.26	22.95

Notes: Mean and standard deviations computed over 250 Monte Carlo replications. ν is the parameter of risk aversion, MPC is the marginal propensity to consume at the moment of death, and k is the curvature of the bequest motive part of the utility function.

S.2.4 Autoencoder on X_2

The use of particular multilayer neural networks as sieve estimators for D^θ can achieve faster rates of convergence than other nonparametric methods. A necessary condition, as stated in Proposition 3 in the main text, is that $\log(p_0/p_\theta)$ admits the following hierarchical representation introduced in Bauer and Kohler (2019):

Definition (Generalized hierarchical interaction model)**.** Let $d \in \mathbb{N}_0$, with $d^* \in \{1, ..., d\}$ and $m : \mathbb{R}^d \to \mathbb{R}$. We say that m admits a generalized hierarchical interaction model of order d^* and level 0, if there exist $a_1, \ldots, a_{d^*} \in \mathbb{R}^d$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that

$$m(x) = f(a_1'x, \ldots, a_{d^*}'x).$$

for all $x \in \mathbb{R}^d$. We say that m satisfies a generalized hierarchical interaction model of order d^* and level $l+1$, if there exist $K \in \mathbb{N}_0$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$ and $f_{1k}, \ldots, f_{d^*k} : \mathbb{R}^d \to \mathbb{R}$ $(k = 1, \ldots, K)$ such that $f_{1k}, \ldots, f_{d^*k}$ $(k = 1, \ldots, K)$ satisfy a generalized hierarchical model of order l and

$$m(x) = \sum_{k=1}^{K} g_k(f_{1k}(x), \ldots, f_{d^*k}(x))$$

for all $x \in \mathbb{R}^d$.

As an example, $\log(p_0/p_\theta)$ satisfies a generalized hierarchical interaction model of order $d^* = 1$ and level 0 when p_θ corresponds to a conditional binary choice model, such as probit or logit, irrespectively of the dimension of the conditioning covariates.

We now provide an intuition on why fitting autoencoders on the inputs, X_i, can be informative of the hierarchical interaction order, d^*. We start by giving some background on autoencoders.

Autoencoders are used as dimension reduction statistical models, and have been referred to as the non-linear version of PCA (e.g. see Bishop (2006)). Autoencoders

are special neural networks that attempt to approximate the inputs, and they have three differentiated parts: encoder, bottleneck, and decoder. The encoder is typically a multilayer feedforward neural network with decreasing number of nodes in each layer. It forges a compressed representation of the inputs into the bottleneck, the hidden layer with the smallest number of nodes. The decoder takes the neurons from the bottleneck and maps it back to the output layer, increasing the number of nodes in each layer. The output layer has exactly as many nodes as the dimension of the input. Fitting an autoencoder involves minimizing the difference between the output layer and the inputs.

Let $X \in \mathbb{R}^d$ be a vector that can be perfectly fit into an autoencoder with $d^* < d$ neurons in the bottleneck. Let $X^* \in \mathbb{R}^{d_*}$ be the output of the neurons in the bottleneck. Hence, we have:

$$X^* = (en^1(\lambda_1' X), \ldots, en^{d*}(\lambda_{d^*}' X))$$

where $en^k : \mathbb{R}^d \to \mathbb{R}$, with $k \in \{1, \ldots, d^*\}$ are d univariate functions that map the inputs X through the encoder into the d^* neurons of the bottleneck. At the same time,

$$X = (de^1(X^*), \ldots, de^d(X^*))$$

where $de^k : \mathbb{R}^{d_*} \to \mathbb{R}$, with $k \in \{1, \ldots, d^*\}$, are d^* univariate functions that map the output of the bottleneck, X^*, into the d neurons in the output layer (which coincides with X) through the decoder. As a result, any function of X, $m(X)$, can be represented as a function g of d^* functions of X. Indeed,

$$m(X) = m(de^1(X^*), \ldots, de^d(X^*)) = g(X_1^*, \ldots, X_{d^*}^*) = g(en^1(\lambda_1' X), \ldots, en^{d*}(\lambda_{d^*}' X)).$$

Hence, m admits a representation as a generalized hierarchical interaction model of some level l (which depends on the exact shape on the autoencoder) and order d^*.

We fit autoencoders of increasing bottleneck dimension in a subset of 12 of the 21 variables in X_2 (excluding the constant) to investigate its underlying dimension, d^*. In particular, we select all binary variables: the gender indicator (1), the health status indicators over the 6 periods of observations (6), and alive/deceased indicators over the last 5 periods of observation (5).

The solid blue line in Figure 3 represents $MSE(d^*) = \|X - \tilde{X}(d^*)\|^2$, where $\tilde{X}(d^*)$ is the output layer of an autoencoder with bottleneck size d^*. The remaining

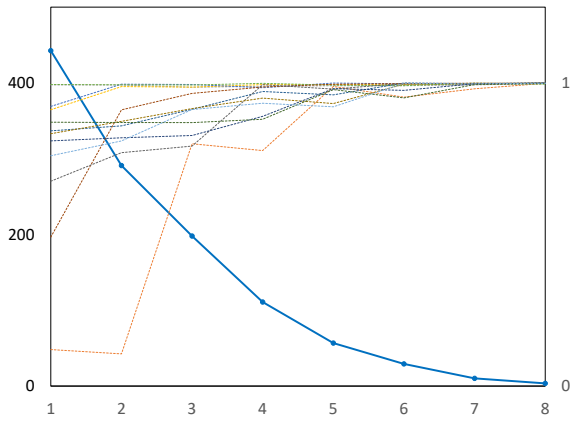


Figure 3: Fitting X_2 through autoencoders with increasing d^*. The blue solid line is the MSE as a function of d^*. The dashed lines represent the autocorrelation of each X_2^k with its prediction from the autoencoder. The left axis corresponds to MSE, the right axis corresponds to autocorrelation, and the x-axis corresponds to d^*.

12 dashed lines correspond to the correlation between the original variable X^k with the prediction from the autoencoder. When $d^* = 4$, MSE has significantly reduced, and the average autocorrelation among all variables is 94.5%.

S.2.5 Fit of the Model

Figure 4 shows the fit of the model in terms of mean asset profiles conditional on cohort and permanent income quintiles, excluding observations above 1% of the mean asset distribution of the actual data.[2] The fit of both DFJ and Adversarial are good, albeit they tend to do best in different parts of the distribution. Adversarial performs remarkably well for all cohorts for the bottom 3 permanent income quintiles. However, for the upper two permanent income quintiles, adversarial can overshoot, especially for the younger indivdiuals in the sample.

We also report the fit of the model separately for men and women in Cohort 2 in Figure 5. Matching the distribution conditional on gender is required in adversarial X_2, but not in DFJ. We can see that adversarial X_2 delivers a good fit for men even

[2]Mean assets are sensitive to small changes in the right hand side tail of the distribution. The trimming strategy for simulated observations under the adversarial estimates accounts for less than 1.75% of the observations, while it is less than 1.5% of observations for DFJ.

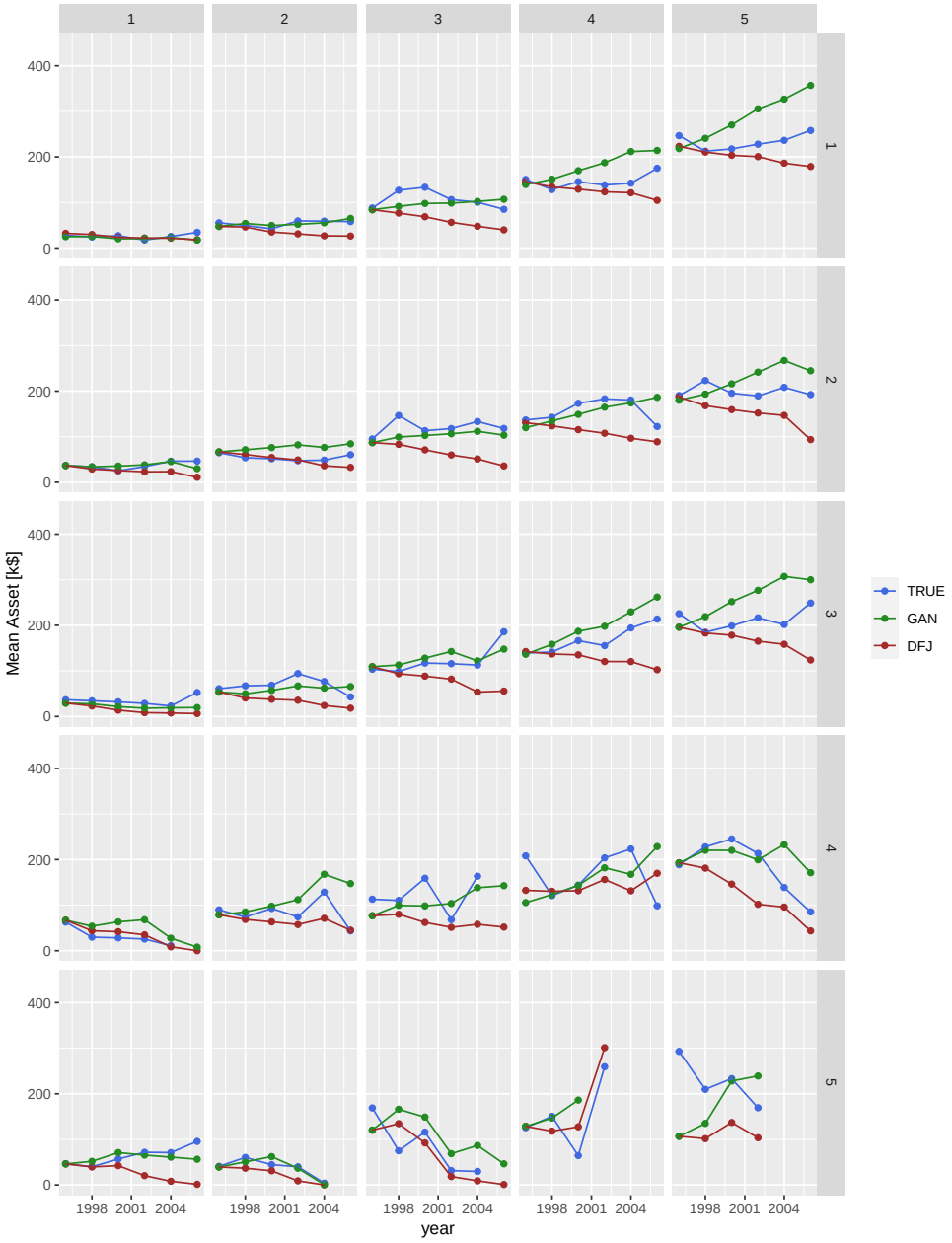


Figure 4: Fit in terms of mean assets by cohort (rows) and PIq (columns) over years. Red is DFJ, green is Adversarial X_2, and blue is actual data.

at the top of the distribution, while DFJ tens to underestimate men's assets often.

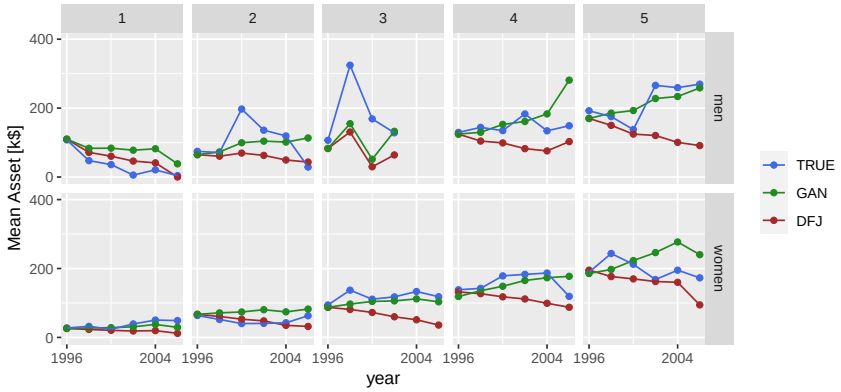


Figure 5: Fit in terms of mean assets in cohort 2 separately for men and women by PIq (columns) over years. Red is DFJ, green is Adversarial X_2, and blue is actual data. Other cohorts exhibit similar patterns.

S.3 EQUIVALENCE TO SMM WHEN D IS LOGISTIC

We start by discussing the statistical properties of the adversarial estimator when D is a logistic regression under high-level conditions, for any choice of $X_i = (1, \tilde{X}_i)$, where $\tilde{X}_i$ is the choice of the researcher.

The goal on this section is two-fold: first, to develop intuition on the properties of the estimator in a case where we can derive expressions analytically. Second, state the asymptotic equivalence result with a SMM estimator when moments are sample means of $\tilde{X}_i$ and optimally weighted. Hence, in this section we abstract from the conditions that ensure that the logistic regression is a regular M-estimator. In the next section, we will spell out all the formal conditions under which we analyze the adversarial framework.

Recall the FOC given in Example 2. Consistency of $\hat{\theta}$ can be established under standard regularity conditions on M-estimation.[3] For simplicity we assume X_i^θ is differentiable with respect to θ.

For any θ, let us define the following limiting discriminator parameter value

$$\lambda_0(\theta) = \arg\max_{\lambda} \mathbb{E}[\log(\Lambda(\lambda' X_i))] + \mathbb{E}[\log(1 - \Lambda(\lambda' X_i^\theta))].$$

We assume the following three high-level assumptions:

1. $\lambda_0(\theta) = 0$ if and only if $\theta = \theta_0$.

[3]For instance, Newey and McFadden (1994, Theorem 2.1).

2. $\sup_\theta \|\hat{\lambda}(\theta) - \lambda_0(\theta)\| = o_p(1)$.

3. $\sqrt{n}(\hat{\lambda}(\theta_0) - \lambda_0(\theta_0)) \rightsquigarrow N(0, \lim_{m,n\to\infty}[1 + \frac{n}{m}]\Omega_\lambda)$.

where $\Omega_\lambda = \mathbb{E}[X_i X_i']^{-1} \operatorname{Var}(X_i) \mathbb{E}[X_i X_i']^{-1}$.

The first condition can be interpreted as an identification assumption. The second condition is uniform consistency of the logit parameters over the space of θ. The third condition states that $\hat{\lambda}$ behaves asymptotically as a regular M-estimator.

Proposition S.1 (Asymptotic equivalence with SMM)**.** *Under Assumptions 1, 2, and 3, as* $n, m \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, \lim_{m,n\to\infty}\left[1 + \frac{n}{m}\right]V\right)$$

where

$$V = \left(\mathbb{E}\left[\frac{\partial X_i^{\theta_0}}{\partial \theta}\right]\mathbb{E}[X_i' X_i]^{-1}\mathbb{E}\left[\frac{\partial X_i^{\theta_0}}{\partial \theta}\right]'\right)^{-1}.$$

In addition,

$$\tilde{\theta} = \arg\min_\theta \left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i - \frac{1}{m}\sum_{i=1}^m \tilde{X}_i^\theta\right)' \Omega_W \left(\frac{1}{n}\sum_{i=1}^n \tilde{X}_i - \frac{1}{m}\sum_{i=1}^m \tilde{X}_i^\theta\right),$$

where Ω_W *is the optimal weighting matrix defined in Gouriéroux et al. (1993, Proposition 5) satisfies*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \rightsquigarrow N\left(0, \lim_{m,n\to\infty}\left[1 + \frac{n}{m}\right]V\right).$$

Proof. Using the properties of the sigmoid function, we have the following expansion

$$\hat{\theta} - \theta_0 = M(\theta^*)^{-1}\left(\hat{\lambda}(\theta_0)' \cdot \frac{1}{m}\sum_{i=1}^m \Lambda(\hat{\lambda}(\theta_0)' X_i^{\theta_0}) \cdot \frac{\partial X_i^{\theta_0}}{\partial \theta}\right)$$

where θ^* lies between $\hat{\theta}$ and θ_0, and

$$\begin{aligned} M(\theta) = \frac{\partial \hat{\lambda}(\theta)}{\partial \theta}\frac{1}{m}\sum_{i=1}^m \Lambda(\hat{\lambda}(\theta)' X_i^\theta)\frac{\partial X_i^\theta}{\partial \theta} + \hat{\lambda}(\theta)'\left(\frac{1}{m}\sum_{i=1}^m \Lambda(\hat{\lambda}(\theta)' X_i^\theta)\frac{\partial^2 X_i^\theta}{\partial \theta^2}\right) \\ + \hat{\lambda}(\theta)'\frac{1}{m}\sum_{i=1}^m \Lambda'(\hat{\lambda}(\theta)' X_i^\theta)\left[\frac{\partial \hat{\lambda}(\theta)}{\partial \theta} X_i^\theta + \hat{\lambda}(\theta)\frac{\partial X_i^\theta}{\partial \theta}\right]\frac{\partial X_i^\theta}{\partial \theta}. \end{aligned}$$

By consistency of $\hat{\theta}$ and conditions 1 and 2 above, we have $\hat{\lambda}(\theta^*) = o_p(1)$. In addition, substituting in the expression of $\frac{\partial \hat{\lambda}}{\partial \theta}$ obtained using the total derivative of the FOC of

the logit maximization (omitted here), we have

$$M(\theta^*) = A(\theta^*)R(\theta^*)^{-1}A(\theta^*) + o_p(1),$$

where

$$A(\theta) = \left(\frac{1}{m}\sum_{i=1}^{m}\Lambda(\hat{\lambda}(\theta)'X_i^\theta)\frac{\partial X_i^\theta}{\partial \theta}\right),$$
$$R(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\Lambda'(\hat{\lambda}(\theta)'X_i)X_i \cdot X_i' + \frac{1}{m}\sum_{i=1}^{m}\Lambda'(\hat{\lambda}(\theta)'X_i^\theta)X_i^\theta \cdot X_i^{\theta\prime}\right).$$

Using the block matrix inversion formula and $\frac{\partial X_i^\theta}{\partial \theta} = (0, \frac{\partial \tilde{X}_i^\theta}{\partial \theta}')'$, we see that, as $n/m \to 0$

$$A(\theta_0)'\Omega_\lambda A(\theta_0) = \frac{1}{2}M(\theta_0),$$

and hence

$$\sqrt{n}(\hat{\theta} - \theta_0) = M(\theta^*)^{-1}\sqrt{n}(\hat{\lambda}(\theta_0) - 0)A(\theta_0) \rightsquigarrow N\left(0, \lim_{m,n\to\infty}\left[1 + \frac{n}{m}\right]V\right).$$

We now move to show the second part of the proposition. Using the notation in Gouriéroux et al. (1993), we define

$$Q(\theta; \tau) = \frac{-1}{2n}\sum_{i=1}^{n}(\tilde{X}_i^\theta - \tau)^2$$

where τ is the auxiliary parameter. We have

$$\hat{\tau}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i^\theta.$$

Using the expression of the asymptotic distribution with the optimal weighting matrix in Gouriéroux et al. (1993, Propositions 4 and 5), we obtain the result. ■

Remark. When $n/m \to C$, there is inflation of the variance proportional to $1 + C$.

REFERENCES

Arjovsky, M. and L. Bottou (2017): "Towards Principled Methods for Training Generative Adversarial Networks," *arXiv preprint arXiv:1701.04862*.

Bauer, B. and M. Kohler (2019): "On Deep Learning as a Remedy for the

Curse of Dimensionality in Nonparametric Regression," *Annals of Statistics*, 47, 2261–2285.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*, Springer.

De Nardi, M., E. French, and J. B. Jones (2010): "Why Do the Elderly Save? The Role of Medical Expenses," *Journal of Political Economy*, 118, 39–75.

Gouriéroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118.

Honoré, B. E. and L. Hu (2017): "Poor (Wo)man's Bootstrap," *Econometrica*, 85, 1277–1301.

Nesterov, Y. (1983): "A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$," in *Sov. Math. Dokl*, vol. 27.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Amsterdam: North-Holland, vol. 4, chap. 36, 2111–2245.

Riedmiller, M. and H. Braun (1993): "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," in *IEEE international conference on neural networks*, IEEE, 586–591.