NBER WORKING PAPER SERIES

ROBUSTNESS CHECKS IN STRUCTURAL ANALYSIS

Sylvain Catherine
Mehran Ebrahimian
Mohammad Fereydounian
David Sraer
David Thesmar

Working Paper 30443
http://www.nber.org/papers/w30443

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2022, Revised October 2024

We thank participants at seminars and conferences for their comments. We are especially grateful to Hui Chen, Dan Green, Maryam Farboodi, Jonathan Parker, Demian Pouzo and Toni Whited, who provided important insights at various stages of this project. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Robustness Checks in Structural Analysis
Sylvain Catherine, Mehran Ebrahimian, Mohammad Fereydounian, David Sraer, and David Thesmar
NBER Working Paper No. 30443
September 2022, Revised October 2024
JEL No. C51, C52, G0

ABSTRACT

This paper introduces a computationally efficient methodology for estimating variants of structural models. Our approach approximates the relationship between moments and parameters, offering a low-cost alternative to traditional estimation methods. We establish general convergence conditions, primarily requiring model-based moments to be continuous functions of parameters. While this continuity does not necessitate a continuous economic model, it does require the model to have only sparse discontinuities, a concept we define. We also provide convergence rate bounds for Kernel and Neural Net approximations, with the latter demonstrating superior performance in higher dimensions.

We apply this methodology to two standard structural models: (1) dynamic corporate finance and (2) life-cycle portfolio choice. We demonstrate the reliability of our approach through simulations and then use it to explore identification, robustness to sample splits and moment selection, and model misspecification. These explorations are computationally infeasible with standard techniques, but become trivial with our methodology.

Sylvain Catherine
Wharton, Finance Department
Steinberg-Dietrich Hall
3620 Locust Walk,
Philadelphia, PA 19104
sylvain.sc.catherine@gmail.com

Mehran Ebrahimian
Department of Finance
Stockholm School of Economics
Sveavägen 65
Box 6501
SE-113 83 Stockholm
Sweden
mehran.ebrahimian@hhs.se

Mohammad Fereydounian
University of Pennsylvania
3401 Walnut Street, Room 407B
Philadelphia, PA 19104
mferey@seas.upenn.edu

David Sraer
Haas School of Business
University of California, Berkeley
545 Student Services Building
Berkeley, CA 94720-1900
and NBER
sraer@berkeley.edu

David Thesmar
MIT Sloan School of Management
100 Main Street, E62-632
Cambridge, MA 02142
and NBER
thesmar@mit.edu

A data appendix is available at http://www.nber.org/data-appendix/w30443

1 Introduction

Robustness checks are a standard feature of applied empirical work in economics and finance. After establishing their main results, researchers typically examine whether alternative mechanisms can explain their findings and provide additional analyses to control for such channels. For instance, they reestimate their main specification across various subsamples, either expecting results to be stable or to vary in a specific way. In other cases, they estimate different specifications to account for alternative channels, anticipating that their main predictions will hold in these amended specifications. These analyses (sample splits, changing the model) are part of the standard toolbox used by applied researchers.

Such robustness checks are rare in structural research, primarily due to computational constraints. As a first example, consider the case of sub-sample splits, which involve reestimating a model for particular periods or groups of observations. In reduced-form work, this analysis has negligible computational cost. However, in structural research, it can be prohibitively expensive, as each new estimation may take days and often requires human monitoring to fine-tune optimization. Second, consider robustness checks in reduced-form work that introduce additional control variables in regressions. Again, these checks are nearly costless. In simulation-based estimation, a related exercise might involve "purging" moments from control variables and estimating variants of the model fitted against such refined moments. Alternatively, the econometrician may want to evaluate parameter robustness to moment selection, varying the number and nature of moments against which the model is estimated. In many cases, the computational burden of reestimating the structural model implies that very few alternatives, if any, can be considered in practice. This paper overcomes this challenge by introducing a computationally efficient methodology for estimating variants of structural models. We provide convergence conditions and characterization, and present two applications using popular models.

Our approach works as follows. We consider the estimation of a structural model $\mathcal{S}$. Given deep parameters θ, the model generates moments $f(\theta)$. In most applications, such dynamic choice models, f is calculated numerically. An estimate θ^* of structural parameters is then obtained by minimizing a distance between simulated moments $f(\theta)$ and empirical moments m:

$$\theta^* = \arg\min_{\theta} (m - f(\theta))^\top W (m - f(\theta)),$$

where W is a weighting matrix. While simulating the economic model (i.e., numerically computing $f(\theta)$) for any given θ can be reasonably fast, the *estimation process* can be computationally costly as it requires a large number of such simulations, as well as human monitoring. This computational burden limits the number of estimations that are feasible for a given research project.

We propose to reduce this computational cost by building an *approximation* of f. Importantly, we do not approximate the numerical solution of the *model* itself, as explored in various works (e.g., Fernandez-Villaverde et al. (2021b), Fernández-Villaverde et al. (2021a), Duarte (2020)). Instead, we directly approximate the *moment function* f.

Our methodology follows three main steps. First, we generate a *training dataset* $\mathcal{D}_n$, comprising n (vectors of) parameters values and their corresponding moment values. This training dataset is fixed once and for all. This step is computationally intensive since n can be large. However, it is *not* more costly than a single model estimation, which also requires a large number of simulations. In the second step, we fit the approximate moment function f_n. Consistent with findings in the literature, we find that Neural Nets perform particularly well, as they mitigate the curse of dimensionality in θ. This step is computationally inexpensive. Finally, in the third step, we estimate parameters by matching moments using the approximate function f_n instead of the true function f:

$$\hat{\theta}_n = \arg\min_{\theta} (m - f_n(\theta))^\top W (m - f_n(\theta)).$$

This step incurs almost no computational cost since f_n is already known.

In section 3, we examine the general applicability of our method. First, we provide conditions under which the moment function f and its approximation f_n lead to the convergence of the estimate $\hat{\theta}_n \to \theta^*$ as the training data size $n \to \infty$. A key condition is the continuity of moments f with respect to θ. Under slightly stronger regularity conditions near the true estimate, we derive a formula for the *approximation error*:

$$\hat{\theta}_n - \theta^* \approx \underbrace{\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)\right)^{-1} \nabla f_n(\hat{\theta}_n)^\top W}_{\equiv \Lambda_n} \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right),$$

where Λ_n is the "approximate sensitivity matrix" (as in Andrews et al. (2017)) of parameters to moments, and $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$ is the moment approximation error. Once the model has been estimated using the approximation, this formula incurs minimal

computational cost.

The choice of the parametric function f_n is crucial for achieving a good approximation. We provide convergence speed results for kernel smoothers and neural nets. The kernel smoother converges at least as fast as $n^{-1/K}$, where K is the dimension of θ. Using classic results from the literature, we find that neural nets converge at least as fast as $\left(\frac{\log n}{n}\right)^{1/4}$. For complex models (where K is large), neural nets converge faster. This intuition is well-documented in the machine-learning literature: Neural nets can "mitigate the curse of dimensionality" by increasing in complexity as the training sample size grows (see e.g. Barron (1994), and more recently Farrell et al. (2021)).

While the continuity of moments f is necessary for convergence, we show in Section 4 that this does not imply the model itself must be continuous. Instead, the model must satisfy a weaker condition, which we call "light discontinuity" in the parameters θ. Intuitively, light discontinuity means that the functions used to compute moments have their discontinuities on a countable number of curves, parameterized by θ. We prove that all models satisfying this condition have continuous moments. We also discuss why standard economic models with non-convexities typically meet this criterion.

We apply our methodology to two standard structural models: (1) a dynamic corporate finance model similar to Hennessy and Whited (2007a) (Section 5.1), and (2) a life-cycle consumption and portfolio choice model, similar to Viceira (2001) and Cocco et al. (2005). In both case, we use neural networks to fit the approximate moment function f_n, since they exhibit faster convergence in high-dimensional settings.

We first assess the validity of the approximate parameter estimates by drawing a "validation" sample of parameters and corresponding moments. Using these validation moments, we estimate the model with the approximation f_n and compare the resulting parameters to the true parameters that generated the validation sample. In both applications, the correlation between the true parameters and our "approximate" estimates in the validation sample is always larger than .95, and mostly larger than .99. While our approach is orders of magnitude faster than SMM (about 1 second compared to an hour), the difference in estimates is small, if not negligible. We also use this validation sample to verify that the approximation error formula accurately measures the true estimation error.

We then use f_n to conduct various robustness checks. First, we examine the sensitivity of parameter estimates to the choice of targeted moments. In structural work,

discussions of identification typically focus on the function f near the SMM estimate θ^* (model comparative statics). However, as noted by Andrews et al. (2017), a more effective diagnostic tool is the assessment of how parameter estimates *vary* with empirical moments, since it reveals how changes in data affect estimates. This approach is rarely employed due to the computational expense of reestimating the model while slowly adjusting empirical moments. Our method allows for this analysis at nearly no cost. In our applications, representing parameters as functions of moments offers a different, and often more accurate, diagnostic for identification.

Second, we evaluate the robustness of parameter estimates to the *selection* of targeted moments. For any structural model, the number of moments one can use to estimate the model is potentially large. If the model correctly represents the data-generating process, the specific moments targeted in estimation do not matter, as long as they identify all parameters. However, in practice, models are always misspecified, so that moment selection matters. To evaluate robustness to moment selection, researchers usually calculate additional moments in the model not used in estimation and compare them to their empirical values. This approach is not ideal, as the choice of baseline moments is arbitrary. Moreover, if the estimated model fails to match these non-targeted moments, it is unclear whether and how matching these moments would affect parameter estimates. A better approach is to reestimate the model across many sets of possible moments. While this approach would be computationally expensive using traditional methods, this becomes feasible using our approximation f_n. In our two applications, we explore thousands of possible moment combinations, and report the resulting distribution of parameter estimates. We find that few estimates are robust to moment selection (especially in the corporate finance model), and isolate the moments that significantly impact estimation.

Third, we test the robustness of estimates across different subsamples. In reduced-form analyses, econometricians often re-estimate regression models on various subsamples to check whether estimates are stable or change in expected ways. In structural work, the equivalent exercise would require reestimating the model for each subsample, which becomes computationally prohibitive if there are many sample splits. However, the low computational cost of our approach makes these checks straightforward. As demonstrated in our corporate finance application, structural parameters exhibit considerable variation over time, providing valuable insights into the model's validity.

Finally, our methodology allows us to assess model misspecification. Specifically, we examine how inference is affected when the data is generated using alternative

models, i.e., not the model used in the baseline estimation. We first simulate many datasets using alternative models with different specifications and a range of parameter values. For each of these datasets, we then reestimate structural parameters using the baseline model. Despite the many structural estimations required, this approach is feasible thanks to our approximation method. More formally, we specify an alternative model with structural parameters θ_a and moments $g_a(\theta_a)$. For each θ_a, we estimate the parameters of the baseline model that best fit the alternative model-generated moments by solving:

$$\hat{\theta}_n(\theta_a, g_a) = \arg\min_{\theta} \left(g_a(\theta_a) - f_n(\theta)\right)^\top W \left(g_a(\theta_a) - f_n(\theta)\right),$$

allowing us to estimate parameters for a wide range of θ_a values and alternative models g_a. For θ_a in the neighborhood of θ^* and g_a in the neighborhood of f, this method aligns with the diagnostic proposed by Andrews et al. (2017), which does not require specifying g_a. Our approach enables a broader exploration of potential misspecification, although it necessitates specifying the nature of misspecification (i.e., to take a stand about the alternative model g_a). We apply this approach to our corporate finance model, considering alternative models where financing may be constrained by cash-flow expectations (Lian and Ma, 2021). We find that, when the baseline model omits these cash-flow constraints, the cost of financing constraints is overestimated by a factor of 2 or 3.

The paper starts with a short literature review (section 2), and then moves on to establish the convergence results (section 3). Section 4 characterizes economic models that satisfy the key condition required in our methodology (moments are continuous functions of structural parameters). Section 5 explores two applications and Section 6 concludes.

2 Related literature

Our paper primarily contributes to recent literature aimed at increasing transparency in structural estimation, with a particular focus on the sensitivity of policy predictions to moment or model misspecification. Andrews et al. (2020b) offers a formal definition of transparency in empirical research and applies it to structural estimation in economics. Andrews et al. (2017) derives a local linear approximation of the relationship between parameter estimates and data moments, providing a diagnostic tool for assessing misspecification bias. This method does not require explicitly specifying the

misspecification but assumes that misspecification bias is small. While our analysis of misspecification is global, it does require to specify alternative models to the baseline model. In a follow-up work, Andrews et al. (2020a) formalizes the link between descriptive analysis and structural estimation using a similar local approximation. Our approach facilitates exploring additional descriptive statistics at low cost (non-targeted moments, model outputs). More broadly, our work connects to the literature on robustness to model misspecification (e.g., Huber (2011); Armstrong and Kolesár (2021); Bonhomme and Weidner (2018)).

In the structural literature, discussions of identification typically revolve around the relationship between moments and structural parameters – the function $f(\theta)$. Table D.1 reviews recent structural papers in corporate and household finance. Of the 43 papers surveyed, 12 show local comparative statics (e.g., plots of f around θ^*), 8 report the Jacobian matrix (derivatives of f around θ^*), and 24 omit both. Four recent papers reports the sensitivity matrix of Andrews et al. (2017), i.e. the local derivative of parameter estimates w.r.t. targeted moments. Our approach eliminates the need for linear approximations, which, as our examples show, are not always valid.

Our paper is also connected to emerging work that seeks to improve numerical solutions for *models* through approximations. For instance, Norets (2012) uses neural networks to approximate the solution of a finite-horizon, dynamic discrete choice model. Similarly, Duarte (2020) describes a new solution method combining ML algorithms and Gradient Descent Algorithm. Chen et al. (2021) and Fonseca et al. (2022) use predictive algorithms to estimate the solution of models, after these are numerically solved on a training sample.[1] Our focus differs significantly from these papers as we aim to approximate *moments* as a function of parameters, rather than approximating *the value or policy functions* of the underlying model. Our approach is driven by our interest in estimation and robustness analysis, rather than solving the model.

Finally, our paper contributes to the vast literature that structurally estimate dynamic models of corporate and household finance (see Strebulaev and Whited (2012) for a survey of the corporate finance literature, and Gomes et al. (2021) for household finance). We assess the robustness of these widely-used models.

[1]For further references, see also Fernández-Villaverde et al. (2021a), Villa and Valaitis (2019), Maliar et al. (2019), Azinovic et al. (2019).

3 Our approach: Presentation and general results

This section outlines our approach and the conditions under which it applies. We begin by showing convergence, then provide a formula to compute the approximation error, and conclude with results on convergence speed. The focus here is on general results that depend solely on the properties of the *moments* of the estimated model, rather than its *solutions*. In Section 4, we discuss conditions on model solutions, with a particular focus on dynamic choice models, which we use as our main applications.

3.1 Description and notations

Let $\mathcal{S}$ be a structural model with deep parameters $\theta \in \mathbb{R}^K$, generating a vector of moments $f(\theta) \in \mathbb{R}^M$, where M is the number of computable moments. These moments do not need to be all empirically observable. Some can be a combination of structural parameters that capture an outcome of interest (e.g., the average loss in market value due to financial constraints in a structural corporate finance model).

In simulation-based estimation, $f(\cdot)$ does not admit a closed-form representation and is computed numerically. For clarity, we assume that $f(\cdot)$ can be exactly computed via extensive simulations, ignoring any simulation error. Let m represent the vector of empirical counterparts to the model-based moments $f(\theta)$. The minimum distance estimator of θ is obtained by the following minimization:

$$\theta^* \in \arg\min_{\theta} \left(m - f(\theta)\right)^\top W \left(m - f(\theta)\right), \tag{1}$$

where W is a weighting matrix, which assigns zero weights to moments that cannot be observed in the data. We call θ^* the *true* parameter estimate, i.e., the solution to Equation (1).

Estimating the model for a specific set of moment values, m, is computationally expensive. When the model lacks a closed-form solution, it must be solved numerically and moments have to be generated through simulation. This process is time-consuming, especially for dynamic and non-linear models. Additionally, optimization algorithms must compute moments $f(\theta)$ across many parameter values to find θ^*, the global minimizer in Equation (1). This issue is exacerbated by the curse of dimensionality, as the parameter space expands exponentially with K.

Due to these limitations, researchers often perform only a few estimations. Fundamentally, the issue that plagues this approach is that it lacks "economies of scale": each new estimation on a different set of empirical moments (whether on moments

estimated for different subsamples or a different set of moments altogether) incurs the same computational cost as the initial estimation.

Our approach creates such economies of scope using an approximation of the function $f(\cdot)$. To achieve this, we propose the following process:

1. **Parameter Bounds:** Define ex ante bounds for each parameter in θ. These bounds also need to be specified in standard estimation techniques. They can be based on expert knowledge or prior findings in the literature. Let $\mathbf{P}_\theta \subset \mathbb{R}^K$ be the set of admissible parameter vectors.

2. **Drawing Parameters:** We select n parameter values $(\theta_i)_{1\leq i\leq n}$ from this set $\mathbf{P}_\theta$. The exact method to build this estimate is not critical. One option is to use a regularly spaced grid. In our applications, we use quasi-random Halton sequences, which maintain uniform density as new points are added. Thus, to increase the size of the training sample from n to n', the researcher only needs to draw $n' - n$ parameter values while retaining the original n draws.

3. **Model Simulations:** For each parameter values θ_i, we solve the model and simulate moments $f(\theta_i)$. This step creates a *training dataset* $\mathcal{D}_n$, consisting of n parameters and their corresponding moments under the model $\mathcal{S}$. Although computationally intensive, this step is performed only once.

4. **Fitting the Approximation:** We use the training data $\mathcal{D}_n$ to find a parametric approximation $f_n(\theta)$ of $f(\theta)$, which can be computed for any θ. We refer to f_n as the *approximate moment function*. While no specific parametrization is required, we focus on neural networks, as they converge faster than kernel methods and yield more accurate results in our applications.

5. **Estimation** Estimate $\hat{\theta}_n$ as the solution of:

$$\hat{\theta}_n \in \underset{\theta\in\mathbf{P}_\theta}{\arg\min}\, (m - f_n(\theta))^\top W (m - f_n(\theta)). \tag{2}$$

   where $\hat{\theta}_n$ is the *approximate* parameter estimate.

Only step 3, building the training dataset $\mathcal{D}_n$, is computationally expensive. However, the size of the training dataset is typically chosen to match the number of model simulations required for a single estimation using SMM. Therefore, this step is no more costly than a standard estimation.

3.2 Convergence

Let m be a vector of empirical moments. To simplify the exposition, we assume that m is measured without statistical error, since this aspect of the inference problem is already covered by well-known formulas (see e.g., Gouriéroux and Montfort (1996)). We make the following assumptions:

Assumption 1. *Suppose that the following hold:*

i. $\mathbf{P}_\theta$ *is nonempty and compact.*

ii. W *is positive-definite.*

iii. There exists a unique $\theta^* \in \mathbf{P}_\theta$ *such that* $m = f(\theta^*)$.

iv. The model $\mathcal{S}$ *generates a continuous moment function* $f(\cdot)$.

The first two assumptions are standard and technical. The third assumption is stronger, as it requires that the model is (1) well-identified and (2) fits the data, though we do not assume that the model is the true data-generating process. The fourth assumption is crucial: The moments need to be *continuous* functions of parameters. This continuity depends on the specific model and moments used. In Section 4, we will derive sufficient conditions on the underlying economic model so that this assumption is verified.

The following theorem formalizes the conditions under which the approximate parameter estimate (the minimizer of program (2)) converges to the true parameter estimate (the minimizer of program (1)):

Theorem 1 (Convergence of approximate parameter estimate)**.** *Suppose Assumption 1 holds and the approximate moment function* f_n *satisfies:*

i. f_n *uniformly converges to* f *as* $n \to \infty$.

ii. f_n *is continuous.*

Then, the approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$ *as* $n \to \infty$.

Proof. See Appendix A.3. □

The conditions (i) and (ii) in Theorem 1, requiring that f_n converges uniformly and is continuous, can be ensured the choice of the approximation f_n.

3.3 Asymptotic approximation error formula

When n is large but finite, our approximate estimator differs from the actual estimator since $f_n \neq f$. This difference can be computed analytically provided n is sufficiently large, i.e., when $\hat{\theta}_n - \theta^*$ is small enough:

Proposition 1. *Suppose the assumptions in Theorem 1 hold. Additionally, assume there exist N_0, L_0, and a neighborhood $\mathbf{O} \subseteq \mathbf{P}_\theta$ around θ^* such that $\mathbf{O}$ is open in $\mathbb{R}^K$ and for $n \geq N_0$, ∇f_n and ∇f exist and are continuous on $\mathbf{O}$, and the sensitivity matrix Λ_n, defined below, exists with $\|\Lambda_n\|_2 \leq L_0$. Then, as $n \to \infty$, we have:*

$$\hat{\theta}_n - \theta^* = \underbrace{\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)\right)^{-1} \nabla f_n(\hat{\theta}_n)^\top W}_{\equiv \Lambda_n} \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right) + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right). \quad (3)$$

Proof. See Appendix A.4. □

A key additional assumption in Proposition 1 is that f is locally differentiable with a continuous derivative around the true parameter. The assumption that f_n is continuously differentiable is easier to satisfy if f_n is constructed appropriately (e.g., using a neural network).

Another crucial assumption in Proposition 1 is the existence and boundedness of the matrix Λ_n. Intuitively, Λ_n is similar to the "sensitivity matrix" in Andrews et al. (2020b): it represents the local derivative of parameters with respect to moments. The existence of Λ_n requires that the model is identified (i.e., ∇f is full-rank), along with some regularity conditions (since the matrix inversion also involves ∇f_n). For simplicity of exposition, we directly require the boundedness of $\|\Lambda_n\|_2$ rather than imposing more primitive conditions. Finally, the second term on the right side of Equation (3), $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$, measures the approximation error on the moments themselves.

The formula introduced in (3) is computationally inexpensive to obtain. While the values of $f(\hat{\theta}_n)$ and its gradient $\nabla f(\hat{\theta}_n)$ must be computed numerically, this requires only a small number of evaluations. The functions f_n and their gradients at the estimate are simple enough to often have closed-form expression.

In our applications below, we demonstrate through simulations that the formula (3) provides an accurate approximation of the true error.

3.4 Two parameterizations for f_n and their speeds of convergence

We now turn to the construction of f_n. We explore two parameterizations that ensure that the conditions in Theorem 1 are satisfied (in particular, that $f_n \to f$ uniformly), and for which we can produce results on convergence speed.

3.4.1 Granularity of the training data

We first introduce a key concept, the *granularity* of the training data. For the parameters in the training sample, $(\theta_i)_{1 \leq i \leq n}$, we define:

$$\delta_n = \max_{\theta \in \mathbf{P}_\theta} \min_{1 \leq i \leq n} \|\theta - \theta_i\|_2, \tag{4}$$

which represents the maximum distance from any point in $\mathbf{P}_\theta$ to its nearest element in the training dataset $\mathcal{D}_n$. Our asymptotic results assume that as n increases, $\delta_n \to 0$: each point on $\mathbf{P}_\theta$ becomes closer and closer to the elements of the training dataset, i.e., the training data become infinitely granular.

How fast does δ_n decrease with n, the size of the training dataset $\mathcal{D}_n$? The answer depends on how the elements of $\mathcal{D}_n$ are structured.

To build intuition, consider the simple case of a regular grid over $\mathbf{P}_\theta$. Suppose the set of possible parameter values is a cube of dimension K, and that $\mathcal{D}_n$ consists of regularly spaced points in this cube. Then, it is easy to see that:

$$\delta_n = \mathcal{O}\left(n^{-\frac{1}{K}}\right).$$

Thus, δ_n decreases with n, but more slowly if parameters have a large dimension K – the standard curse of dimensionality.[2]

In our application, we use a uniform draw over $\mathbf{P}_\theta$ via a Halton sequence instead of a regular grid. The advantage of this method is that granularity increases uniformly by simply adding more points to the existing sample, unlike a regular grid that would require redrawing the entire sample to increase n, which is computationally costly.

With uniform draws, the upper bound on convergence is the same as with a regular grid, although only in probability, as shown in the following lemma:

[2]To see this, assume edges of $\mathbf{P}_\theta$ have length E_k and each edge has q regularly spaced points. Then, the number of points in the training sample is $n = q^K$. The maximum distance between the grid and any point of $\mathbf{P}_\theta$ is $\delta_n = \frac{\sqrt{\sum_k E_k^2}}{2q}$. Thus, $\delta_n \propto n^{-\frac{1}{K}}$.

Lemma 1. *Suppose* $\mathbf{P}_\theta = [a_1, b_1] \times \ldots \times [a_K, b_K] \subset \mathbb{R}^K$*, with* $a_i < b_i < \infty$ *for* $1 \leq i \leq K$*, and assume* $\mathcal{D}_n$ *is generated by draws of* θ *from a uniform distribution over* $\mathbf{P}_\theta$*. Then, for any value of* $0 < p < 1$*, with probability* $1 - p$*, the maximum distance to the nearest training data point,* δ_n*, defined in* (4)*, satisfies:*

$$\delta_n = (-\log p)^{\frac{1}{K}} \cdot \mathcal{O}\left(n^{-\frac{1}{K}}\right). \tag{5}$$

Proof. See Appendix A.5. □

This lemma shows that, even when requiring a high confidence (i.e., p close to 0), the upper bound on δ_n decreases nearly as fast as $n^{-1/K}$. Thus, the granularity of uniform random draws is, asymptotically, similar to regular grids.

3.4.2 Kernel approximation

Next, we examine the properties and convergence speed of kernel smoothing (Li and Racine, 2023), a natural parameterization for f_n. We use a Nadaraya-Watson weighted average formula, while requiring specific properties of the kernel function that best fit our setting. Given training data $(\theta_i)_{1 \leq i \leq n}$, the kernel smoothing function is defined as:

$$f_n(\theta) = \frac{\sum_{i=1}^n k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i=1}^n k_n(\theta, \theta_i)}, \qquad k_n(\theta, \theta') = \eta\left(\frac{\|\theta - \theta'\|_2^2}{\lambda_n^2}\right), \tag{6}$$

where the weight function $\eta(\cdot)$ is a real-valued, continuously differentiable function that accepts scalar inputs. Moreover, $\eta(x) > 0$ for $0 \leq x < 1$, $\eta(x) = 0$ for $x \geq 1$, and $\eta(x)$ is non-increasing for $0 \leq x \leq 1$. The scaling parameter λ_n is set to be larger than the granularity of the training dataset δ_n, i.e., $\lambda_n = \gamma\, \delta_n$ for some constant $\gamma > 1$.

The following proposition summarizes the properties of kernel smoothing as $\delta_n \to 0$:

Proposition 2. *Let* $\hat{\theta}_n$ *be the approximate parameter estimate based on the kernel-smoothing function in* (6)*. Suppose Assumption 1 holds. Then, as* $n \to \infty$ *and the training data become more granular (*$\delta_n \to 0$*), we have:*

i. The approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$.

ii. Furthermore, ∇f_n *exists and is continuous on any neighborhood* $\mathbf{O} \subseteq \mathbf{P}_\theta$ *around* θ^* *that is open in* $\mathbb{R}^K$*. Assuming the rest of assumptions in Proposition 1 hold,*

the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(\delta_n\right). \tag{7}$$

iii. Assuming the conditions in part (ii) hold, if the training data lie on a regular grid, the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(n^{-\frac{1}{K}}\right). \tag{8}$$

Finally, as shown in Lemma 1, the same bound holds in probability when the training data is uniformly drawn from a bounded box in $\mathbb{R}^K$.

Proof. See Appendix A.6. □

Proposition 2 shows that, under regularity conditions, the convergence speed is at least of the same order as δ_n, the granularity of the training data. Additionally, if the training data are evenly spaced or drawn uniformly, the convergence error is given by $n^{-1/K}$. This is the classic curse of dimensionality: a large number of parameters dramatically slows down convergence speed.

3.4.3 Neural network approximation

We now explore the properties and convergence speed of approximate moments when f_n represents a neural network. To leverage existing results from the literature, we focus here on a training dataset generated by random draws of θ in $\mathbf{P}_\theta$. As we show below, neural networks converge faster than kernel, especially for high-dimensional parameters.

Assume each component function f_n^r, $1 \leq r \leq M$ of f_n is a shallow neural network (NN) of the class:

$$f_n^r(\theta) = c_0^r + \sum_{i=1}^{N} c_i^r \phi\left((a_i^r)^\top \theta + b_i^r\right), \tag{9}$$

where N is the number of nodes, ϕ is a sigmoid function which is infinitely differentiable, a_i^r is a K-dimensional vector, while c_0^r, c_i^r, and b_i^r are scalars.

The exact convergence speed of f_n depends on several factors, including the loss function used to train the network, the distribution of θ, or the smoothness of the true function f. For simplicity, we rely here implicitly on the assumptions needed to show convergence and convergence speed of f_n in Barron (1994). These assumptions yield

the following result for the approximate parameter estimate $\hat{\theta}_n$, labeled as "informal" since we do not fully specify the assumptions:

Proposition 3 (Informal). *Let $\hat{\theta}_n$ be the approximate parameter estimate based on a neural network described in (9), which is appropriately trained over the training dataset $\mathcal{D}_n$, generated by random draws of θ in P_θ. Moreover, suppose Assumption 1 holds. Then, as the training data size $n \to \infty$:*

i. The approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$.

ii. Suppose the neural net training procedure is appropriately set up such that f_n, together with f and θ^*, satisfies the assumptions of Proposition 1 and $\|\nabla f_n(\theta)\|_2$ is uniformly bounded across n and θ. Moreover, let the number of neural net nodes increase as $N = \sqrt{n/(K \log n)}$. Then, the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(\sqrt{M}\left(\frac{K \log n}{n}\right)^{\frac{1}{4}}\right). \tag{10}$$

Proof. See Appendix A.7. □

This informal proposition leverages both our earlier results and two important findings from the early machine-learning literature. First, Cybenko (1989) shows that neural networks uniformly converge toward continuous functions, which allows us to apply Theorem 1, establishing that $\hat{\theta}_n \to \theta^*$. This property, along with some regularity conditions, enables us to apply Proposition 1, which relates $\hat{\theta}_n - \theta^*$ to $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$. Finally, we use Barron (1994) (and its associated regularity conditions) to establish an upper bound on the convergence of $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$, which leads to our final result.

Proposition 3 shows that the convergence speed of the neural network approximation depends less on the dimension K compared to the kernel approximation, where K enters multiplicatively but not as a power of n. This result echoes standard findings in machine learning. Neural networks, by increasing the number of nodes, can better handle complexity and mitigate the curse of dimensionality.

While this result holds for a single layer NN, it is common to use, in practice, deeper NNs, rather than the shallow but very wide structure suggested by Barron (1994). At the very least, the above result provides an upper bound for the convergence error of deeper neural networks (i.e., a lower bound for the convergence speed). See, for instance, Farrell et al. (2021) for recent convergence speed results in the case

of Multi-Layer Perceptrons (MLPs). Using such advanced results seems promising, but beyond the scope of this paper.

4 What ensures the continuity of moments

A key assumption to ensure convergence of our approach is that the true moments of the estimated model $\mathcal{S}$, $f(\theta)$, are continuous. This may look like a restrictive assumption, but we show here that it does not require the estimated model to be continuous. We start with a generic characterization of models, and then move on to the specific class of dynamic choice models, the focus of our application.

4.1 Continuity of model, continuity of moments

We start with a generic model parametrized by $\theta \in \mathbb{R}^k$, with variables $x \in \mathbb{R}^d$ (some of these variables may be endogenous to the model, while others may be exogenous). For instance, in the corporate finance model we explore below, x could include capital, debt, and productivity shocks while θ contains deep parameters such as adjustment costs and productivity volatility.

Let $h(x, \theta)$ be an M-dimensional vector of functions of the variables x, whose expectation defines the moments of interest. Thus, the moment function f is:

$$f(\theta) = \mathbb{E}_x[h(x, \theta)].$$

h does not need to be continuous for f to be continuous in θ. Instead, we provide below a weaker sufficient condition that guarantees the continuity of f, based on the following definition:

Definition 1 (Light discontinuity)**.** *Consider the function $h(x, \theta)$, that is defined over $\mathbf{X} \times \mathbf{P} \subseteq \mathbb{R}^d \times \mathbb{R}^k$, and let $\mathbf{D}$ be the set of all discontinuity points of h, i.e.,*

$$\mathbf{D} = \{(x, \theta) \in \mathbf{X} \times \mathbf{P} \mid h \textit{ is discontinuous at } (x, \theta)\}. \tag{11}$$

Then, h is called lightly discontinuous *over θ (i.e., over its second input) if $\mathbf{D}$ can be expressed in terms of a countable number of functions F_r as follows:*

$$\mathbf{D} = \bigcup_{r=1}^{\infty} \{(x, \theta) \mid x_{\mathbf{I}} = F_r(x_{-\mathbf{I}}, \theta), \textit{ for some } \emptyset \neq \mathbf{I} \subseteq [d]\}, \tag{12}$$

where $[d] = \{1, \ldots, d\}$*, and for* $\mathbf{I} = \{i_1, \ldots, i_k\} \subseteq [d]$ *with* $i_1 < \ldots < i_k$*,* $x_{\mathbf{I}}$ *is the subvector* $x_{\mathbf{I}} = (x_{i_1}, \ldots, x_{i_k})$*, and* $-\mathbf{I} = [d] \setminus \mathbf{I}$*. Moreover, the domain of each function* F_r *is a subset of all valid inputs* $(x_{-\mathbf{I}}, \theta)$*, where* $(x, \theta) \in \mathbf{X} \times \mathbf{P}$*.*

Of course, all continuous functions are also lightly discontinuous. To further visualize this condition, consider the case where x is a scalar, i.e., $d = 1$. The set of discontinuity points becomes $\mathbf{D} = \bigcup_{r=1}^{\infty} \{(x, \theta) \mid x = F_r(\theta)\}$, meaning h has a countable number of discontinuities for each θ. Definition 1 generalizes to cases where x has more than one dimension. For $d > 1$, it requires that the discontinuity points of h form a countable number of (zero-volume) "curves" in $\mathbb{R}^d$ for each θ (as opposed to "points" in the case where $d = 1$). This condition is commonly satisfied by economic models with non-convexities, as discussed below.

The following theorem establishes that when f is composed of moments of "lightly discontinuous" functions, it is continuous. This key result, which requires additional but easily met technical conditions, is stated as follows:

Theorem 2 (Continuity of the moment function)**.** *Consider the (real- or vector-valued) function* $h(x, \theta)$ *defined over the convex set* $\mathbf{X} \times \mathbf{P} \subseteq \mathbb{R}^d \times \mathbb{R}^k$ *and the measure* μ_x *defined over* $\mathbf{X}$ *(which may or may not be a probability measure) corresponding to model variables* x*, where* μ_x *is dominated by the Lebesgue measure on* $\mathbb{R}^d$*. Additionally, assume there exists a function* g *defined over* $\mathbf{X}$ *such that* $\|h(x, \theta)\|_2 \leq g(x)$ *for all* $(x, \theta) \in \mathbf{X} \times \mathbf{P}$*, where* $h(x, \theta)$ *and* $g(x)$ *are measurable functions of* x*, and* $\int g(x)\, d\mu_x < \infty$*. Then, if* $h(x, \theta)$ *is lightly discontinuous over* θ*:*

i. The moment function $f(\theta) = \mathbb{E}_x[h(x, \theta)] = \int h(x, \theta) d\mu_x$ *is continuous.*

ii. As a special case of part (i), suppose μ_x *is a probability measure and the event* $A(x, \theta)$ *can be described by the indicator function* $h(x, \theta) = \mathbf{1}_{A(x,\theta)}$*, i.e.,* $h(x, \theta) = 1$ *if* $A(x, \theta)$ *occurs and* $h(x, \theta) = 0$ *otherwise. Then* $f(\theta) = \mathbb{P}_x[A(x, \theta)] = \mathbb{E}_x[h(x, \theta)]$ *is continuous.*

Proof. See Appendix A.8. □

Theorem 2 is a key result of this paper. A straightforward application is when h is continuous, as is often the case in economic models where variables and outcomes are continuous functions of parameters. In such cases, moments are continuous in deep parameters, ensuring that our approximation method converges. More interestingly, when h is discontinuous but its discontinuities can be described by a countable number of functions of $x_{-\mathbf{I}}$ and θ, the moments remain continuous.

This property applies to all economic models with non-convexities. For example, consider a neoclassical investment model with fixed adjustment costs, where capital k is the only state variable. These models feature an (s, S) rule, where next-period capital k' is discontinuous for low and high productivity levels $z = \underline{z}(k, \theta)$ and $z = \overline{z}(k, \theta)$, respectively. Between these bounds, no investment occurs, but when productivity crosses these thresholds, the firm discontinuously invests or disinvests. The policy function of these models thus satisfies the light discontinuity condition of Definition 1. Thus, despite the discontinuities in the policy function k', the moment of any continuous function of k and z will be continuous in θ, as shown by the theorem.

The intuition of the theorem is clear from the following one-dimensional example, inspired by (s, S) rules:

Example 1. *Assume $x > 0$ is a scalar and consider the function $h(x, \theta) = x\mathbf{1}_{x>\theta}$. The discontinuities of h occur at $x = \theta$, i.e., the set of discontinuity points can be written as $\{(x, \theta) \mid x = F_1(\theta)\}$, where the function $F_1 : [0, \infty) \to [0, \infty)$ is given by $F_1(\theta) = \theta$. This discontinuity set is of the form introduced by Definition 1, making h lightly discontinuous over θ. For instance, such a function h could be generated by an investment model with fixed adjustment costs and time-varying productivity x, where the firm only invests when productivity is sufficiently high. Theorem 2 shows that $\mathbb{E}_x[h(x, \theta)]$ is continuous. The reason is clear in this simple example. If G is the c.d.f. of x, we have $\mathbb{E}_x[h(x, \theta)] = \int_\theta x \, dG(x)$, which is not only continuous but differentiable with respect to θ. The same intuition holds for a higher number of breakpoints, provided they are countable.*

The second part of the theorem is a direct corollary of the first. If an econometrician computes the probability of an event $h(x, \theta) > 0$, this probability will be continuous in θ if $\mathbf{1}_{h(x,\theta)>0}$ is lightly discontinuous over θ. The discontinuity in $\mathbf{1}_{h(x,\theta)>0}$ occurs at the "boundary" of the region defined by $h(x, \theta) > 0$. By crossing this boundary, either by a jump in h or smoothly, $\mathbf{1}_{h(x,\theta)>0}$ "jumps" between 0 and 1. If this boundary can be described by the form introduced in Definition 1, then $\mathbf{1}_{h(x,\theta)>0}$ is lightly discontinuous over θ.

Consider again Example 1, but now focusing on $\mathbb{P}_x[h(x, \theta) > 0]$. The boundary of $h(x, \theta) = x\mathbf{1}_{x>\theta} > 0$ (considering the domain $x > 0$) is $x = \theta$, forming the discontinuity set $\{(x, \theta) \mid x = \theta\}$, which satisfies Definition 1. Theorem 2 then predicts that $\mathbb{P}_x[h(x, \theta) > 0]$ is continuous, which can be verified directly by noting that $\mathbb{P}_x[h(x, \theta) > 0] = \int_\theta dG(x)$, a clearly continuous (and differentiable) function of θ.

Functions that are not lightly continuous are unlikely to be used in structural

estimation. For example, consider the case where $h(x, \theta) = x\mathbf{1}_{\theta>0}$. In this case, $\mathbb{E}_x[h(x, \theta)] = \mathbb{E}_x[x] \cdot \mathbf{1}_{\theta>0}$ is clearly not continuous in θ. Our theorem does not apply because the function $x\mathbf{1}_{\theta>0}$ is *not* lightly discontinuous over θ– its discontinuity occurs for all values of x when $\theta = 0$. However, such functions are rare in economics and even less likely to be used in empirical analyses.

Finally, note that the continuity results in Theorem 2 hold for a broader notion of discontinuity, as described in Appendix A.8. We focus on the more restrictive Definition 1 because it is directly applicable to economic models.

4.2 Dynamic choice models

A downside of the previous result is that it relies on characterizing the model's output rather than its primitives. This means the model must first be solved to show that h is lightly discontinuous.

To characterize the model's primitives instead, we focus on a class of dynamic choice models for several reasons. First, these models generally lack closed-form solutions and must be solved numerically, making them a natural fit for our method. Second, they are widely used in the literature (e.g., in education (Keane and Wolpin, 1997), household finance (Gourinchas and Parker, 2002), corporate finance (Hennessy and Whited, 2007b), etc.). Finally, our two applications below come from this class of models.

More precisely, we consider the following class of models. Let $0 < \beta < 1$, $s \in \mathbf{P}_a$, $z \in \mathbf{P}_z$, and $\theta \in \mathbf{P}_\theta$, where each of $\mathbf{P}_a$, $\mathbf{P}_z$, and $\mathbf{P}_\theta$ lies in finite-dimensional Euclidean space. The class of models is described by the following Bellman equation:

$$\begin{aligned} V(s, z; \theta) = \sup_{a \in \mathbf{P}_a} \quad & \pi(a, s, z; \theta) + \beta\, \mathbb{E}_{z'}[\, V(a, z'; \theta) \mid z], \\ \text{s.t.} \quad & M_i(a, s, z; \theta) \leq 0, \quad 1 \leq i \leq N. \end{aligned} \tag{13}$$

where $\pi(\cdot)$ and $M_i(\cdot)$ are real-valued functions defined over $\mathbf{P}_a^2 \times \mathbf{P}_z \times \mathbf{P}_\theta$. Moreover, $\mu_{z'|z}$ is the probability measure defined over $\mathbf{P}_z$ corresponding to the random variable z' given z. The following theorem applies to cases where the model's primitives (payoff function, constraints) are continuous.

Theorem 3. *Consider the maximization problem introduced in* (13) *and assume the following conditions hold:*

i. $\mathbf{P}_\theta$, $\mathbf{P}_z$, *and* $\mathbf{P}_a$ *are nonempty and compact.*

ii. $\mathbf{P}_a$ *is convex and contains an open set.*

iii. $\mathbf{P}_z$ *is measurable and* $\mu_{z'|z}$ *is dominated by the Lebesgue measure.*

iv. π *and* M_i *are continuous.*

v. $M_i(a, s, z; \theta)$ *is convex in* a*.*

vi. For every $(s, z; \theta) \in \mathbf{P}_a \times \mathbf{P}_z \times \mathbf{P}_\theta$*, there exists* $a \in \mathbf{P}_a$ *such that* $M_i(a, s, z; \theta) < 0$ *for all* $1 \leq i \leq N$*.*

Then, there exists a unique and continuous $V(\cdot)$ *that solves the dynamic model in* (13)*, and the optimal solution is attained at some* $a^*(s, z; \theta)$ *that satisfies the constraints. Moreover, if* $a^*(s, z; \theta)$ *is unique for every* $(s, z; \theta) \in \mathbf{P}_a \times \mathbf{P}_z \times \mathbf{P}_\theta$*, then* $a^*(\cdot)$ *is continuous.*

Proof. See Appendix A.9. □

This theorem ensures that, as long as the model's primitives are continuous, the value function and the optimal policy functions are continuous as well. For this theorem to hold, profit and constraint functions need to be continuous. Condition (v) is typically satisfied in most problems. For instance, an equity issuance constraint in a corporate finance model corresponds to $M = -\pi$, and since profits in these models are continuous and concave, Condition (v) is met. Similarly, in models where debt d is constrained by firm value, the constraints correspond to $M = -\lambda V + d$, which is convex as long as V is concave.

Theorem 3 excludes cases where M and π are not continuous, such as models with lumpy investment costs or fixed stock market participation costs. These models feature "inaction bands" with discontinuous jumps; for instance, households may not buy stocks unless their wealth exceeds a certain threshold, after which they invest. Such policy functions are lightly discontinuous, as in Example 1. However, the literature has yet to demonstrate that solutions of these models generically include inaction bands with jumps (see, e.g., Elsby and Michaels (2019)).

5 Applications

We apply our method to two cases, examining its implementation and evaluating its performance. We then demonstrate its usefulness in studying robustness through identification analysis, moment selection, sample splits, and misspecification.

5.1 Dynamic corporate finance model

5.1.1 Model, parameters and moments

We use a standard model of firm dynamics with collateral constraints, as described in Catherine et al. (2022b). Time is discrete. Every period, the firm draws a productivity z_t, and chooses capital k_t and debt d_t to maximize a discounted sum of per-period cash flows, subject to a financing constraint. The discount rate is r.

At time t, the firm's profit is: $\pi_t = e^{z_t(1-\alpha)}k_t^{\alpha}$. Productivity follows an AR(1) process $z_t = \rho_z z_{t-1} + \eta_t$. The variance of the innovation term η_t is σ_z^2.

Investment $i_t = k_{t+1} - (1-\delta)k_t$ entails a convex cost $\frac{\gamma}{2}\frac{i_t^2}{k_t}$, where δ is the depreciation rate. Profits, net of interest payments and depreciation (δk_t), are taxed at a rate $\tau = 1/3$. This tax rate applies both to negative and positive profits so that firms receive a tax credit when their accounting profits are negative. The firm can hold cash (when $d_t < 0$) or issue debt ($d_t > 0$), which is risk-free and incurs an interest rate r.

Financing frictions arise from two constraints: (1) a collateral constraint that limits its borrowing $d_{t+1} \leq \lambda k_{t+1}$ (2) a linear cost (ξ) of equity issuance, such that negative cash-flows to equity are multiplied by $(1+\xi)$.

In total, there are seven structural parameters: $\theta = (\delta, \gamma, \alpha, \rho_z, \sigma_z, \xi, \lambda)$. We also compute the "value loss of financing constraints", which represents the difference in the mean log value of constrained firms compared to similar unconstrained firms (i.e., $\xi = 0$).

The literature has used a variety of moments to identify these parameters. We define f as the function linking the parameters θ to a set of 17 moments used in previous work. The first seven moments are : mean(investment/assets), mean(profit/assets), mean(equity issuance/assets), mean(leverage), mean autocorrelation of investment, std(log growth sales) and std(log growth 5yr sales). They correspond to the moments used in Catherine et al. (2022b), who also discuss how they identify the model's parameters. An additional 10 moments, used in the literature, are described in Appendix B.1. Throughout our analysis, the dataset we use is a COMPUSTAT extract from 1971–2019.

5.1.2 Training the approximate moment function f_n

We now construct the approximate moment function f_n. We restrict structural parameters to a compact box $\mathcal{P}$ with the following ranges: $\delta \in [0; .2]$, $\gamma \in [0; .3]$, $\alpha \in [.5; .9]$,

$\rho_z \in [.5; .98]$, $\sigma_z \in [0.2; 2]$, $\xi \in [0; .3]$ and $\lambda \in [0; .6]$. We generate a training dataset $\mathcal{D}_n$ using a Halton sequence of $n = 50,000$ parameters drawn from $\mathcal{P}$. For each draw, we solve the model, and calculate the 17 corresponding moments using simulations. This step takes approximately 140 hours with our numerical setup.

We also generate a separate "validation" dataset of 1,000 parameters using a Halton sequence in $\mathcal{P}$. Some draws leads to models that cannot be identified. For instance, if the cost of equity issuance is sufficiently large, firms won't issue any equity. As a result, the ratio of equity issuance to assets cannot identify the cost of equity issuance beyond a threshold. This non-identification is not specific to our approximation, i.e. the true model is also not identified for these draws.

To detect these "badly identified" draws, we compute for each draw in the validation sample $(J^\top(\theta_i)WJ(\theta_i))^{-1/2}$, where J is the Jacobian matrix of the true model and W is the inverse of the variance-covariance matrix of the 17 moments, estimated by bootstrapping the full sample. We then drop all draws for which one of the diagonal elements of this matrix is 10 times larger than the standard error of the full-sample parameter estimates (obtained through real SMM, see below).[3] We end up with a validation sample of 215 observations.

We train a neural network f_n using this training data. In Appendix B.4, we compare this neural network approximation to a kernel method and show that the neural network achieves a better fit, consistent with the faster convergence of neural networks in high-dimensional problems (see Proposition 3). The neural network is a Multi-Layer Perceptron (MLP) with 5 layers and 512,256,128,64,32 nodes. Implementation details are in Appendix B.3. This architecture is more complex than in proposition 3, as deep NNs have been shown to converge faster when the target function f is smooth enough (Farrell et al., 2021), especially for larger dimensions.

5.1.3 Approximate moment estimation: performance

To validate our approach, we use two methods: (1) comparing the approximate estimates to the true parameters using the validation sample and (2) comparing the approximate estimates to actual SMM estimates on real data.

For each draw in the validation sample, we estimate the seven structural parameters by targeting the first seven moments described in Section 5.1.1 using the approximate moment function f_n. Figure 1 shows scatter plots comparing the estimated

[3] While this selection criterion is somewhat arbitrary, we have experimented with alternative definitions and found that this did not affect our assessment of the estimates' precision.

and true parameters across all the draws. R^2 values are above 95% for all parameters. We also observe that some parameters are better estimated than others. This is because, despite the filter applied to the validation sample, some draws still lead to poorly identified models. This is in particular true for equity issuance costs, as low equity issuance is consistent with a large range of equity issuance cost parameters. Nevertheless, Figure 1 establishes that the method performs well overall.

Next, we compare the approximate estimates to actual SMM estimates on moments from real data. The advantage of this approach is that we know that the true model is well-identified when matched on real empirical moments. We identify the true SMM parameters using a standard procedure (Tik Tak algorithm): (1) we initialize the algorithm by evaluating the SMM objective at 50,000 different starting points (2) we run Nelder-Mead optimizations at the 50 best starting points using at most 200 function evaluations. The model is estimated by targeting the full-sample moments shown in column 1 of Appendix Table B.1. Table 1 reports true SMM estimates and approximate estimates, while Appendix Table B.1 provide moment fits.[4] Table 1 shows that true and approximate parameter estimates are all within a few percentage points of each other. Line 3 applies the approximation error correction from Section 3.3, which further reduce the bias of the approximate estimate.

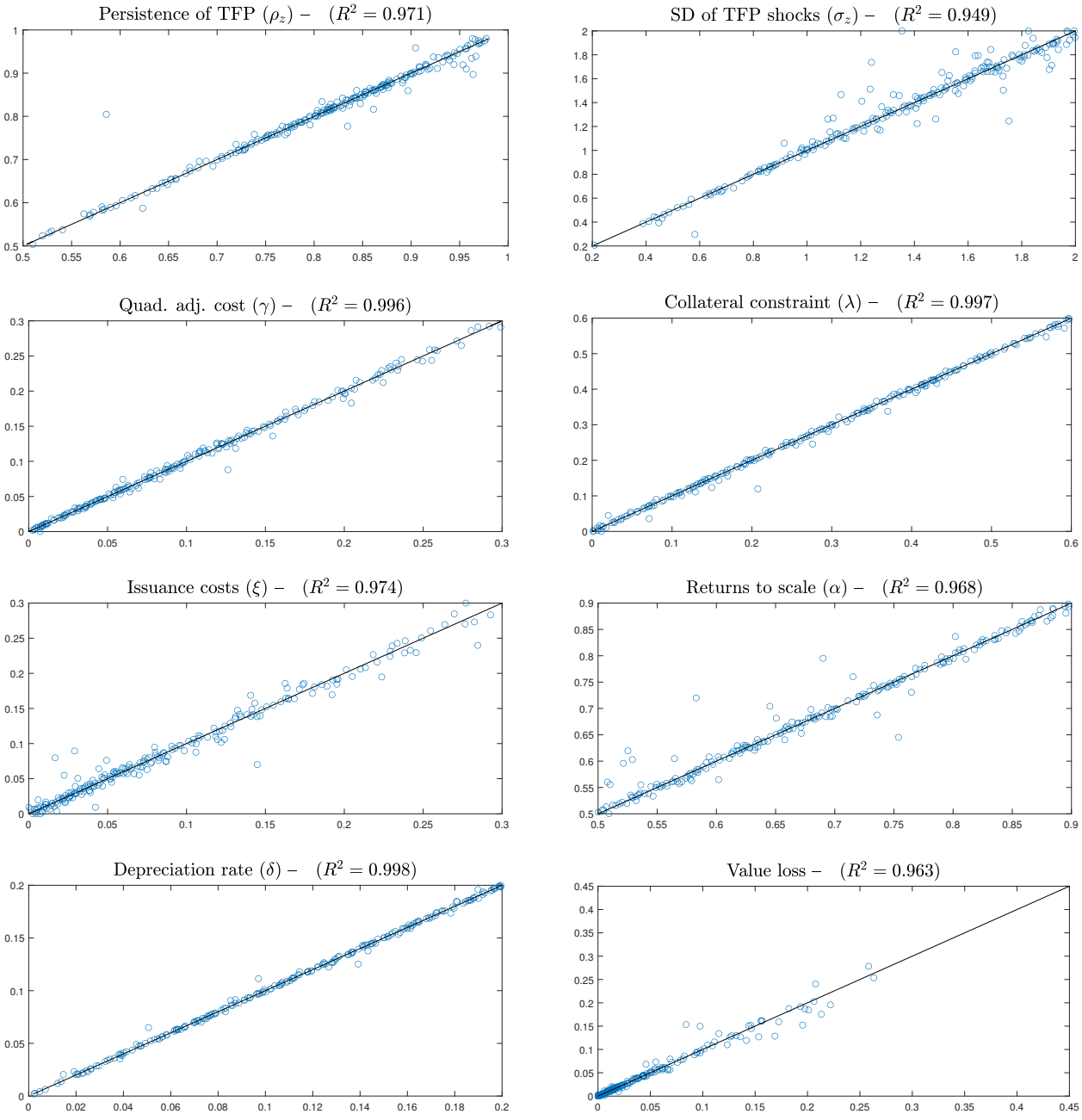
5.1.4 Identification diagnostic

Our method can also assess how moments affect parameter estimates, a computationally intensive task requiring re-estimating the model for many different values of the targeted moments.

For brevity, Figure 2 illustrates this approach by focusing on a specific moment, the volatility of 5-year sales growth. The yellow line shows how this simulated moment changes as a function of the different parameters. This is the typical "comparative static" figure reported in structural work. In contrast, the blue line corresponds to our identification diagnostic: we re-estimate the model many times by targeting many possible values for the volatility of 5-year sales growth and report the relation between the approximate parameter estimates and the targeted moment values. Every point on the blue line thus corresponds to a different estimation. The red line

[4]Appendix Figure B.2 shows that our approach is faster than the standard SMM by several orders of magnitude. The computing times we report exclude the simulation of the training sample, which is long but required for both estimations. For the true SMM, it takes an additional 17 minutes for the estimation to converge to its final value. In contrast, the approximation-based estimation converges in less than a second (provided the NN has been estimated).

Figure 1: Out-of-sample Performance



Notes. This figure shows the precision, in the validation sample, of our benchmark approximate SMM across estimated parameters. For each draw $\theta, f(\theta)$ in the validation sample, we use neural nets to construct the approximate moment function f_n and estimate parameters $\widehat{\theta}_n$. The x-axis reports the true parameters θ, and the y-axis reports the estimated parameters $\widehat{\theta}_n$.

Table 1: Parameter Estimates: true vs. approximate SMM

	ρ_z	σ_z	γ	λ	ξ	α	δ	value loss
true SMM	.7178	1.1566	.0413	.1087	.0378	.8155	.0670	.0250
- s.e., local deriv.	.0068	.0508	.0024	.0030	.0021	.0074	.0008	.0019
approx. SMM	.7270	1.1413	.0439	.1070	.0378	.8141	.0669	.0254
- s.e., local fit deriv.	.0070	.0560	.0024	.0029	.0017	.0082	.0008	.0018
approx. SMM, corrected	.7179	1.1569	.0412	.1087	.0381	.8158	.0671	.0252
- s.e., local fit deriv.	.0065	.0525	.0022	.0030	.0016	.0077	.0008	.0017

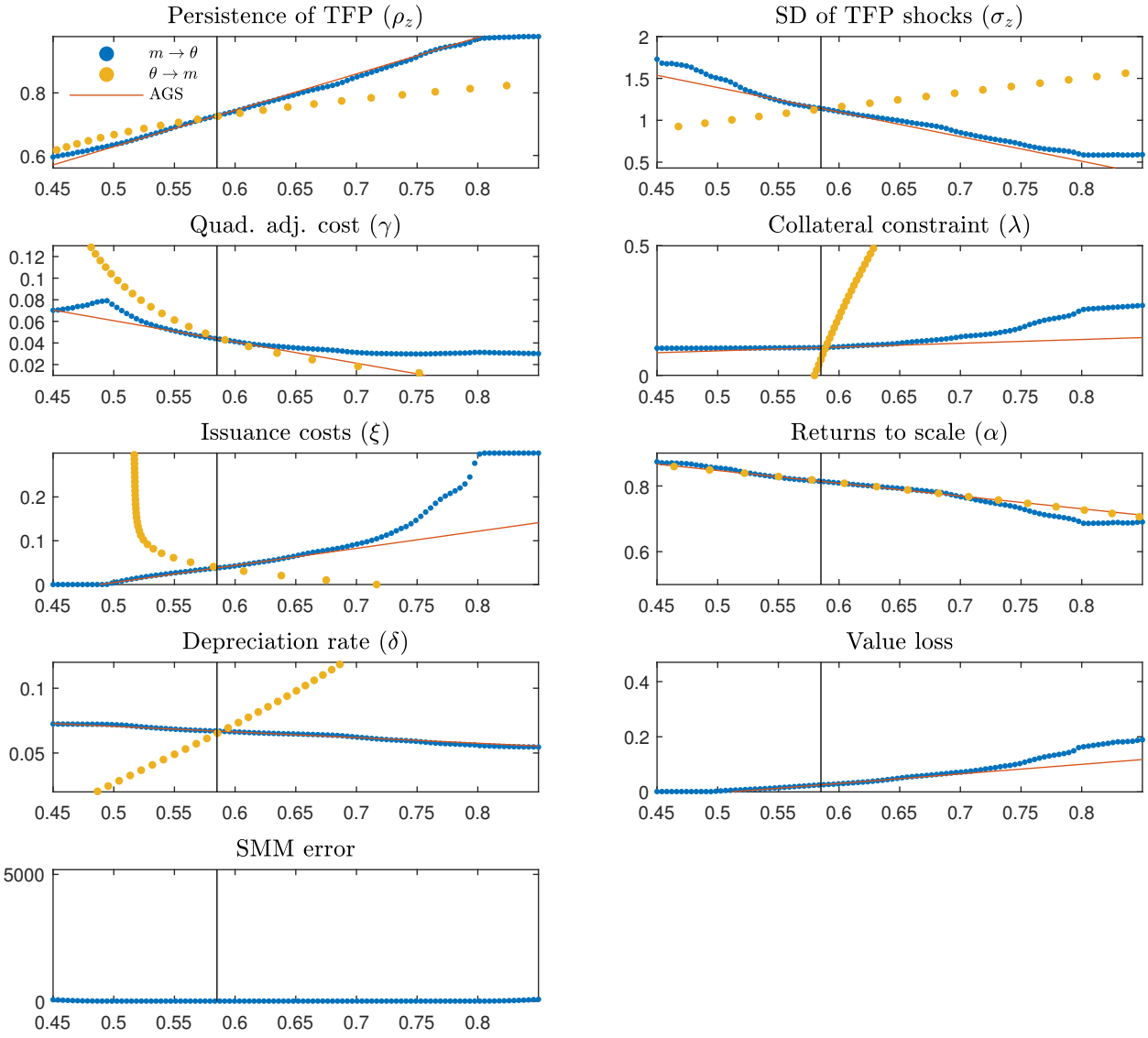
Notes. The table reports parameter estimates of the corporate finance model presented in Section 5.1.1. The first line corresponds to parameter estimates using a standard SMM technique (as described in the text). The second line shows parameter estimates using NN fit as approximate moment function. The third line shows the approximate SMM estimates, corrected using the approximation error formula from Section 3.3. 's.e., local deriv.' corresponds to the standard errors of the true SMM parameters, calculated using the Jacobian matrix of the true model f. 's.e., local fit deriv.' corresponds to the standard errors of the approximate SMM parameters, calculated using the Jacobian matrix of the approximate moment function f_n. We use the delta method to calculate the standard error of value loss.

shows the local linear approximation from Andrews et al. (2017), which corresponds to a linearized version of our technique around the SMM estimate and does not require re-estimating the model.

Figure 2 illustrate the advantages of our methodology. First, the standard "comparative static" figure reported in the literature can be misleading. For instance, the yellow line on the top-right panel shows that a larger volatility of TFP shocks (σ_z) increases the volatility of 5-year sales growth. One could use this analysis to conclude that samples with larger volatility of 5-year sales growth would lead to infer a higher σ_z. This conclusion would be incorrect. The blue line on the same panel shows that a higher volatility of 5-year sales growth instead leads to infer *a lower* σ_z. The reason for this discrepancy is that the "comparative static" approach ignores that parameter estimates jointly depend on all targeted moments. This is easily seen in our context. For a given volatility of 1-year sales growth, a larger volatility of 5-year sales growth implies that TFP shocks are more persistent (i.e., a higher ρ, as shown in the top-left panel). Because TFP persistence is higher, capital becomes more responsive to TFP shocks, which, in turn, increases the volatility of 1-year sales growth. To match the data, the model then needs to *reduce* σ, the volatility of TFP shocks.

Second, Figure 2 also illustrates that the local linear approximation in Andrews et al. (2017) can fail to hold outside of a close vicinity of the actual moments used in the estimation. For instance, consider the case where the volatility of 5-year sales growth was 80%, instead of its empirical value of 58%. The linear approximation on

Figure 2: Relationship between parameters to 5-year growth volatility



Notes. This figure plots parameter values on the y-axis and the volatility of 5-year sales growth on the x-axis. The yellow line draws local comparative statics, i.e. how variations in one parameter around its estimated value – holding other parameters fixed at their estimated value – affect the moment value. The blue line plots how variations in the moment value – holding other moments fixed at their empirical value – affect the estimation of each parameter. For each new set of moment, parameters are re-estimated using our approximation technique. The red line is the local approximation of Andrews et al. (2017). The black vertical line shows the value of the moment in the data.

the red line suggests that the collateral constraint parameter (λ) would be close to the one we estimate in the main estimation, leading to the conclusion that financial constraints would be about the same in this alternative sample. Our diagnostic shows instead that λ would be then about twice larger, suggesting instead significantly lower financial constraints.

5.1.5 Robustness to moment selection

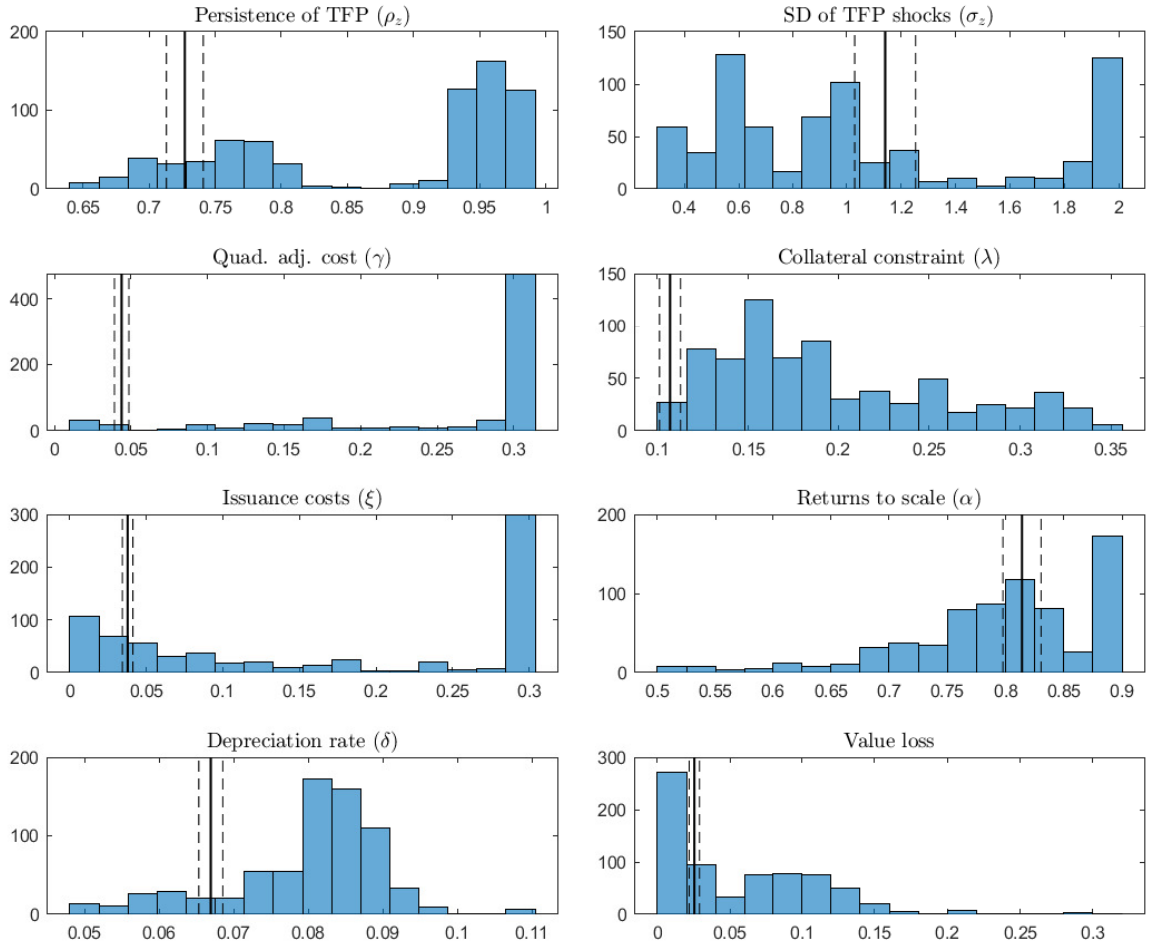
Methods of moments do not provide guidance on which moments to target in estimation. We can however use our approximation approach to evaluate the robustness of parameter estimates to moment selection. Conceptually, this is straightforward: we consider many combinations of possible targeted moments and re-estimate the model for each combination. Robustness is established if parameter values remain stable across most estimations. This robustness exercise cannot be done with standard estimation techniques, as it requires many estimations. Our approximation approach makes this feasible at low computational cost.

Concretely, we re-estimate our model by targeting both (a) the 7 moments used in our baseline *and* (b) any possible subset of the remaining 10 moments. This requires $2^{10} = 1,024$ separate estimations. Of these, we only consider estimations that correspond to reasonably well-identified sets of moments. We drop cases where the standard errors for an estimated parameter is more than 10 times larger than the standard error of the baseline true SMM.[5] This leaves us with 722 estimations. Figure 3 reports the distribution of parameter estimates. The solid black lines correspond to the baseline estimates and dashed lines show the 95% confidence interval. A baseline estimate robust to moment selection would have a large share of alternative estimates close to the baseline value.

Most parameter estimates for this model are not robust to moment selection. For instance, the estimated variance of innovations to z, σ_z, which has a benchmark estimate of 1.1, admits a wide range of alternative estimates – from 0.4 and 1.1 – depending on which moments are targeted in the estimation. We also see that in many alternative estimations, the parameter estimate for σ_z hits its upper bound of 1.5. Similar conclusions are obtained for all parameters.

[5] As before, we compute standard errors through the classic SMM formula $(J'WJ)^{-1/2}$, using as weight matrix the inverse of the variance-covariance matrix of empirical moments.

Figure 3: Histogram of estimates across 722 sets of targeted moments (Corporate Finance Model)



Notes. This figure explores the sensitivity of parameter estimates to moment selection. Our baseline estimation targets seven moments $(m_i)_{i\in\{1..7\}}$. We consider a set of 10 additional moments used in the literature to estimate similar models: $(m_i)_{i\in\{8..17\}}$ described in Section 5.1.2. We construct all possible sets of moments that contain the seven baseline moments and any combination of the 10 additional moments. These sets of moments are then used as targeted moments in an approximate SMM. This results in 1,024 sets of parameter estimates. After dropping cases where estimates are poorly-identified – where the standard errors for an estimated parameter is more than 10 times larger than the standard errors of the baseline estimate – we end up with 702 estimates. Each panel in the figure shows the distribution of parameters across these 702 estimations. The vertical black line and dashed lines show the baseline parameter estimates, together with their 95% confidence interval.

5.1.6 Sample splits

Our method also allows for re-estimating the model across subsamples, a common practice in reduced-form analysis but computationally expensive for structural models. Such sample splits can be useful as they can help test for specific mechanisms or identify useful trends in the data.

Figure 4 shows time-varying estimates of parameters: every year t, we re-estimate the model by targeting moments estimated over the $[t-5; t+4]$ period. This analysis reveals interesting trends. For instance, the collateral constraint parameter λ goes from .2 in the 1970s to close to 0 in the 2000s. One explanation is the well-documented increase in cash holdings over this period (Bates et al., 2009): a reduction in net leverage leads the model to believe that firms are more financially constrained and that the collateral parameter λ is thus smaller. Similarly, the estimated depreciation rate steadily declines from 8% to 4%. One possible explanation is the rise in intangible capital over the period (Crouzet and Eberly, 2018). As the model does not feature intangible capital, a reduction in physical investment rate can only be attributed to a decline in depreciation rate.

5.2 Model Misspecification

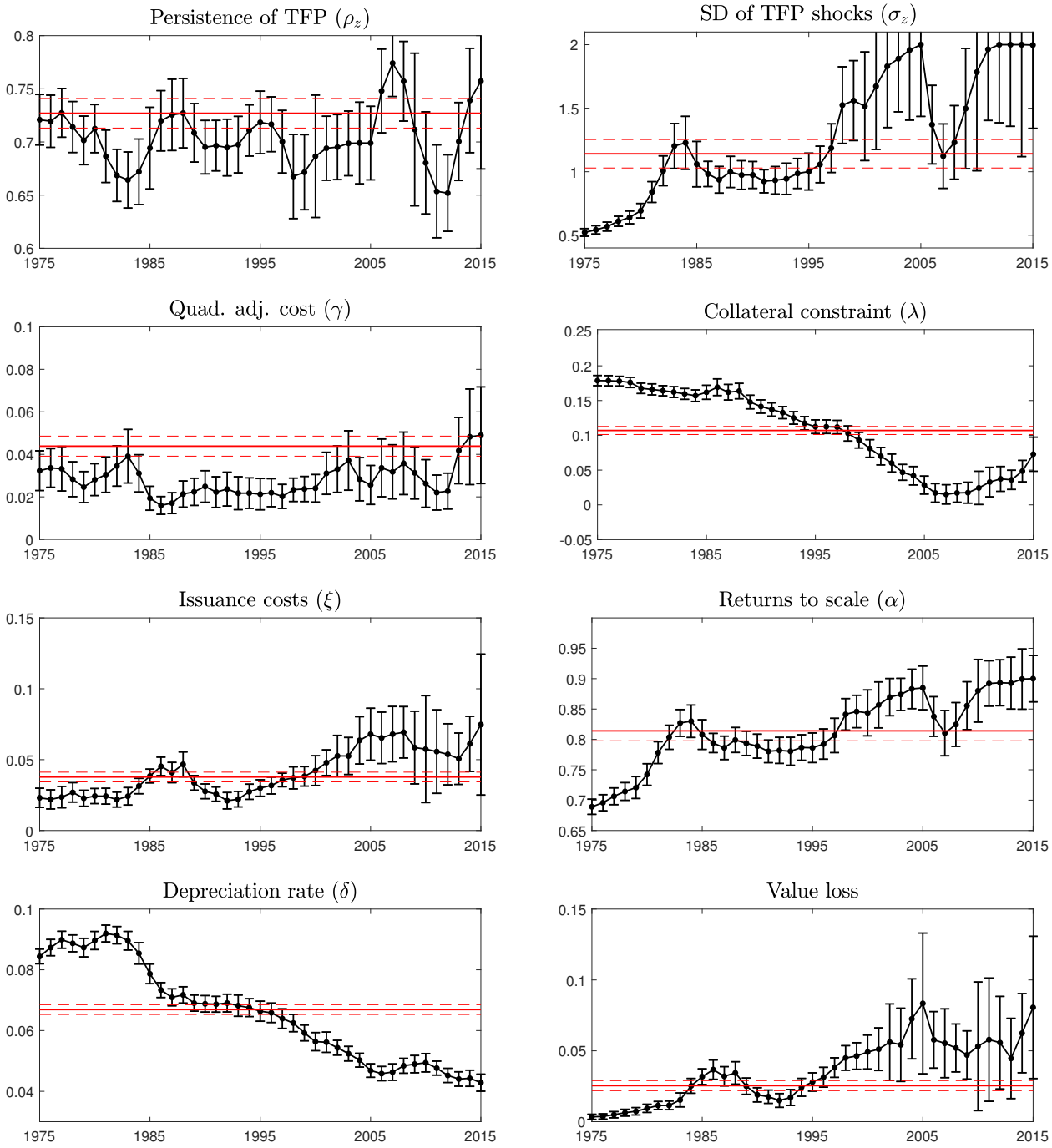
Finally, we explore model misspecification by simulating alternative models and re-estimating the baseline model using moments generated from these alternatives. The baseline estimate will be robust to misspecification if it is close to these alternative estimates. This approach is similar to Catherine et al. (2022b).

We illustrate our approach in the context of the recent corporate finance literature, which shows that financial constraints often take the form of cash-flow constraints rather than collateral constraints (e.g. Lian and Ma (2021), Greenwald (2019)). Our objective is to assess the robustness of the baseline estimates to this particular source of misspecification.

To do so, we augment the baseline model of Section 5.1.1 to introduce this new channel. We rewrite the debt constraint as: $d_t < \lambda k_t + \lambda_2 . \mathrm{E}[e^{z_t(1-\alpha)}] k_t^{\alpha}$. We then simulate 40 versions of this alternative model, where baseline parameters are the baseline estimates of Table 1, and λ_2 increases uniformly from 0 (no misspecification) to 2 (large misspecification). This provides us with 40 sets of alternative moments.

We then estimate the baseline model (i.e., the model with $\lambda_2 = 0$) 40 separate times by targeting these 40 different sets of moments. Figure 5 plots the estimated value loss from financing constraints (y-axis) against the value of λ used to simulate

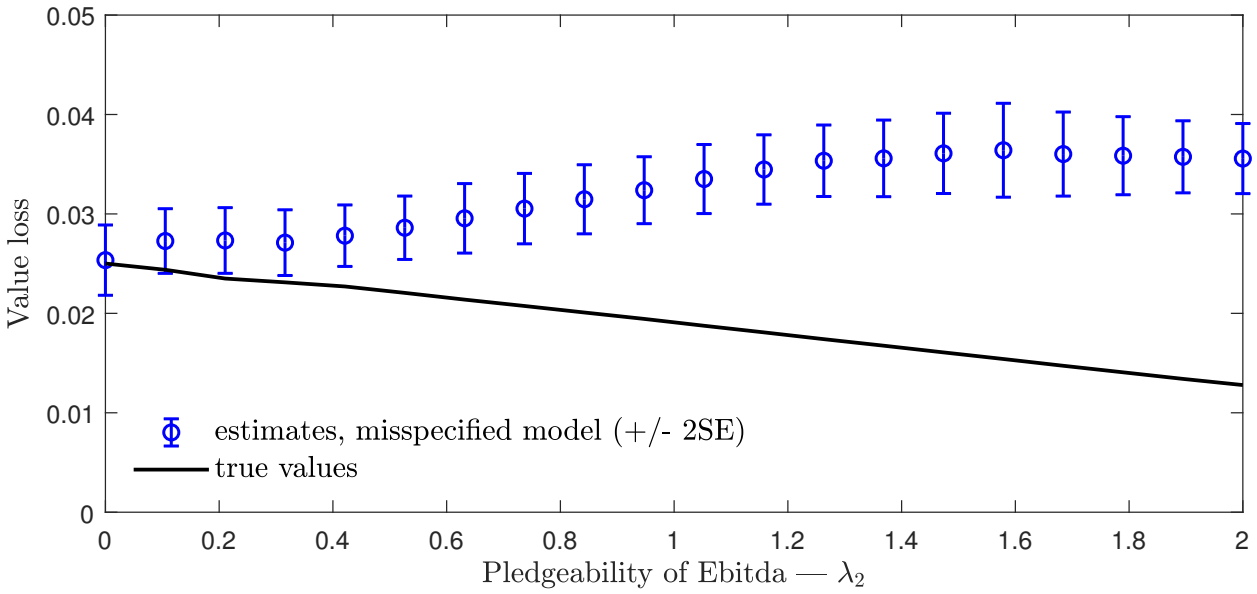
Figure 4: Time Series Estimates (Corporate Finance Model)



Notes. This figure shows the sensitivity of parameter estimates to the sample period used to compute the targeted moment. For each year t on the x-axis, we compute the seven baseline moments $(m_i)_{i\in\{1..7\}}$ on the sample period $[t-5, t+4]$. We then re-estimate the model using the benchmark approximate SMM that targets the moments measured for this sub-period. The y-axis reports the resulting parameter estimates and their 95% confidence interval. The horizontal solid and dashed red lines corresponds to the baseline estimates obtained when computing moments on the entire sample, together with their 95% confidence interval.

the targeted moments (x-axis). The black line corresponds to the actual value loss of financing constraints in the correctly-specified model with both collateral and cash-flow constraints, which intuitively decreases as λ_2 increases.

Figure 5: Model Mispecification: Estimating Model with Collateral Constraints on Data Generated by Cash-flow Constraints (Corporate Finance Model)



Notes. This figure explores the sensitivity of estimation to model misspecification. We consider an augmented version of our corporate finance model where firms can pledge a multiple λ_2 of their EBITDA so that their debt constraint takes the form $d_t < \lambda k_t + \lambda_2 . \mathrm{E}[e^{z_t(1-\alpha)}] k_t^{\alpha}$. We assume that the correctly specified model is the augmented model where the baseline parameters are set to their estimated value in Table 1 and λ_2 takes various values from 0 (the baseline model) to 2. For each possible value of λ_2, we re-estimate the baseline parameter values using the mis-specified model that targets these moments (simulated with the correctly-specified model). The x-axis plots the values of λ_2 used in each estimation. The blue circles report the value loss from financial constraint estimated with the approximation to the misspecified model while we target moments generated by the correctly-specified model. The black line reports the value loss from financial constraint in the correctly-specified model.

Figure 5 shows that estimates of value loss (from financial constraints) are robust to values of λ_2 below 0.6. However, as λ_2 increases, the misspecified model wrongly infers increasing value losses (while the true value losses are in fact going down). A simple explanation is that as λ_2 increases, firms can avoid issuing equity and the average equity to asset ratio in the economy goes down. The baseline model infers from this reduced equity issuance that the costs of equity issuance is going up, which leads to a large value loss from financial constraint. This analysis illustrates how we can leverage our methodology to evaluate the effect of specific model misspecification.

5.3 Dynamic portfolio choice model

We now apply our method to a life-cycle portfolio choice model.

5.3.1 Model description

The model is a variation of Catherine et al. (2022a), excluding housing investment. Households choose between investing in a risk-free asset and the stock market over their life cycle. Compared to Viceira (2001) and Cocco et al. (2005), the model includes countercyclical income risk (as in Guvenen et al. (2014)) and a realistic Social Security system. Details on labor income, stock market, and Social Security specifications are in Appendix C.1.

Given processes for labor, capital and social security incomes, the agent maximizes the discounted sum of utilities from consumption C_t:

$$V_{t_0} = \mathrm{E} \sum_{t=t_0}^{T} \beta^{t-1} \left(\prod_{k=0}^{t-1} (1 - m_k) \right) \frac{C_t^{1-\gamma}}{1-\gamma}, \tag{14}$$

where γ is the coefficient of relative risk-aversion, m_k the mortality rate at age k, β the subjective discount factor and T the maximum lifespan. Mortality rates and age of death are calibrated. Short selling or leveraging the stock market is not allowed, so that the wealth share of equity is between 0 and 1. Finally, owning any equity at date t costs Φ times the macro component of wages (see Appendix C.1 for details).

With other parameters calibrated, we estimate three model parameters: the discount factor β, risk aversion γ, and the cost of owning equity Φ ($\theta = (\gamma, \beta, \Phi)$).

As in Catherine et al. (2022a), parameter estimation is based on matching three moments from the Survey of Consumer Finances (1989–2016): (1) the wealth-to-labor income ratio, (2) the percentage of households holding equity, and (3) the stock share of wealth among equity holders. The construction of these moments is explained in Appendix C.2, and their role in identifying θ is discussed in Catherine et al. (2022a).

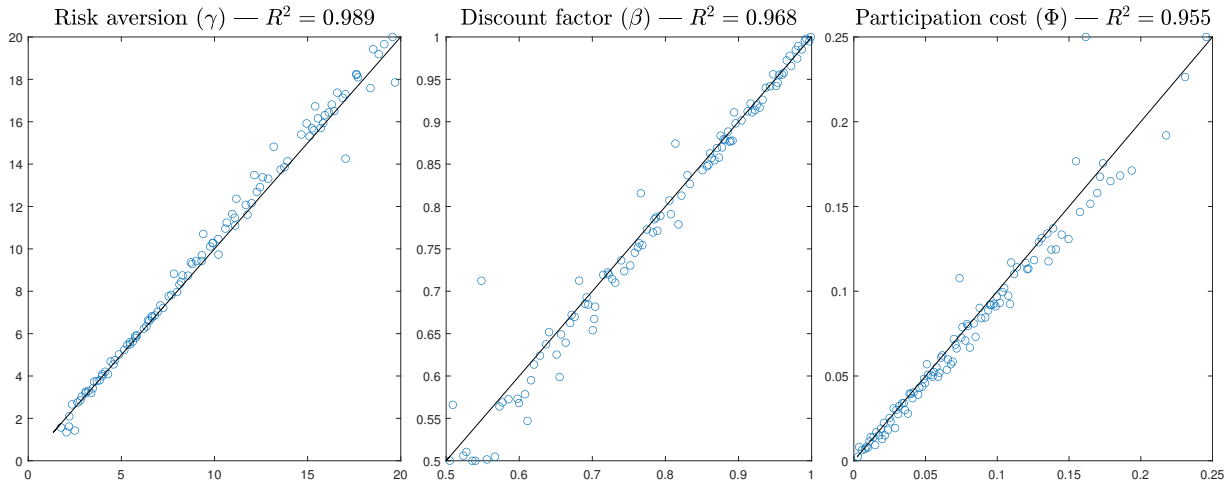
5.3.2 Training the aproximate moment function f_n

Similar to the corporate finance model, we build a training sample to train the approximate moment function f_n. The sample contains 1,992 random parameter draws θ_i and the corresponding moments $f(\theta_i)$ obtained by simulating the model. We also construct a validation sample of 106 well-identified parameter values to evaluate the precision of the approximation f_n. Exact details about the construction of these samples are provided in Appendix C.3. Note that the training and validation samples are smaller than in our corporate finance example, so we can assess the effect of sample size on our methodology.

Next, we train the approximate moment function f_n using the training sample and a neural network (a 5-layer MLP with 256, 128, 64, 32 and 16 nodes).

The results are similar to the corporate finance model, as shown in Figure 6, the counterpart to Figure 1. For each draw in the validation sample, the scatter plot compares the true parameters (x-axis) to the parameters estimated using f_n (y-axis). The out-of-sample R^2 is above 95% for all three parameters. Appendix C.4 shows that the kernel smoother performs as well as the deep NN in this application. This can be due to the lower-dimensional setting (3 parameters) or possibly to the smaller training sample, which may reduce the relative precision of the NN.

Figure 6: Performance of Estimation using Benchmark Approximation (Household Finance Model)



Notes. This figure shows, for the household finance model, the precision, in the validation sample, of our benchmark approximate SMM across estimated parameters. For each draw $\theta, f(\theta)$ in the validation sample, we use neural nets to construct the approximate moment function f_n and estimate parameters $\widehat{\theta}_n$ which match $f(\theta)$. The x-axis reports the true parameters θ, while the y-axis reports the estimated parameters $\widehat{\theta}_n$.

As in the corporate finance application, we also test our method on real data, using the three core empirical moments from the SCF (see Appendix C.2). We estimate parameters using (1) standard SMM, (2) the approximate NN moment function, and (3) the approximate moment method corrected with the error estimate formula from (3).

Table 2 shows that the approximate estimates are close to the standard SMM estimates.[6] Both approaches give similar values for risk-aversion (γ about 8.3) and discount factor (β about .91). The participation cost Φ is estimated at .0053 in the

[6] In Appendix C.5, we also show that the approximate method is orders of magnitude faster than running the true SMM.

Table 2: Parameter Estimates: true vs. approximate SMM

	γ	β	Φ
true SMM	8.328	0.9106	0.0053
- s.e., local deriv.	0.064	0.0021	0.0002
approx. SMM	8.621	0.9048	0.0056
- s.e., local fit deriv.	0.083	0.0024	0.0003
approx. SMM, corrected	8.371	0.9098	0.0052
- s.e., local fit deriv.	0.077	0.0023	0.0003
estimation, lower bound	1.01	0.5	0
estimation, upper bound	20	1	0.25

Notes. This table reports the parameter estimates of the household finance model presented in Section C.1. We report parameter estimates using the true SMM (first line), the benchmark approximate SMM (second line), and the approximate SMM corrected with the first-order error estimate (third line). The benchmark approximation is a neural net. γ is risk-aversion. β is the subjective discount factor. Φ is the participation cost.

true SMM, and .0056 in the approximate SMM, a small economic difference of about \$50 a year. The third line confirms that the approximation error formula (3) usefully corrects the approximate SMM.

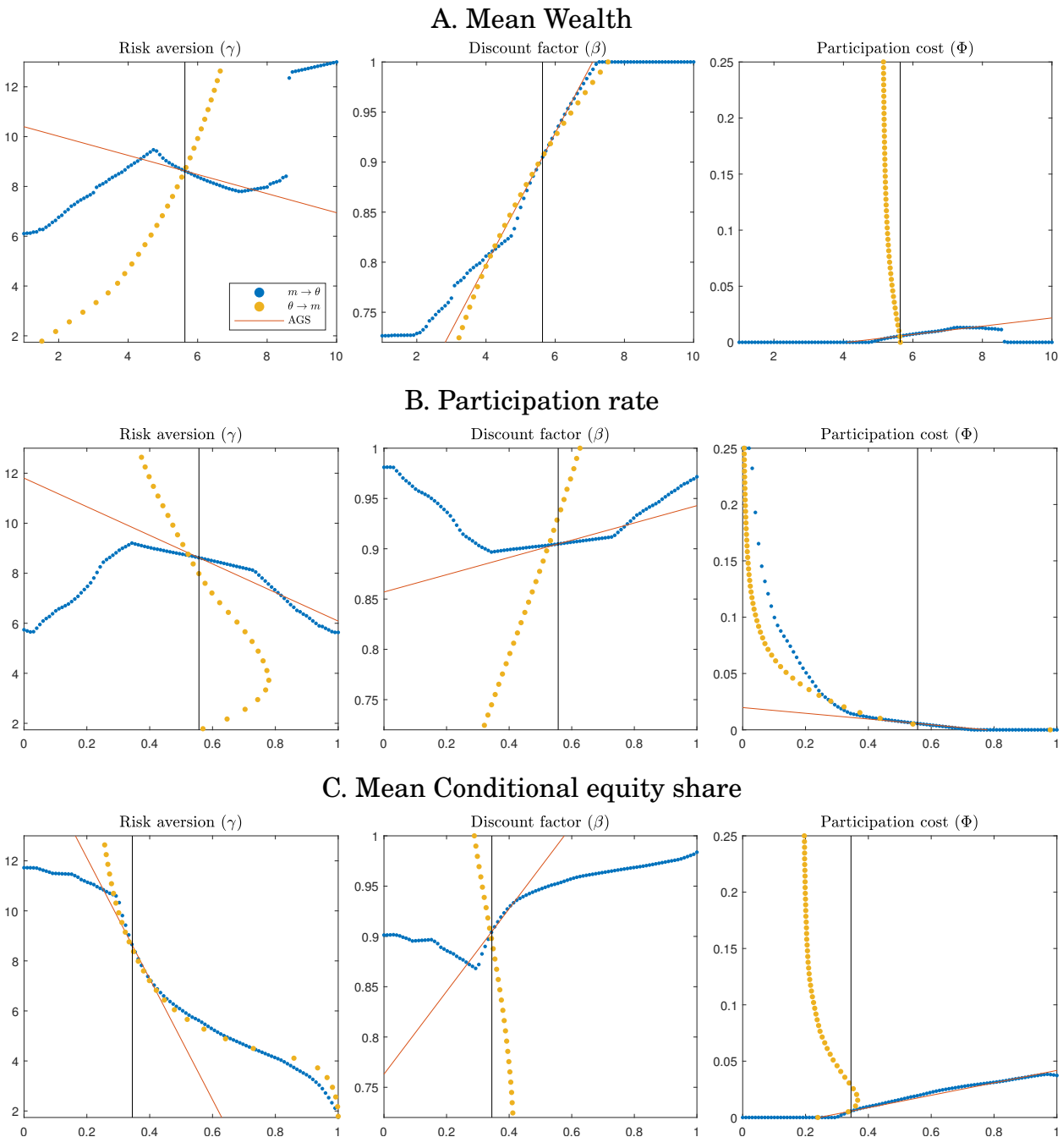
5.4 Identification Diagnostic

Figure 7, analog to Figure 2, shows how changes in targeted moments affect estimated parameters (blue line). This analysis is made possible by the low computational cost of the approximate SMM. There are nine figures, one for each moment-parameter pair. The figure includes nine panels, each representing a moment-parameter pair.

Similar takeaways emerge from this analysis. First, relying on local comparative statics from $f(\theta)$ can be misleading. For instance, in the top left panel, comparative statics suggest that higher mean wealth leads to increased risk aversion (yellow line), as more risk-averse households save more. However, the actual estimation (blue line) shows the opposite: holding the participation rate and equity share constant, the model needs to *reduce* risk aversion to match higher household savings and equity investment.

Second, the linear approximation from Andrews et al. (2017) can fail when con-

Figure 7: Sensitivity of Parameters to Moments (Household Finance Model)



Notes. This figure plots parameter values on the y-axis and the value of moment on the x-axis. Panel A (resp. B and C) corresponds to mean wealth) (resp. participation rate and conditional equity share). The yellow line draws the function $f(\theta)$ – moments as functions of parameter values. The blue line plots how variations in moment values – holding other moments fixed – affect parameter estimates. Each dot on the blue line corresponds to a separate estimation. Finally, the red line corresponds to the local linear approximation of the blue line around the parameter estimates, i.e. the "sensitivity matrix" of Andrews et al. (2017). The black vertical line indicates the value of a moment in the data.

sidering moment values that differ significantly from their empirical values. For example, an equity share of 0.6 (instead of 0.35 in the SCF) would lead to inferring a risk-aversion of 5 in actual estimation; the linear approximation would lead to estimate it at 2 instead.

5.5 Robustness to Moment Selection

As in our corporate finance application, we analyze the robustness of parameter estimates to moment selection. To do this, we expand our set of moments from the initial three to include an additional 64 moments. The first group of four moments deals with normalization and non-normality: (1) median wealth, (2) median conditional equity share, (3) an alternative definition of mean conditional equity share, normalized by financial wealth rather than net worth, and (4) the median of this alternative conditional equity share. The first two moments address the fat upper tails of stock holdings and total wealth, while the last two adjust for differences between net worth and financial wealth, which are identical in the model but not in the data.

The second group consists of 60 moments, which are the baseline moments (mean wealth, stock participation, and equity share) broken down by 20 age groups, ranging from ages 23–25 to 80–82.[7] The procedure is explained in detail in Catherine (2021).

We then re-estimate the model using 72 alternative sets of moments. Unlike the corporate finance case, where all combinations were explored, the portfolio choice model cannot simultaneously match equity shares normalized by both total wealth and financial wealth. Thus, we focus on the following moment combinations:

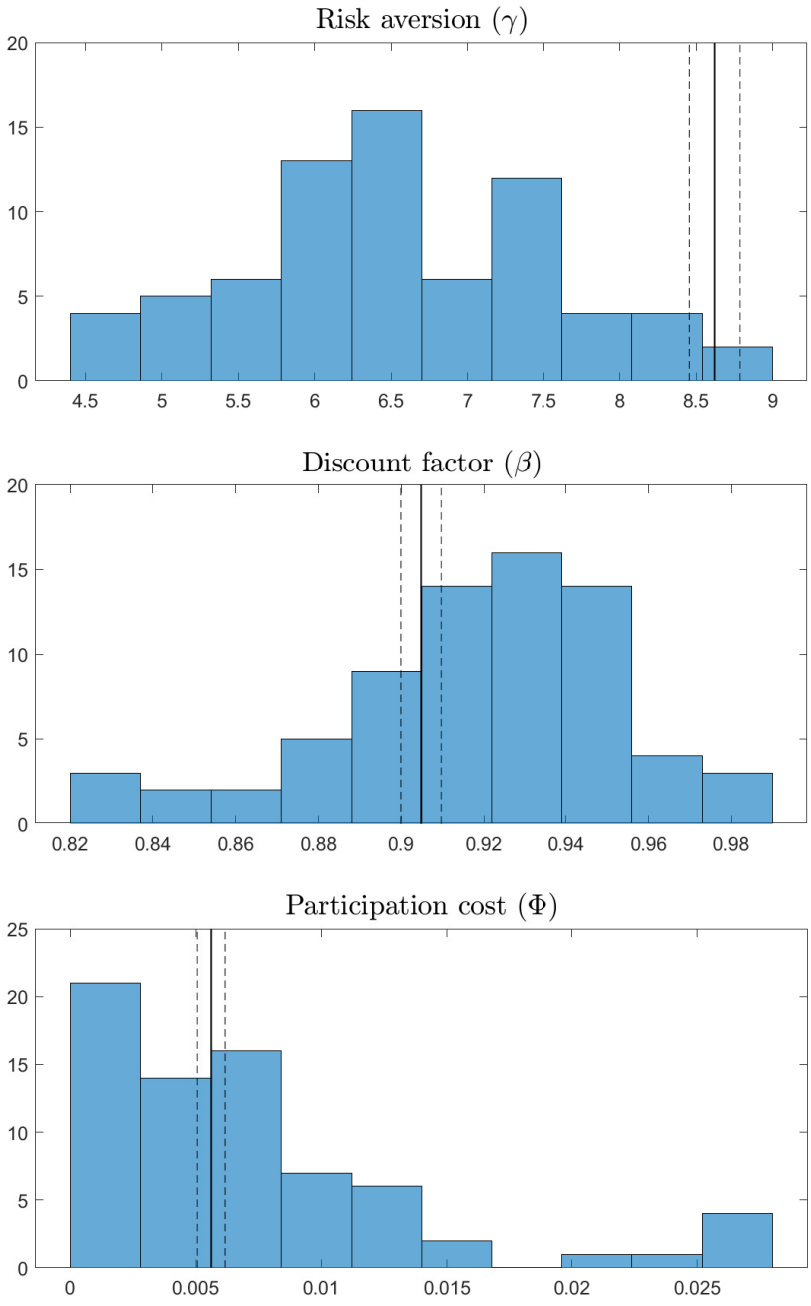
- 12 combinations of the three baseline moments: (a) mean wealth (overall or by age group), (b) mean participation (overall or by age group), and (c) mean equity share (overall, by age group, or normalized by financial wealth).
- 36 combinations that add either (a) median wealth, (b) median equity share, or (c) median equity share normalized by financial wealth to the previous 12.
- 24 combinations that add both median wealth and one of the two median equity share measures to the previous combinations.

Figure 8 shows the distribution of parameter estimates across all 72 sets of moments. Most combinations produce consistent risk-aversion estimates, with the dis-

[7] We compute the life-cycle of *unconditional* equity share, instead of conditional equity share, as the model often returns zero participation for some age groups, making the conditional mean equity share undefined.

tribution peaking at 6–7, although this peak is substantially lower than the baseline estimate of 8.5. The discount factor (β) estimates range from 0.9 to 0.96, closely clustering around the baseline estimate of 0.93. For participation costs, the distribution peaks at 0, significantly lower than the baseline estimate of 0.005.

Figure 8: Histogram of estimates across 72 sets of targeted moments (Household Finance Model)



Notes. This figure explores the sensitivity of parameter estimates to moment selection. Our baseline estimation targets three moments $(m_i)_{i\in\{1..3\}}$. We consider 72 alternative set of moments described in Section 5.5. Each panel in the figure shows the distribution of parameters across these 72 estimations. The vertical black line and dashed lines report the baseline parameter estimates, together with their 95% confidence interval.

6 Conclusion

This paper presents a fast and straightforward method for conducting robustness checks and identification diagnostics in structural estimation. Our approach involves estimating an approximation of the relation between model parameters and moments using a training dataset. Once this approximation is fitted, it allows for the rapid estimation of structural parameters. In our two applications—corporate finance and household finance models—we demonstrate that this "approximate SMM" drastically reduces computational costs compared to standard SMM methods, with minimal loss of precision.

This reduction in computational burden opens up three important exercises that were previously challenging to conduct. First, we can now easily assess parameter robustness to moment selection. Second, sample-split analyses become feasible, allowing for quicker assessments of model robustness and validity across different subsamples. Lastly, the approximate SMM enables us to explore the sensitivity of baseline estimates to misspecification bias. By simulating various alternative models, we can evaluate how deviations from the baseline model affect parameter estimates.

References

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 2017, *132* (4).

_ , _ , and _ , "On the Informativeness of Descriptive Statistics for Structural Estimates," *Econometrica (Matthew Gentzkow's Fisher-Schultz Lecture)*, 2020, *88* (6), 2231–2258. Working Paper.

_ , _ , and _ , "Transparency in Structural Research," *Journal of Business and Economic Statistics (invited discussion paper)*, 2020, *38* (4), 711–722.

Armstrong, Timothy B. and Michal Kolesár, "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, January 2021, *12* (1), 77–108.

Azinovic, Marlon, Luca Gaegauf, and Simon Scheidegger, "Deep equilibrium nets," 2019.

Barron, Andrew R, "Approximation and estimation bounds for artificial neural networks," *Machine learning*, 1994, *14*, 115–133.

Bates, Thomas, Kathleen Kahle, and René Stulz, "Why Do U.S. Firms Hold So Much More Cash than They Used To?," *Journal of Finance*, 2009, *64*, 1985–2021.

Bonhomme, Stéphane and Martin Weidner, "Minimizing Sensitivity to Model Misspecification," Papers 1807.02161, arXiv.org July 2018.

Catherine, Sylvain, "Countercyclical Labor Income Risk and Portfolio Choices over the Life Cycle," *The Review of Financial Studies*, 12 2021. hhab136.

_ , Paolo Sodini, and Yapei Zhang, "Countercyclical Income Risk and Portfolio Choices: Evidence from Sweden," *Working Paper*, 2022.

_ , Thomas Chaney, Zongbo Huang, David Sraer, and David Thesmar, "Quantifying Reduced-Form Evidence on Collateral Constraints," *Journal of Finance*, 2022.

Chen, Hui, Antoine Didisheim, and Simon Scheidegger, "Deep Structural Estimation:With an Application to Option Pricing," Cahiers de Recherches Economiques du Département d'économie 21.14, Université de Lausanne, Faculté des HEC, Département d'économie February 2021.

Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout, "Consumption and Portfolio Choice over the Life Cycle," *The Review of Financial Studies*, 02 2005, *18* (2), 491–533.

Crouzet, Nicolas and Janice Eberly, "Intangibles, Investment, and Efficiency,"

American Economic Review: AEA Papers and Proceedings, 2018, *108*, 426–431.

Cybenko, George, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, 1989, *2* (4), 303–314.

Duarte, Victor, "Machine Learning for Continuous-Time Finance," 2020.

Elsby, Michael W.L. and Ryan Michaels, "Fixed adjustment costs and aggregate fluctuations," *Journal of Monetary Economics*, 2019, *101*, 128–147.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra, "Deep Neural Networks for Estimation and Inference," *Econometrica*, 2021, *89* (1), 181–213.

Fernández-Villaverde, Jesús, Galo Nuño, George Sorg-Langhans, and Maximilian Vogler, "A Deep Learning Algorithm for High-Dimensional Dynamic Programming Problems," 2021.

Fernandez-Villaverde, Jesus, Mahdi Ebrahimi Kahou, Jesse Perla, and Arnav Sood, "Exploiting Symmetry in High-Dimensional Dynamic Programming," 2021.

Folland, Gerald B., *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., New York: Wiley, 1999.

Fonseca, Julia, Victor Duarte, Aaron Goodman, and Jonathan Parker, "Benchmarking Global Optimizers," NBER Working Papers 29559 2022.

Gomes, Francisco, Michael Haliassos, and Tarun Ramadorai, "Household Finance," *Journal of Economic Literature*, September 2021, *59* (3), 919–1000.

Gouriéroux, Christian and Alain Montfort, *Simulation-Based Econometric Methods* 1996.

Gourinchas, Pierre-Olivier and Jonathan A. Parker, "Consumption over the Life-Cycle," *Econometrica*, 2002, *70* (1), 47–89.

Greenwald, Dan, "Firm Debt Covenants and the Macroeconomy: The Interest Coverage Channel," Technical Report 2019.

Guvenen, Fatih, Serdar Ozkan, and Jae Song, "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 2014, *122* (3), 621 – 660.

Halmos, P.R., *Measure Theory* Graduate Texts in Mathematics, Springer New York, 2013.

Hennessy, Christopher A. and Toni M. Whited, "How Costly Is External Financing? Evidence from a Structural Estimation," *The Journal of Finance*, 2007, *62* (4), 1705–1745.

_ **and** _ , "How Costly is External Financing? Evidence From a Structural Estimation," *Journal of Finance*, 2007.

Huber, Peter J., *Robust Statistics*, Berlin, Heidelberg: Springer Berlin Heidelberg,

Kato, Tosio, *Perturbation theory for linear operators*, Vol. 132, Springer Science & Business Media, 2013.

Keane, Michael and Kenneth Wolpin, "The Career Decisions of Young Men," *Journal of Political Economy*, 1997, *105* (3).

Lang, Robert, "A note on the measurability of convex sets," *Archiv der Mathematik*, 1986, *47*, 90–92.

Lang, S., *Real and Functional Analysis* Graduate Texts in Mathematics, Springer New York, 2012.

Li, Qi and Jeffrey Scott Racine, *Nonparametric econometrics: theory and practice*, Princeton University Press, 2023.

Lian, Chen and Yueran Ma, "Anatomy of Corporate Borrowing Constraints," *Quarterly Journal of Economics*, 2021, *136*.

Maliar, Lilia, Serguei Maliar, and Pablo Winant, "Will Artificial Intelligence Replace Computational Economists Any Time Soon?," CEPR Discussion Papers 14024, C.E.P.R. Discussion Papers September 2019.

Munkres, James R., *Topology*, 2nd ed., Prentice Hall, 2000.

Norets, Andriy, "Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations," *Econometric Reviews*, 2012, *31* (1), 84–106.

Rockafellar, R Tyrrell and Roger J-B Wets, *Variational analysis*, Vol. 317, Springer Science & Business Media, 2009.

Rudin, Walter, *Principles of mathematical analysis* 1953.

Stokey, Nancy L, Robert E Lucas, and Edward C Prescott, "Recursive Methods in Economic Dynamics," 1989.

Strebulaev, Ilya A. and Toni M. Whited, "Dynamic Models and Structural Estimation in Corporate Finance," *Foundations and Trends(R) in Finance*, November 2012, *6* (1–2), 1–163.

Viceira, Luis M., "Optimal Portfolio Choice for Long-Horizon Investors with Nontradable Labor Income," *The Journal of Finance*, 2001, *56* (2), 433–470.

Villa, Alessandro T. and Vytautas Valaitis, "Machine Learning Projection Methods for Macro-Finance Models," *International Political Economy: Investment & Finance eJournal*, 2019.

APPENDIX – FOR ONLINE PUBLICATION

A Proofs

In this section, we provide the mathematical proofs of the statements in the main body of the paper. We begin by outlining the necessary conventions, definitions, and elementary facts in Appendix A.1. Next, in Appendix A.2, we present lemmas and their proofs, which are used to establish the statements in the subsequent sections.

A.1 Preliminaries

A.1.1 Conventions

Unless otherwise deducible from the context, we assume that each set mentioned in this section lies in some finite-dimensional Euclidean space. The transpose operation is denoted by $\cdot^\top$. If $x \in \mathbb{R}^d$, we consider x as a column vector, i.e., x is $d \times 1$. The probability of an event A with respect to the randomness of x is denoted by $\mathbb{P}_x[A]$ or simply $\mathbb{P}[A]$, and the expectation of a function $g(x)$ of the random variable x is denoted by $\mathbb{E}_x[g(x)]$ or simply $\mathbb{E}[g(x)]$. The set of positive integers is denoted by $\mathbb{N}$, and the set of real numbers by $\mathbb{R}$. The set difference is denoted by $\mathbf{A} \setminus \mathbf{B}$, defined as $\mathbf{A} \setminus \mathbf{B} = \{x \in \mathbf{A} \mid x \notin \mathbf{B}\}$. The Cartesian product of sets is denoted by $A \times B$, and $A^2 = A \times A$. The volume of a set $\mathbf{A}$ is denoted by $\mathrm{vol}(\mathbf{A})$. The minimum and maximum of finitely many real numbers are denoted by $\min(r_1, \ldots, r_n)$ and $\max(r_1, \ldots, r_n)$, respectively.

A.1.2 Definitions

1. **Indexed sequences:** We consider sequences of objects like $\{a_k\}_{k \in \mathbf{N}}$, where the corresponding index set $\mathbf{N} \subseteq \mathbb{N}$ has infinitely many elements. When it is clear from the context, we omit the index set and simply refer to the sequence as a_k. Moreover, the corresponding convergence is considered based on $\mathbf{N}$. For instance, when the elements a_k are vectors in $\mathbb{R}^d$, the convergence of a_k to a, denoted by $a_k \to a$, simply means that for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that for every $k \in \mathbf{N}$ with $k \geq N$, we have $\|a_k - a\|_2 < \epsilon$. If this does not hold, we write $a_k \nrightarrow a$.
2. **Extended real system:** We consider $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\} = [-\infty, \infty]$. The corresponding extended arithmetic and conventions can be found in Rockafellar and Wets (2009, Chapter 1).
3. **Minimizers and maximizers:** For $h : \mathbb{R}^d \to \overline{\mathbb{R}}$, note that $\inf h$ and $\sup h$ always exist in $\overline{\mathbb{R}}$. Moreover, the set of minimizers and maximizers of h are defined as:
$$\arg\min h = \{x \in \mathbb{R}^d \mid h(x) = \inf h\}, \quad \arg\max h = \{x \in \mathbb{R}^d \mid h(x) = \sup h\}. \tag{A.1}$$
4. **Neighborhood of infinity:** Define:
$$\mathcal{N}_\infty = \{\mathbf{N} \subseteq \mathbb{N} \mid \mathbb{N} \setminus \mathbf{N} \text{ is finite}\}, \quad \mathcal{N}_\infty^\# = \{\mathbf{N} \subseteq \mathbb{N} \mid \mathbf{N} \text{ is infinite}\}. \tag{A.2}$$
5. **Properness**: A function $h : \mathbb{R}^d \to \overline{\mathbb{R}}$ is called *proper* if there exists $x \in \mathbb{R}^d$ such that $h(x) < \infty$ and, for all $x \in \mathbb{R}^d$, we have $-\infty < h(x)$.
6. **Level sets:** For a real-valued function $h : \mathbf{X} \to \overline{\mathbb{R}}$ and $\alpha \in \mathbb{R}$, the *level set* of h corresponding to α is defined as $\mathrm{lev}_{\leq \alpha} h = \{x \in \mathbf{X} \mid h(x) \leq \alpha\}$.

7. **Level-boundedness:** A sequence of functions $\{h_i\}_{i\in\mathbb{N}}$, with $h_i : \mathbf{X} \to \overline{\mathbb{R}}$, is called *eventually level-bounded* if there exists $\mathbf{N} \in \mathcal{N}_\infty$ such that for every $i \in \mathbf{N}$, h_i is *level-bounded*, i.e., for every $\alpha \in \mathbb{R}$, the level set $\operatorname{lev}_{\leq\alpha} h$ is bounded (possibly empty).

8. **Painlevé-Kuratowski convergence of sets:** (Rockafellar and Wets, 2009, Definition 4.1) Consider the sequence of sets $\{\mathbf{C}_i\}_{i\in\mathbb{N}}$, with the notions of $\liminf_i \mathbf{C}_i$ and $\limsup_i \mathbf{C}_i$ defined as:

$$\limsup_{i\to\infty} \mathbf{C}_i = \left\{x \mid \exists \mathbf{N} \in \mathcal{N}_\infty^{\#},\ \exists x_k \in \mathbf{C}_k \text{ for } k \in \mathbf{N},\ \text{with } x_k \to x\right\}, \tag{A.3}$$

$$\liminf_{i\to\infty} \mathbf{C}_i = \left\{x \mid \exists \mathbf{N} \in \mathcal{N}_\infty,\ \exists x_k \in \mathbf{C}_k \text{ for } k \in \mathbf{N},\ \text{with } x_k \to x\right\}. \tag{A.4}$$

   Then $\{\mathbf{C}_i\}_{i\in\mathbb{N}}$ is said to converge to $\mathbf{C}$ if $\mathbf{C} \subseteq \liminf_{i\to\infty} \mathbf{C}_i$ and $\limsup_{i\to\infty} \mathbf{C}_i \subseteq \mathbf{C}$, or equivalently $\mathbf{C} = \liminf_{i\to\infty} \mathbf{C}_i = \limsup_{i\to\infty} \mathbf{C}_i$. In such a case, we write $\mathbf{C}_i \to \mathbf{C}$.

9. **Kuratowski continuity of set-valued functions:** (Equivalent to Rockafellar and Wets (2009, Definition 5.4)) Consider a set-valued function $\mathbf{P}(x)$, where $x \in \mathbb{R}^d$ and $\mathbf{P}(x) \subseteq \mathbb{R}^k$, and let $\mathbf{X} \subseteq \mathbb{R}^d$. Then $\mathbf{P}(x)$ is called *outer semi-continuous* at $x_0 \in \mathbf{X}$ relative to $\mathbf{X}$ if, for every sequence $\{x_i\}_{i\in\mathbb{N}}$ in $\mathbf{X}$ with $x_i \to x_0$, we have $\limsup_{i\to\infty} \mathbf{P}(x_i) \subseteq \mathbf{P}(x_0)$. It is *inner semi-continuous* if $\mathbf{P}(x_0) \subseteq \liminf_{i\to\infty} \mathbf{P}(x_i)$.

   $\mathbf{P}(x)$ is called *continuous* at $x_0 \in \mathbf{X}$ relative to $\mathbf{X}$ if it is both outer and inner semi-continuous at x_0 relative to $\mathbf{X}$. This is equivalent to having $\mathbf{P}(x_i) \to \mathbf{P}(x_0)$ (in the Painlevé-Kuratowski sense) for every sequence $\{x_i\}_{i\in\mathbb{N}}$ in $\mathbf{X}$ with $x_i \to x_0$. If $\mathbf{P}(x)$ is (outer semi- or inner semi-) continuous at every $x_0 \in \mathbf{X}$ relative to $\mathbf{X}$, then $\mathbf{P}(\cdot)$ is (outer semi- or inner semi-) continuous relative to $\mathbf{X}$.

10. **Graph of set-valued functions:** Consider a set-valued function $\mathbf{P}(x)$, where $x \in \mathbb{R}^d$ and $\mathbf{P}(x) \subseteq \mathbb{R}^k$. The *graph* of $\mathbf{P}$ is defined as $\operatorname{gph} \mathbf{P} = \{(x, u) \in \mathbb{R}^d \times \mathbb{R}^k \mid u \in \mathbf{P}(x)\}$.

11. **Inverse image under set-valued functions:** Consider a set-valued function $\mathbf{P}(x)$, where $x \in \mathbb{R}^d$ and $\mathbf{P}(x) \subseteq \mathbb{R}^k$. The *inverse image* of a subset $\mathbf{Y} \subseteq \mathbb{R}^k$ is defined as $\mathbf{P}^{-1}(\mathbf{Y}) = \{x \in \mathbb{R}^d \mid \mathbf{P}(x) \cap \mathbf{Y} \neq \emptyset\}$.

12. **Open and closed balls:** Given $\theta_0 \in \mathbb{R}^d$ and $r > 0$, the open and closed balls around θ_0 with radius r are denoted by $\mathbf{B}(\theta_0, r)$ and $\bar{\mathbf{B}}(\theta_0, r)$, respectively, and are defined as follows:

$$\mathbf{B}(\theta_0, r) = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta_0\|_2 < r\}, \quad \bar{\mathbf{B}}(\theta_0, r) = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta_0\|_2 \leq r\}. \tag{A.5}$$

13. **Diameter of sets:** Let $\mathbf{C} \subseteq \mathbb{R}^d$. The diameter of $\mathbf{C}$ is defined as $\operatorname{diam}(\mathbf{C}) = \sup_{x,y\in\mathbf{C}} \|x - y\|_2$.

14. **Preimages:** For a function $h : \mathbf{X} \to \mathbf{Y}$ and $\mathbf{B} \subseteq \mathbf{Y}$, the *preimage* of $\mathbf{B}$ under h is defined as $h^{\text{pre}}(\mathbf{B}) = \{x \in \mathbf{X} \mid h(x) \in \mathbf{B}\}$.

15. **Supremum norm on functions:** For a real-valued function $h : \mathbf{X} \to \mathbb{R}$, the supremum norm $\|\cdot\|_\infty$ is defined as $\|h\|_\infty = \sup_{x\in\mathbf{X}} |h(x)|$.

16. **Epigraphs:** For $h : \mathbb{R}^d \to \overline{\mathbb{R}}$, the *epigraph* of h is defined as $\operatorname{epi} h = \left\{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} \mid \alpha \geq h(x)\right\}$.

17. **Cluster points:** A point a is called a *cluster point* of the sequence a_i if there exists a subsequence of a_i that converges to a.

18. **Interior of sets:** For $\mathbf{A} \subseteq \mathbb{R}^d$, the *interior* of $\mathbf{A}$, denoted by $\operatorname{int}(\mathbf{A})$, is the union of all subsets of $\mathbf{A}$ that are open in $\mathbb{R}^d$.

19. **Absolute continuity (dominance) of measures:** Consider measures μ and ν on the same measurable space $(\mathbf{X}, \mathcal{F})$. Then, μ is *dominated* by ν (or *absolutely continuous* with respect to ν), denoted by $\mu \ll \nu$, if for every $\mathbf{A} \in \mathcal{F}$, $\nu(\mathbf{A}) = 0$ implies that $\mu(\mathbf{A}) = 0$.

20. **Negligible sets:** Consider a measure space $(\mathbf{X}, \mathcal{F}, \mu)$. A set $\mathbf{A} \subseteq \mathbf{X}$ is called *negligible* with respect to μ (or μ-negligible) if $\mathbf{A} \in \mathcal{F}$ and $\mu(\mathbf{A}) = 0$.

21. **Jacobian matrices:** Consider the function $h : \mathbf{X} \to \mathbb{R}^k$ with $\mathbf{X} \subseteq \mathbb{R}^d$. Given the existence of an open neighborhood $\mathbf{O} \subseteq \mathbf{X}$ around a, the Jacobian matrix of h at a, denoted by $\nabla h(a)$ (if it exists), is a $k \times d$ matrix whose ij-th element is $\partial h^i / \partial x^j(a)$. Here, $\partial h^i / \partial x^j$ represents the partial derivative of the i-th component of h with respect to the j-th element of the input.

22. **Uniform convergence of functions:** Consider the functions $h_i, h : \mathbf{X} \to \mathbb{R}^k$. The sequence $\{h_i\}_{i \in \mathbb{N}}$ *uniformly converges* to h on $\mathbf{A} \subseteq \mathbf{X}$ if, for every $\epsilon > 0$, there exists $\mathbf{N} \in \mathcal{N}_\infty$ such that $\|h_i(x) - h(x)\|_2 < \epsilon$ for all $i \in \mathbf{N}$ and all $x \in \mathbf{A}$. This is denoted by $h_i \xrightarrow{\text{unif}} h$ on $\mathbf{A}$.

23. **Integrable functions:** Consider a measure μ and a function h. The function h is called *integrable* with respect to μ, or in short μ-integrable, if it is μ-measurable and $\int h \, d\mu < \infty$.

24. **Continuity of functions over topological spaces:** A function $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are topological spaces, is called *continuous* if for every open subset $\mathbf{V} \subseteq \mathcal{Y}$, the preimage $h^{\text{pre}}(\mathbf{V}) \subseteq \mathcal{X}$ is open. Equivalently, $h^{\text{pre}}(\mathbf{V})$ is closed in $\mathcal{X}$ if $\mathbf{V}$ is closed in $\mathcal{Y}$.

25. **Continuity of functions over Euclidean spaces:** A function $h : \mathbf{X} \to \mathbb{R}^k$, where $\mathbf{X} \subseteq \mathbb{R}^d$, is called *continuous* at $\bar{x} \in \mathbf{X}$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that if $x \in \mathbf{B}(\bar{x}, \delta) \cap \mathbf{X}$, we have $\|h(x) - h(\bar{x})\|_2 < \epsilon$. Equivalently, h is continuous at $\bar{x} \in \mathbf{X}$ if for every sequence $\{x_i\}_{i \in \mathbb{N}}$ in $\mathbf{X}$ with $x_i \to \bar{x}$, we have $h(x_i) \to h(\bar{x})$. Moreover, h is called continuous on $\mathbf{A} \subseteq \mathbf{X}$ if it is continuous at every $\bar{x} \in \mathbf{A}$. Finally, h is called continuous if it is continuous on $\mathbf{X}$.

26. **Lower semi-continuity**: A function $h : \mathbb{R}^d \to \overline{\mathbb{R}}$ is called a *lower semi-continuous* function if for every $x_0 \in \mathbb{R}^d$, we have $\liminf_{x \to x_0} h(x) = h(x_0)$.

27. **Lipschitz continuity:** A function $h : \mathbf{X} \to \mathbb{R}^k$ is called *Lipschitz continuous* if there exists a constant $L \geq 0$ such that for every $x, y \in \mathbf{X}$, we have $\|h(x) - h(y)\|_2 \leq L \|x - y\|_2$. The constant L is called the Lipschitz constant (factor).

28. **Uniform continuity:** A function $h : \mathbf{X} \to \mathbb{R}^k$ is called *uniformly continuous* if for every $\epsilon > 0$, there exists $\delta > 0$ such that for every $x, y \in \mathbf{X}$ with $\|x - y\|_2 < \delta$, we have $\|h(x) - h(y)\|_2 < \epsilon$.

A.1.3 Elementary Facts

In the following, we present some elementary and well-known mathematical facts that are used throughout the proofs. For more details on of these facts, we refer to Munkres (2000); Rudin (1953); Rockafellar and Wets (2009); Lang (2012); Folland (1999); Kato (2013); Halmos (2013).

1. Bolzano-Weierstrass Theorem: Every bounded sequence in $\mathbb{R}^d$, has a cluster point.

2. If $\mathbf{A} \times \mathbf{B}$ is convex then $\mathbf{A}$ and $\mathbf{B}$ are convex.

3. Compactness in $\mathbb{R}^d$ is equivalent to being closed and bounded. If $\mathbf{A}$ and $\mathbf{B}$ are compact sets, then both $\mathbf{A} \times \mathbf{B}$ and $\mathbf{A} \cap \mathbf{B}$ are compact. Additionally, if $\mathbf{C}$ is a closed subset of $\mathbf{A}$, then $\mathbf{C}$ is also compact.

4. A continuous function that is considered over a compact set, attains its maximum and minimum at some points in that compact set, and consequently is bounded.
5. A continuous function over a compact set is uniformly continuous.
6. The image of a compact set under a continuous function is compact.
7. The uniform convergence of a sequence of multi-component functions holds if and only if it holds for each component.
8. If h_i and g_i are sequences of bounded functions with $h_i \xrightarrow{\text{unif}} h$ and $g_i \xrightarrow{\text{unif}} g$, then $h_i + g_i$ and $h_i g_i$ are sequences of bounded functions with $h_i + g_i \xrightarrow{\text{unif}} h + g$ and $h_i g_i \xrightarrow{\text{unif}} hg$.
9. Pointwise maximization of finite number of convex function is convex.
10. Pointwise maximization of finite number of continuous function is continuous.
11. The level set of a convex function with a convex domain is a convex set.
12. The space of real-valued continuous functions over a compact set, equipped with the $\|\cdot\|_\infty$ norm, forms a Banach space, i.e., it is a complete normed space.
13. For a real-valued function h, suppose ∇h exists and is continuous on an open subset $\mathbf{O}$ of the domain of h and $x_0 = \arg\min_{x \in \mathbf{O}} h(x)$. Then $\nabla h(x_0) = 0$.
14. Taylor's First-Order Expansion: For a continuously differentiable function $h : \mathbf{S} \to \mathbb{R}^k$, where $\mathbf{S} \subseteq \mathbb{R}^d$ is open and convex, suppose $x, x_0 \in \mathbf{S}$. Then,

$$h(x) = h(x_0) + \nabla h(x_0)(x - x_0) + \mathcal{O}\left(\|x - x_0\|_2^2\right). \tag{A.6}$$

15. Mean Value Inequality: For a differentiable function $h : \mathbf{S} \to \mathbb{R}^k$, where $\mathbf{S} \subseteq \mathbb{R}^d$ is open and convex, if $\|\nabla h(x)\|_2 \leq L$ for all $x \in \mathbf{S}$, then h is Lipschitz continuous with Lipschitz factor L.
16. For events A and B, if $A \Rightarrow B$, then $\mathbb{P}[A] \leq \mathbb{P}[B]$.
17. Dominated Convergence Theorem: Let $\{h_i\}_{i \in \mathbb{N}}$ be a sequence of real-valued measurable functions on a measure space $(\mathbf{X}, \mathcal{F}, \mu)$ such that $h_i(x) \to h(x)$ holds almost everywhere on $\mathbf{X}$ with respect to μ. Moreover, assume there exists a μ-integrable function $g(x)$ such that $|h_i(x)| \leq g(x)$ for every $i \in \mathbb{N}$ and $x \in \mathbf{X}$. Then h_i and h are μ-integrable and $\int h_i d\mu \to \int h d\mu$.
18. Considering the Lebesgue measure, a countable union of negligible sets is negligible.
19. Given the function h whose domain lies in $\mathbb{R}^{d_1}$ and range in $\mathbb{R}^{d_2}$, the graph of h, i.e., the set $\{(x, y) \mid y = h(x)\}$, has zero Lebesgue measure in $\mathbb{R}^{d_1 + d_2}$.
20. Consider the continuous function $h(x, y)$, whose domain is a compact set, and the measure μ_x that is defined over the domain where x lies. Moreover, assume that μ_x is dominated by the Lebesgue measure. Then the mapping $y \mapsto \int h(x, y)\, d\mu_x$ is continuous.
21. A continuous function over a compact set is integrable with respect to the Lebesgue measure.

A.2 Intermediate Proof Steps

In this section, we present and prove key formal statements that serve as intermediate steps for the results of this paper. Collecting these steps here aims to improve organization and ensure conciseness in the final proofs.

Theorem A.1 (Theorem 7.33 in Rockafellar and Wets (2009)). *Consider $h_i, h : \mathbb{R}^d \to \overline{\mathbb{R}}$, and suppose the sequence of functions $\{h_i\}_{i\in\mathbb{N}}$ is eventually level-bounded, and h_i and h are lower semi-continuous and proper. Moreover, suppose we have $\operatorname{epi} h_i \to \operatorname{epi} h$ (in the Painlevé-Kuratowski sense). Then $\inf h_i \to \inf h$ (finite), while for i in some index set $\mathbf{N} \in \mathcal{N}_\infty$, the sets $\arg\min h_i$ are nonempty and form a bounded sequence, i.e.,*

$$\limsup_{i\to\infty} (\arg\min h_i) \subseteq \arg\min h. \tag{A.7}$$

Lemma A.1. *Consider the sequence of sets $\{\mathbf{A}_i\}_{i\in\mathbb{N}}$, where $\emptyset \neq \mathbf{A}_i \subseteq \mathbf{A}$ for every $i \in \mathbb{N}$ and $\mathbf{A} \subseteq \mathbb{R}^d$ is bounded. Moreover, suppose $\limsup_{i\to\infty} \mathbf{A}_i \subseteq \{a\}$, where the limit is considered in the Painlevé-Kuratowski sense. Then, for a sequence $\{a_i\}_{i\in\mathbb{N}}$ with $a_i \in \mathbf{A}_i$ for each $i \in \mathbb{N}$, we have $a_i \to a$.*

Proof of Lemma A.1. First, observe that $\limsup_{i\to\infty} \mathbf{A}_i \neq \emptyset$ because $\mathbf{A}_i \neq \emptyset$ and thus there exists $a_i \in \mathbf{A}_i \subseteq \mathbf{A}$ for $i \in \mathbb{N}$. Therefore, the sequence $\{a_i\}_{i\in\mathbb{N}}$ is bounded and has a cluster point in $\mathbb{R}^d$ due to the Bolzano-Weierstrass Theorem (see item 17 of Appendix A.1.2 and item 1 of Appendix A.1.3). Note that any cluster point of $\{a_i\}_{i\in\mathbb{N}}$ lies in $\limsup_{i\to\infty} \mathbf{A}_i$ by the definition provided in item 8 of Appendix A.1.2, and thus $\limsup_{i\to\infty} \mathbf{A}_i \neq \emptyset$. Hence, $\limsup_{i\to\infty} \mathbf{A}_i = \{a\}$.

Consider $\{a_i\}_{i\in\mathbb{N}}$ with $a_i \in \mathbf{A}_i$. To prove $a_i \to a$, by contradiction suppose $a_i \nrightarrow a$ (see item 1 of Appendix A.1.2). This means there exists $\epsilon > 0$ and $\mathbf{N} \in \mathcal{N}_\infty^\#$ such that $\|a_k - a\|_2 \geq \epsilon$ for all $k \in \mathbf{N}$.

On the other hand, note that $\{a_k\}_{k\in\mathbf{N}}$ is a bounded sequence and thus has a cluster point in $\mathbb{R}^d$ due to the Bolzano-Weierstrass Theorem. This cluster point lies in $\limsup_{i\to\infty} \mathbf{A}_i = \{a\}$ by definition and is therefore a. This means $\{a_k\}_{k\in\mathbf{N}}$ has a subsequence that converges to a, which contradicts $\|a_k - a\|_2 \geq \epsilon$ for all $k \in \mathbf{N}$, as obtained above. □

Lemma A.2. *Suppose $\mathbf{A} \subseteq \mathbb{R}^d$ is compact and convex, and $\operatorname{int} \mathbf{A} \neq \emptyset$. Then, for every $a \in \mathbf{A} \setminus \operatorname{int} \mathbf{A}$, there exists a sequence $\{a_n\}_{n\in\mathbb{N}}$ in $\operatorname{int} \mathbf{A}$ such that $a_n \to a$.*

Proof of Lemma A.2. Since $\operatorname{int} \mathbf{A} \neq \emptyset$, there exists $a_0 \in \operatorname{int} \mathbf{A}$. Let $e_1, \ldots, e_d$ denote the unit vectors parallel to the d standard axes. Since $\operatorname{int} \mathbf{A}$ is open, there exist numbers $\alpha_k > 0$ for $1 \leq k \leq d$ such that $b_k = a_0 + \alpha_k e_k \in \operatorname{int} \mathbf{A}$ for $1 \leq k \leq d$. Note that $a \neq b_k$ because $b_k \in \operatorname{int} \mathbf{A}$, while $a \in \mathbf{A} \setminus \operatorname{int} \mathbf{A}$. Next, define the following convex cone:

$$\mathbf{C} = \left\{ a + \sum_{j=1}^{d} \frac{t_j}{d}(b_j - a) \;\middle|\; t_j \in [0,1] \text{ for } 1 \leq j \leq d \right\}. \tag{A.8}$$

Due to the convexity of $\mathbf{A}$, we have $\mathbf{C} \subseteq \mathbf{A}$. Therefore, $\operatorname{int} \mathbf{C} \subseteq \operatorname{int} \mathbf{A}$. By letting $t_j = 1/n$, observe that $a_n = a + \sum_{j=1}^{d}(b_j - a)/nd \in \operatorname{int} \mathbf{C} \subseteq \operatorname{int} \mathbf{A}$, and $a_n \to a$. □

Lemma A.3. *(An extension of Rockafellar and Wets (2009, Example 5.10)) Consider nonempty sets $\mathbf{X} \subseteq \mathbb{R}^d$, $\mathbf{A} \subseteq \mathbb{R}^k$, and the functions $g_i : \mathbf{A} \times \mathbf{X} \to \mathbb{R}$ for $1 \leq i \leq N$. Define the set-valued function $\mathbf{P}(x) \subseteq \mathbb{R}^k$ with $x \in \mathbb{R}^d$, where for $x \notin \mathbf{X}$, we have $\mathbf{P}(x) = \emptyset$, and for $x \in \mathbf{X}$:*

$$\mathbf{P}(x) = \{a \in \mathbf{A} \mid g_i(a, x) \leq 0, \quad \textit{for all } 1 \leq i \leq N\}. \tag{A.9}$$

Suppose the following conditions hold:

i. $\mathbf{X}$ *is compact.*

ii. $\mathbf{A}$ *is compact and convex with* $\operatorname{int}(\mathbf{A}) \neq \emptyset$.

iii. g_i *is continuous.*

iv. $g_i(a, x)$ *is convex in* a.

v. For every $x \in \mathbf{X}$*, there exists* $a \in \mathbf{A}$ *such that* $g_i(a, x) < 0$ *for all* $1 \leq i \leq N$.

Then $\mathbf{P}(\cdot)$ *is continuous relative to* $\mathbf{X}$ *(in the Kuratowski sense), and* $\mathbf{P}(x)$ *is a nonempty and closed subset of* $\mathbf{A}$ *for every* $x \in \mathbf{X}$.

Proof of Lemma A.3. Define $g(a, x) = \max(g_1(a, x), \ldots, g_N(a, x))$ and observe that g is continuous and $g(a, x)$ is convex in a (see items 9 and 10 of Appendix A.1.3). Moreover, $g(a, x) \leq 0$ holds if and only if $g_i(a, x) \leq 0$ holds for every $1 \leq i \leq N$, and a similar statement is true for $g(a, x) < 0$. Hence, we can rewrite $\mathbf{P}(x)$ for $x \in \mathbf{X}$ as follows:

$$\mathbf{P}(x) = \{a \in \mathbf{A} \mid g(a, x) \leq 0\}. \tag{A.10}$$

Moreover, note that the graph of $\mathbf{P}$ (defined in item 10 of Appendix A.1.2) is:

$$\operatorname{gph} \mathbf{P} = \{(x, a) \in \mathbb{R}^d \times \mathbb{R}^k \mid a \in \mathbf{P}(x)\} = \{(x, a) \in \mathbf{X} \times \mathbf{A} \mid g(a, x) \leq 0\}. \tag{A.11}$$

To prove that $\mathbf{P}$ is continuous relative to $\mathbf{X}$ (in the Kuratowski sense), it suffices to show that it is both outer and inner semi-continuous relative to $\mathbf{X}$ (see item 9 of Appendix A.1.2). With this in mind, the proof is presented through the following steps:

Step 1. Proving $\mathbf{P}(x)$ is nonempty, convex, and a closed subset of $\mathbf{A}$ for every $x \in \mathbf{X}$:
Suppose $x \in \mathbf{X}$. Then, $\mathbf{P}(x)$ is nonempty due to item (v) of the assumptions. Additionally, $\mathbf{P}(x)$ is the preimage of the closed set $(-\infty, 0]$ under the continuous mapping $a \mapsto g(a, x)$ and is therefore a closed subset of $\mathbf{A}$ (see item 24 of Appendix A.1.2). Moreover, $\mathbf{P}(x)$ is the level set of a convex mapping $a \mapsto g(a, x)$ over the convex domain $\mathbf{A}$, and thus $\mathbf{P}(x)$ is a convex set (see item 11 of Appendix A.1.3).

Step 2. Proving $\mathbf{P}$ is outer semi-continuous relative to $\mathbf{X}$:
$\operatorname{gph} \mathbf{P}$ in (A.11) is the preimage of the closed set $(-\infty, 0]$ under the continuous mapping $(x, a) \mapsto g(a, x)$ and thus is a closed subset of $\mathbf{X} \times \mathbf{A}$. Since $\mathbf{X} \times \mathbf{A}$ is compact, $\operatorname{gph} \mathbf{P}$ is compact (see item 3 of Appendix A.1.3).

From Step 1, $\mathbf{P}(x)$ is closed in the compact set $\mathbf{A}$ and is thus compact and consequently closed in $\mathbb{R}^k$. Therefore, $\mathbf{P}$ is a closed-valued mapping. According to Rockafellar and Wets (2009, Theorem 5.7(b)), a closed-valued mapping $\mathbf{P}$ is outer semi-continuous relative to $\mathbf{X}$ if and only if, for every compact set $\mathbf{B} \subseteq \mathbb{R}^k$, its inverse image $\mathbf{P}^{-1}(\mathbf{B})$ (defined in item 11 of Appendix A.1.2) is a closed subset of $\mathbf{X}$.

As a result, it suffices to prove that $\mathbf{P}^{-1}(\mathbf{B})$ is compact. To this end, let $\operatorname{Proj} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ denote the projection mapping of (x, a) onto x, i.e., $\operatorname{Proj}(x, a) = x$. Using this notation, we have:

$$\mathbf{P}^{-1}(\mathbf{B}) = \operatorname{Proj}\left(\operatorname{gph} \mathbf{P} \cap \mathbf{X} \times \mathbf{B}\right). \tag{A.12}$$

Note that $\operatorname{gph}\mathbf{P} \cap \mathbf{X} \times \mathbf{B}$ is compact because both $\mathbf{X} \times \mathbf{B}$ and $\operatorname{gph}\mathbf{P}$ are compact. Moreover, projection is a continuous mapping. Therefore, $\mathbf{P}^{-1}(\mathbf{B})$ in (A.12) is the image of a compact set under a continuous mapping and is thus compact (see item 6 of Appendix A.1.3).

Step 3. Finding $\operatorname{int}\mathbf{P}(\bar{x})$ and proving $\operatorname{int}\mathbf{P}(\bar{x}) \neq \emptyset$ for all $\bar{x} \in \mathbf{X}$:
Suppose $\bar{x} \in \mathbf{X}$. Then, the following holds due to the continuity of the mapping $a \mapsto g(a, \bar{x})$:

$$\operatorname{int}\mathbf{P}(\bar{x}) = \{a \in \operatorname{int}\mathbf{A} \mid g(a,\bar{x}) < 0\} \cup \operatorname{int}\left(\{a \in \mathbf{A} \mid g(a,\bar{x}) = 0\}\right) = \{a \in \operatorname{int}\mathbf{A} \mid g(a,\bar{x}) < 0\}. \tag{A.13}$$

Note that the rightmost equality in (A.13) holds because $\operatorname{int}\{a \in \mathbf{A} \mid g(a,\bar{x}) = 0\} = \emptyset$. To see why this is true, suppose by contradiction that there exists an open set $\mathbf{O} \subseteq \mathbf{A}$ such that $g(a,\bar{x}) = 0$ for all $a \in \mathbf{O}$. Let $a_1 \in \mathbf{O}$. From the assumptions, there is also $a_2 \in \mathbf{A}$ such that $g(a_2,\bar{x}) < 0$. Furthermore, there exists $0 < \lambda < 1$ such that the point $a_0 = \lambda a_1 + (1-\lambda)a_2$ also lies in $\mathbf{O}$. Since $a_0 \in \mathbf{O}$, we have $g(a_0,\bar{x}) = 0$. However, this contradicts the fact that, due to the convexity of $g(\cdot,\bar{x})$, we have $g(a_0,\bar{x}) \leq \lambda g(a_1,\bar{x}) + (1-\lambda)g(a_2,\bar{x}) = (1-\lambda)g(a_2,\bar{x}) < 0$. Hence, (A.13) holds.

Due to item (v) of the assumptions, we know there exists $\hat{a} \in \mathbf{A}$ such that $g(\hat{a},\bar{x}) < 0$. Based on this, we claim that $\operatorname{int}\mathbf{P}(\bar{x}) \neq \emptyset$, i.e., there exists $a \in \operatorname{int}\mathbf{A}$ such that $g(a,\bar{x}) < 0$. Suppose, by contradiction, that for every $a \in \operatorname{int}\mathbf{A}$, we have $g(a,\bar{x}) \geq 0$. This would imply that $\hat{a} \in \mathbf{A} \setminus \operatorname{int}\mathbf{A}$. Since $\mathbf{A}$ is compact and convex with $\operatorname{int}\mathbf{A} \neq \emptyset$, we can apply Lemma A.2 and conclude that there exists a sequence $\{a_n\}_{n\in\mathbb{N}}$ in $\operatorname{int}\mathbf{A}$ such that $a_n \to \hat{a}$. As a result, $g(a_n,\bar{x}) \geq 0$ for every $n \in \mathbb{N}$. Since g is continuous, $g(a_n,\bar{x}) \to g(\hat{a},\bar{x})$, and thus, from $g(a_n,\bar{x}) \geq 0$, we conclude that $g(\hat{a},\bar{x}) \geq 0$. This contradicts the assumption that $g(\hat{a},\bar{x}) < 0$. Hence, the claim is proven, i.e., $\operatorname{int}\mathbf{P}(\bar{x}) = \{a \in \operatorname{int}\mathbf{A} \mid g(a,\bar{x}) < 0\} \neq \emptyset$.

Step 4. Proving $\mathbf{P}$ is inner semi-continuous relative to $\mathbf{X}$:
Rockafellar and Wets (2009, Theorem 5.9 (a)) states that, given $\mathbf{P}$ is convex-valued (shown in Step 1), to prove that $\mathbf{P}$ is inner semi-continuous relative to $\mathbf{X}$, it suffices to show that for every $\bar{x} \in \mathbf{X}$, we have $\operatorname{int}\mathbf{P}(\bar{x}) \neq \emptyset$ (shown in Step 3) and for all $\bar{a} \in \operatorname{int}\mathbf{P}(\bar{x})$, there exists an open neighborhood $\mathbf{W}$ around $(\bar{x},\bar{a})$ such that $\mathbf{W} \cap (\mathbf{X} \times \mathbb{R}^k) \subseteq \operatorname{gph}\mathbf{P}$.

Let $\bar{a} \in \operatorname{int}\mathbf{P}(\bar{x})$, with $\operatorname{int}\mathbf{P}(\bar{x})$ obtained in (A.13). Thus, $g(\bar{a},\bar{x}) < 0$. Since $(x,a) \mapsto g(a,x)$ is a continuous mapping over $\mathbf{X} \times \operatorname{int}\mathbf{P}(\bar{x})$, there exists an open neighborhood $\mathbf{W} \subseteq \mathbb{R}^d \times \operatorname{int}\mathbf{P}(\bar{x})$ around $(\bar{x},\bar{a})$ such that $g(a,x) < 0$ for all $(x,a) \in \mathbf{W} \cap (\mathbf{X} \times \operatorname{int}\mathbf{P}(\bar{x}))$. Observe that $\mathbf{W} \cap (\mathbf{X} \times \operatorname{int}\mathbf{P}(\bar{x})) = \mathbf{W} \cap (\mathbf{X} \times \mathbb{R}^k)$. Therefore, we have shown that for all $(x,a) \in \mathbf{W} \cap (\mathbf{X} \times \mathbb{R}^k)$, we have $g(a,x) < 0$, and thus $(x,a) \in \operatorname{gph}\mathbf{P}$. Hence, $\mathbf{W} \cap (\mathbf{X} \times \mathbb{R}^k) \subseteq \operatorname{gph}\mathbf{P}$. □

Lemma A.4. *Consider the function $f : \mathbf{A} \times \mathbf{X} \to \mathbb{R}$ and the following maximization problem:*

$$v(x) = \sup_{a \in \mathbf{Q}(x)} f(a,x), \tag{A.14}$$

where $\mathbf{Q}(x) \subseteq \mathbf{A}$ is a set-valued function with $\mathbf{Q}(x) = \emptyset$ if $x \notin \mathbf{X}$. Suppose the following conditions hold:

i. $\mathbf{A}$ *and* $\mathbf{X}$ *are nonempty and compact.*

ii. f *is continuous.*

iii. $\mathbf{Q}(\cdot)$ *is continuous relative to* $\mathbf{X}$ *(in the Kuratowski sense).*

iv. $\mathbf{Q}(x)$ *is a nonempty and closed subset of* $\mathbf{A}$ *for every* $x \in \mathbf{X}$.

Then, $v(\cdot)$ is continuous, and the problem (A.14) attains its optimal solution at some $a^*(x) \in \mathbf{Q}(x)$. Moreover, if $a^*(x)$ is unique for every $x \in \mathbf{X}$, then $a^*(\cdot)$ is continuous.

Proof of Lemma A.4. Note that $\mathbf{Q}(x)$ is a nonempty compact set (see item 3 of Appendix A.1.3), and thus problem (A.14) is a maximization of a continuous function over a compact set. Therefore, $v(x)$ is finite, and this optimal value is attained at some $a^*(x) \in \mathbf{Q}(x)$ (see item 4 of Appendix A.1.3). Hence, it remains to prove the continuity of $v(\cdot)$ and $a^*(\cdot)$. To this end, let $\mathbf{A} \subseteq \mathbb{R}^{d_1}$ and $\mathbf{X} \subseteq \mathbb{R}^{d_2}$. Denote by $\delta : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \overline{\mathbb{R}}$ the indicator function of $\mathbf{Q}(x)$ similar to Rockafellar and Wets (2009, Chapter 1), i.e., for $a \in \mathbb{R}^{d_1}$ and $x \in \mathbb{R}^{d_2}$, define:

$$\delta(a, x) = \begin{cases} 0 & \text{if } x \in \mathbf{X} \text{ and } a \in \mathbf{Q}(x), \\ \infty & \text{otherwise.} \end{cases} \tag{A.15}$$

Moreover, consider an extension of $-f(\cdot)$ to the function $h : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \overline{\mathbb{R}}$, where:

$$h(a, x) = \begin{cases} -f(a, x) & \text{if } (a, x) \in \mathbf{A} \times \mathbf{X}, \\ \infty & \text{if } (a, x) \notin \mathbf{A} \times \mathbf{X}. \end{cases} \tag{A.16}$$

Using (A.15) and (A.16), we can rewrite the problem in (A.14) as an equivalent unconstrained minimization problem as follows:

$$v(x) = \inf_{a \in \mathbb{R}^{d_1}} \left[h(a, x) + \delta(a, x) \right], \quad a^*(x) \in \arg\min_{a \in \mathbb{R}^{d_1}} \left[h(a, x) + \delta(a, x) \right]. \tag{A.17}$$

Consider a sequence $\{x_i\}_{i \in \mathbb{N}}$, with $x_i \to x_0$, where $x_i \in \mathbf{X}$ for $i \geq 0$, and let $a_i^* = a^*(x_i)$, $v_i = v(x_i)$, $h_i(a) = h(a, x_i)$, $\delta_i(a) = \delta(a, x_i)$, and $f_i(a) = f(a, x_i)$ for $i \geq 0$. To demonstrate the continuity of $v(\cdot)$ and $a^*(\cdot)$, it suffices to show that $v(x_i) \to v(x_0)$ and $a^*(x_i) \to a^*(x_0)$, i.e., $v_i \to v_0$ and $a_i^* \to a_0^*$ (see item 25 of Appendix A.1.2). We will show this by applying Theorem A.1 through the following steps:

Step 1. Proving level-boundedness, properness, and lower semi-continuity of $h_i + \delta_i$, $i \geq 0$:
Observe that δ_i for $i \geq 0$ is proper if and only if $\mathbf{Q}(x_i)$ is nonempty, level-bounded if and only if $\mathbf{Q}(x_i)$ is bounded, and lower semi-continuous if and only if $\mathbf{Q}(x_i)$ is closed. Since $\mathbf{A}$ is compact and $\mathbf{Q}(x_i)$ is a nonempty, closed subset of $\mathbf{A}$ due to the assumptions, $\mathbf{Q}(x_i)$ is nonempty, closed, and bounded.

Moreover, note that h_i for $i \geq 0$ is level-bounded because $\mathbf{A}$ is bounded, lower semi-continuous because f_i is continuous, and proper because $\mathbf{A}$ is nonempty.

Given this, we can first conclude that $h_i + \delta_i$ for $i \geq 0$ is proper and level-bounded, which results in the sequence $\{h_i + \delta_i\}_{i \in \mathbb{N}}$ being eventually level-bounded. Additionally, $h_i + \delta_i$ for $i \geq 0$ is lower semi-continuous due to the additive property of lower semi-continuity for proper functions (Rockafellar and Wets, 2009, Proposition 1.39).

Step 2. Proving $\operatorname{epi}(h_i + \delta_i) \to \operatorname{epi}(h_0 + \delta_0)$:
We prove this directly through the definition of Painlevé-Kuratowski convergence of sets, as mentioned

in item 8 of Appendix A.1.2. To this end, it suffices to show both of the following inclusions:

$$\limsup_{i\to\infty} \left(\mathrm{epi}(h_i+\delta_i)\right) \subseteq \mathrm{epi}(h_0+\delta_0), \tag{A.18}$$

$$\mathrm{epi}(h_0+\delta_0) \subseteq \liminf_{i\to\infty} \left(\mathrm{epi}(h_i+\delta_i)\right). \tag{A.19}$$

We have for $i \geq 0$:

$$(a,\beta) \in \mathrm{epi}(h_i+\delta_i) \quad \Leftrightarrow \quad (a,\beta) \in \mathrm{epi}\, h_i,\ a \in \mathbf{Q}(x_i). \tag{A.20}$$

Moreover, note that due to the continuity of $\mathbf{Q}(\cdot)$ in the Kuratowski sense (item 9 of Appendix A.1.2), we have $\mathbf{Q}(x_i) \to \mathbf{Q}(x_0)$ (in the Painlevé-Kuratowski sense). This means:

$$\limsup_{i\to\infty} \mathbf{Q}(x_i) \subseteq \mathbf{Q}(x_0), \tag{A.21}$$

$$\mathbf{Q}(x_0) \subseteq \liminf_{i\to\infty} \mathbf{Q}(x_i). \tag{A.22}$$

Step 2–Part 1. Proving (A.18):
To prove (A.18), assuming $(a,\beta) \in \limsup_{i\to\infty} (\mathrm{epi}(h_i+\delta_i))$, we show $(a,\beta) \in \mathrm{epi}(h_0+\delta_0)$. Suppose $(a,\beta) \in \limsup_{i\to\infty} (\mathrm{epi}(h_i+\delta_i))$. Therefore, there exists $\mathbf{N} \in \mathcal{N}_\infty^\#$ and $(a_k,\beta_k) \in \mathrm{epi}(h_k+\delta_k)$ for $k \in \mathbf{N}$ such that $(a_k,\beta_k) \to (a,\beta)$. From (A.20), we have $(a_k,\beta_k) \in \mathrm{epi}\, h_k$ and $a_k \in \mathbf{Q}(x_k)$ for $k \in \mathbf{N}$. Hence, we found $\mathbf{N} \in \mathcal{N}_\infty^\#$ and a sequence $\{a_k\}_{k\in\mathbf{N}}$ such that $a_k \in \mathbf{Q}(x_k)$ for $k \in \mathbf{N}$ and $a_k \to a$. This means $a \in \limsup_{i\to\infty} \mathbf{Q}(x_i)$, and due to (A.21), we conclude that $a \in \mathbf{Q}(x_0)$.

Next, recall that $(a_k,\beta_k) \in \mathrm{epi}\, h_k$, which means $\beta_k \geq h_k(a_k)$. Hence,

$$\lim_{k\to\infty} \beta_k \geq \lim_{k\to\infty} h_k(a_k) \quad \Rightarrow \quad \beta \geq \lim_{k\to\infty} h_k(a_k) = \lim_{k\to\infty} h(a_k,x_k) \overset{\text{E0}}{=} h(a,x_0) = h_0(a), \tag{A.23}$$

where the equality E0 in (A.23) holds because $(a_k,x_k) \to (a,x_0)$ and $h(\cdot)$ is continuous on $\mathbf{A}\times\mathbf{X}$. Having $\beta \geq h_0(a)$ from (A.23) means $(a,\beta) \in \mathrm{epi}\, h_0$. As a result, we have shown $a \in \mathbf{Q}(x_0)$ and $(a,\beta) \in \mathrm{epi}\, h_0$. By applying (A.20), we have $(a,\beta) \in \mathrm{epi}(h_0+\delta_0)$. Hence, (A.18) is proven.

Step 2–Part 2. Proving (A.19):
To prove (A.19), assuming $(a,\beta) \in \mathrm{epi}(h_0+\delta_0)$, we show $(a,\beta) \in \liminf_{i\to\infty} (\mathrm{epi}(h_i+\delta_i))$. Suppose $(a,\beta) \in \mathrm{epi}(h_0+\delta_0)$. Therefore, from (A.20) we have $(a,\beta) \in \mathrm{epi}\, h_0$ and $a \in \mathbf{Q}(x_0)$. Due to (A.22), there exists $\mathbf{N}' \in \mathcal{N}_\infty$ and $a'_k \in \mathbf{Q}(x_k)$ for $k \in \mathbf{N}'$ such that $a'_k \to a$. Letting $\beta'_k = \max(\beta, h_k(a'_k))$, we can write:

$$\lim_{k\to\infty} \beta'_k = \max(\beta, \lim_{k\to\infty} h_k(a'_k)) = \max(\beta, \lim_{k\to\infty} h(a'_k,x_k)) \overset{\text{E1}}{=} \max(\beta, h(a,x_0)) = \max(\beta, h_0(a)) \overset{\text{E2}}{=} \beta, \tag{A.24}$$

where the equality E1 in (A.24) holds because $(a'_k,x_k) \to (a,x_0)$ and $h(\cdot)$ is continuous on $\mathbf{A}\times\mathbf{X}$. Additionally, the equality E2 holds because $(a,\beta) \in \mathrm{epi}\, h_0$, i.e., $\beta \geq h_0(a)$. Moreover, notice that $\beta'_k \geq h_k(a'_k)$ and thus $(a'_k,\beta'_k) \in \mathrm{epi}\, h_k$. Recall that $a'_k \in \mathbf{Q}(x_k)$. Applying (A.20) then results in $(a'_k,\beta'_k) \in \mathrm{epi}(h_k+\delta_k)$. Hence, we found $\mathbf{N}' \in \mathcal{N}_\infty$ and a sequence $\{(a'_k,\beta'_k)\}_{k\in\mathbf{N}'}$ such that $(a'_k,\beta'_k) \in \mathrm{epi}(h_k+\delta_k)$ for $k \in \mathbf{N}'$ and $(a'_k,\beta'_k) \to (a,\beta)$. This means $(a,\beta) \in \liminf_{i\to\infty} (\mathrm{epi}(h_i+\delta_i))$. Hence, (A.19) is proven.

Step 3. Proving $v_i \to v_0$ and $a_i^* \to a_0^*$:
Through Steps 1 and 2, we have shown that the sequence of functions $\{h_i+\delta_i\}_{i\in\mathbb{N}}$ and its limit function

$h_0 + \delta_0$ satisfy the conditions of Theorem A.1. Hence, we conclude that $v_i \to v_0$ and:

$$\limsup_{i\to\infty} (\arg\min(h_i + \delta_i)) \subseteq \arg\min(h_0 + \delta_0). \tag{A.25}$$

Given the uniqueness of $a^*(x)$ for every $x \in \mathbf{X}$, we have $\arg\min(h_i + \delta_i) = \{a_i^*\}$ for $i \geq 0$, and thus $\limsup_{i\to\infty}\{a_i^*\} \subseteq \{a_0^*\}$. This leads to $a_i^* \to a_0^*$ due to Lemma A.1. □

Lemma A.5. *Let $\mathbf{P} = [a_1, b_1] \times \ldots \times [a_K, b_K] \subset \mathbb{R}^K$, with $a_i < b_i < \infty$ for $1 \leq i \leq K$. Then, there exists a constant $\xi(\mathbf{P}) > 0$ such that for all $x \in \mathbf{P}$ and all $0 \leq r \leq \operatorname{diam}(\mathbf{P})$, we have:*

$$\operatorname{vol}\left(\mathbf{B}(x, r) \cap \mathbf{P}\right) \geq \xi(\mathbf{P}) \cdot \operatorname{vol}\left(\mathbf{B}(x, r)\right). \tag{A.26}$$

Proof of Lemma A.5. Suppose $r < r_0$ for sufficiently small r_0, e.g., let $r_0 = \frac{1}{4}\min_{1\leq i\leq K}(b_i - a_i)$. Then it is straightforward to verify that the intersection $\operatorname{vol}(\mathbf{B}(x, r) \cap \mathbf{P})$ attains its minimum value when x is a vertex of the box $\mathbf{P}$, and this minimum value is $2^{-K} \cdot \operatorname{vol}(\mathbf{B}(x, r))$ at any vertex. Hence, if $r < r_0$, for all $x \in \mathbf{P}$, we have:

$$\operatorname{vol}\left(\mathbf{B}(x, r) \cap \mathbf{P}\right) \geq \frac{1}{2^K} \cdot \operatorname{vol}\left(\mathbf{B}(x, r)\right). \tag{A.27}$$

Now let:

$$f(x, r) = \frac{\operatorname{vol}\left(\mathbf{B}(x, r) \cap \mathbf{P}\right)}{\operatorname{vol}\left(\mathbf{B}(x, r)\right)}, \tag{A.28}$$

and define $\xi(\mathbf{P})$ through the following minimization problem:

$$\begin{aligned} \xi(\mathbf{P}) = \inf \quad & f(x, r) \\ \text{s.t.} \quad & x \in \mathbf{P} \\ & r \in [r_0, \operatorname{diam}(\mathbf{P})]. \end{aligned} \tag{A.29}$$

Note that f is continuous, and thus (A.29) is a minimization of a continuous function over the compact set $\mathbf{F} = \mathbf{P} \times [r_0, \operatorname{diam}(\mathbf{P})]$, and therefore attains its optimal solution at some point $(\bar{x}, \bar{r}) \in \mathbf{F}$ (see items 3 and 4 of Appendix A.1.3). Moreover, observe that $f(x, r)$ returns positive values on $\mathbf{F}$, and thus $\xi(\mathbf{P}) = f(\bar{x}, \bar{r}) > 0$. Letting x_0 be a vertex of $\mathbf{P}$, note that $(x_0, r_0) \in \mathbf{F}$, which results in:

$$\xi(\mathbf{P}) \leq f(x_0, r_0) = \frac{1}{2^K}. \tag{A.30}$$

Given $x \in \mathbf{P}$, due to (A.29), Equation (A.26) holds for $r_0 \leq r \leq \operatorname{diam}(\mathbf{P})$, and due to (A.27) and (A.30), Equation (A.26) holds for $r < r_0$. Therefore, Equation (A.26) holds for all $x \in \mathbf{P}$ and all $r \leq \operatorname{diam}(\mathbf{P})$. □

Definition A.1 (Negligible discontinuity)**.** *Consider the function $h(x, y)$ defined over $\mathbf{X} \times \mathbf{Y} \subseteq \mathbb{R}^d \times \mathbb{R}^k$ and let $\mathbf{D}$ be the set of all discontinuity points of h, i.e.,*

$$\mathbf{D} = \{(x, y) \in \mathbf{X} \times \mathbf{Y} \mid h \text{ is discontinuous at } (x, y)\}. \tag{A.31}$$

Moreover, let $\mathbf{D}_{y_0}$ be the cross-section of $\mathbf{D}$ with the space $y = y_0$, that is:

$$\mathbf{D}_{y_0} = \{x \in \mathbf{X} \mid (x, y_0) \in \mathbf{D}\} \subseteq \mathbb{R}^d. \tag{A.32}$$

Then, $h(\cdot)$ is called negligibly discontinuous *over y (i.e., over its second input) if, for every $y_0 \in \mathbf{Y}$, the*

set $\mathbf{D}_{y_0}$ is ν_d-negligible, i.e., $\nu_d(\mathbf{D}_{y_0}) = 0$, where ν_d is the Lebesgue measure over $\mathbb{R}^d$.

Lemma A.6. *Consider the function $h(x, y)$ defined over the convex set $\mathbf{X} \times \mathbf{Y} \subseteq \mathbb{R}^d \times \mathbb{R}^k$, whose outputs lie in $\mathbb{R}^s$, and the measure μ_x defined over $\mathbf{X}$ with $\mu_x \ll \nu_d$, where ν_d is the restriction of the Lebesgue measure over $\mathbb{R}^d$ on $\mathbf{X}$. Moreover, assume there exists a real-valued function g defined over $\mathbf{X}$ such that $\|h(x, y)\|_2 \leq g(x)$ for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$, where $h(x, y)$ and $g(x)$ are measurable functions of x, and $\int g(x)\, d\mu_x < \infty$. Finally, suppose h is negligibly discontinuous over y. Then, $f(y) = \int h(x, y)\, d\mu_x$ is continuous.*

Proof of Lemma A.6. Note that the convexity of $\mathbf{X} \times \mathbf{Y}$ leads to the convexity of $\mathbf{X}$ (see item 2 of Appendix A.1.3), and thus $\mathbf{X}$ is Lebesgue measurable (Lang, 1986). As a result, the restriction of the Lebesgue measure over $\mathbb{R}^d$ on $\mathbf{X}$, namely ν_d, is well-defined. Moreover, since $\mu_x \ll \nu_d$, the measurability of a set or function with respect to μ_x and ν_d coincides. We proceed through the following steps:

Step 1. Proving for the case $s = 1$:
With $s = 1$, meaning $h(x, y)$ is a real-valued function, suppose a sequence $\{y_i\}_{i\in\mathbb{N}}$ in $\mathbf{Y}$ with $y_i \to \bar{y}$ and $\bar{y} \in \mathbf{Y}$ is given. To prove that $f(y)$ is continuous, it suffices to show that $f(y_i) \to f(\bar{y})$ (see item 25 of Appendix A.1.2).

Let $h_i(x) = h(x, y_i)$ and $\bar{h}(x) = h(x, \bar{y})$. Therefore, $h_i(\cdot)$ and $\bar{h}(\cdot)$ are real-valued functions on $\mathbf{X}$. Next, we claim that:

$$h_i(x) \longrightarrow \bar{h}(x), \quad \text{almost everywhere on } \mathbf{X} \text{ with respect to } \mu_x, \tag{A.33}$$

Suppose (A.33) holds. Additionally, note that $|h_i(x)| \leq g(x)$ for all $i \in \mathbb{N}$ due to the assumptions. As a result, $\int h_i(x)\, d\mu_x \longrightarrow \int \bar{h}(x)\, d\mu_x$ holds by the Dominated Convergence Theorem (see item 17 of Appendix A.1.3). This implies that $f(y_i) \to f(\bar{y})$.

Hence, it remains to prove (A.33), which is done as follows: Using the notation of Definition A.1, let $\mathbf{D}$ denote the set of discontinuity points of $h(\cdot)$ and consider the cross-section of $\mathbf{D}$ with $y = \bar{y}$, denoted by $\mathbf{D}_{\bar{y}}$. Letting $x \in \mathbf{X} \setminus \mathbf{D}_{\bar{y}}$, the definition of $\mathbf{D}_{\bar{y}}$ implies that h is continuous at $(x, \bar{y})$. Recall that $(x, y_i) \to (x, \bar{y})$. This leads to $h(x, y_i) \to h(x, \bar{y})$, or equivalently $h_i(x) \to \bar{h}(x)$, due to the continuity of $h(\cdot)$ at $(x, \bar{y})$. Hence, we have shown that $h_i(x) \longrightarrow \bar{h}(x)$ holds for $x \in \mathbf{X} \setminus \mathbf{D}_{\bar{y}}$.

To complete the proof of (A.33), it remains to show that $\mu_x(\mathbf{D}_{\bar{y}}) = 0$. Note that $\nu_d(\mathbf{D}_{\bar{y}}) = 0$ due to the negligible discontinuity of h over y. Given that $\mu_x \ll \nu_d$, we conclude that $\mu_x(\mathbf{D}_{\bar{y}}) = 0$ (see item 19 of Appendix A.1.2).

Step 2. Proving for the case $s > 1$:
For $1 \leq r \leq s$, let h^r and f^r be the r-th component of the functions h and f. Moreover, observe that $f^r(y) = \int h^r(x, y)\, d\mu_x$ and $|h^r(x, y)| \leq \|h(x, y)\|_2 \leq g(x)$ for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$. Additionally, the set of discontinuity points of h^r is a subset of the set of discontinuity points of h, which results in h^r being negligibly discontinuous over y. Therefore, we can apply Step 1 to each component and conclude that f^r is continuous for every $1 \leq r \leq s$. Hence, f is continuous. □

A.3 Proof of Theorem 1

Proof. We proceed through the following steps:

Step 1. Converting the problem into an equivalent unconstrained one:
Define:

$$\Theta_n = \underset{\theta \in \mathbf{P}_\theta}{\arg\min} \, (m - f_n(\theta))^\top W (m - f_n(\theta)), \tag{A.34}$$

where Θ_n is the set of all solutions to the minimization problem in (A.34). By arbitrarily picking $\hat{\theta}_n \in \Theta_n$, the goal is to show that $\hat{\theta}_n \to \theta^*$. Our approach is to apply Theorem A.1. To this end, we first introduce a reformulation of (A.34) that fits Theorem A.1, where the domain of the functions is extended to the whole space, i.e., the corresponding minimization problem is unconstrained. Having $\mathbf{P}_\theta \subseteq \mathbb{R}^K$, we use function extension to convert (A.34) into an equivalent unconstrained minimization by considering $h_n : \mathbb{R}^K \to \overline{\mathbb{R}}$ such that:

$$h_n(\theta) = \begin{cases} (m - f_n(\theta))^\top W (m - f_n(\theta)) & \text{if } \theta \in \mathbf{P}_\theta, \\ \infty & \text{if } \theta \notin \mathbf{P}_\theta. \end{cases} \tag{A.35}$$

As a result, using (A.35) together with the notation introduced in item 3 of Appendix A.1.2, the minimization problem (A.34) can be re-written as:

$$\Theta_n = \arg\min h_n. \tag{A.36}$$

Similarly, define $h(\theta) = (m - f(\theta))^\top W (m - f(\theta))$ if $\theta \in \mathbf{P}_\theta$, and $h(\theta) = \infty$ if $\theta \notin \mathbf{P}_\theta$. Having defined h_n and h as outlined above, we proceed with the following steps to verify the conditions of Theorem A.1 and apply it to complete the proof.

Step 2. Proving properness, level-boundedness, and lower semi-continuity of h_n and h:
We first refer to items 5, 7, and 26 of Appendix A.1.2 for the corresponding definitions. Observe that the functions h_n and h are proper because $\mathbf{P}_\theta$ is nonempty. Additionally, h_n is level-bounded due to the boundedness of $\mathbf{P}_\theta$. Therefore, the sequence $\{h_n\}_{n \in \mathbb{N}}$ is eventually level-bounded. Finally, h_n and h are lower semi-continuous due to the continuity of f_n and f on $\mathbf{P}_\theta$, respectively.

Step 3. Proving that $h_n \xrightarrow{\text{unif}} h$ on $\mathbf{P}_\theta$:
We first refer to item 22 of Appendix A.1.2 for the definition of uniform convergence. Next, denote the ij-th element of W by W_{ij} and the r-th component of m, f_n, and f by m^r, f_n^r, and f^r, respectively. Now, letting $\theta \in \mathbf{P}_\theta$, we can write:

$$h_n(\theta) = (m - f_n(\theta))^\top W (m - f_n(\theta)) = \sum_{i,j \in \{1,\ldots,M\}} W_{ij} (m^i - f_n^i(\theta))(m^j - f_n^j(\theta)). \tag{A.37}$$

Note that f_n^r is a continuous function over the compact set $\mathbf{P}_\theta$ and thus is bounded (see item 4 of Appendix A.1.3). Moreover, $f_n \xrightarrow{\text{unif}} f$ on $\mathbf{P}_\theta$ implies that the corresponding component functions also exhibit uniform convergence, i.e., $f_n^r \xrightarrow{\text{unif}} f^r$ on $\mathbf{P}_\theta$ for every $1 \leq r \leq M$ (see item 7 of Appendix A.1.3). Therefore, $h_n(\theta)$ in (A.37) can be constructed step by step, where each step involves a summation or a

product of two uniformly convergent sequences of bounded functions, resulting in a sequence of functions with the same properties (see item 8 of Appendix A.1.3). Hence, $h_n \xrightarrow{\text{unif}} h$ on $\mathbf{P}_\theta$.

Step 4. Proving $\operatorname{epi} h_n \to \operatorname{epi} h$:
The uniform convergence $h_n \xrightarrow{\text{unif}} h$ on $\mathbf{P}_\theta$ from Step 3, combined with the lower semi-continuity of h_n from Step 2, results in the epigraphical convergence of h_n to h on $\mathbf{P}_\theta$, by Rockafellar and Wets (2009, Proposition 7.15). Note that every point (θ, α) in either $\operatorname{epi} h_n$ or $\operatorname{epi} h$ satisfies $\theta \in \mathbf{P}_\theta$. Thus, $\operatorname{epi} h_n \to \operatorname{epi} h$ (in the Painlevé-Kuratowski sense).

Step 5. Proving $\theta_n \to \theta^*$:
Due to Steps 1, 2, and 4, we have shown that the sequence of functions $\{h_n\}_{n\in\mathbb{N}}$ and its limit function h satisfy the conditions of Theorem A.1. By applying Theorem A.1, we conclude that Θ_n is nonempty, and $\limsup_{n\to\infty} \Theta_n \subseteq \arg\min h$. Moreover, given the uniqueness of θ^* satisfying $m = f(\theta^*)$ and the positive-definiteness of W, we conclude that θ^* is the unique minimizer of h, i.e., $\arg\min h = \{\theta^*\}$. Hence, $\limsup_{n\to\infty} \Theta_n \subseteq \{\theta^*\}$. Since $\hat{\theta}_n \in \Theta_n \subseteq \mathbf{P}_\theta$ and $\mathbf{P}_\theta$ is bounded, this implies $\hat{\theta}_n \to \theta^*$ by Lemma A.1. □

A.4 Proof of Proposition 1

Proof. Since $\mathbf{O}$ is open, there exists an open ball $\mathbf{B}$ centered at θ^* such that $\mathbf{B} \subseteq \mathbf{O}$. Next, Theorem 1 implies that $\hat{\theta}_n \to \theta^*$. Therefore, there exists N_1 such that $\hat{\theta}_n \in \mathbf{B}$ for $n \geq N_1$. For the rest of the proof, suppose $n \geq \max(N_0, N_1)$. Since ∇f_n exists and is continuous on $\mathbf{B} \subseteq \mathbf{O}$, the derivative of the objective function of the minimization problem in (2) exists and is continuous on $\mathbf{B}$, and thus it vanishes at the corresponding minimizer $\hat{\theta}_n \in \mathbf{B}$ (see item 13 of Appendix A.1.3). Therefore,

$$\nabla f_n(\hat{\theta}_n)^\top W \left(f(\theta^*) - f_n(\hat{\theta}_n)\right) = 0. \tag{A.38}$$

Moreover, since $\hat{\theta}_n \to \theta^*$ and ∇f exists and is continuous on an open and convex set $\mathbf{B} \subseteq \mathbf{P}_\theta$ containing $\hat{\theta}_n$, we can write the following first-order Taylor expansion of f in the vicinity of $\hat{\theta}_n$ (see item 14 of Appendix A.1.3):

$$f\left(\theta^*\right) = f(\hat{\theta}_n) - \nabla f(\hat{\theta}_n)(\hat{\theta}_n - \theta^*) + \mathcal{R}, \tag{A.39}$$

where $\mathcal{R} = \mathcal{O}(\|\hat{\theta}_n - \theta^*\|_2^2)$. Next, replacing $f\left(\theta^*\right)$ from (A.39) into (A.38) results in:

$$\nabla f_n(\hat{\theta}_n)^\top W \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n) - \nabla f(\hat{\theta}_n)(\hat{\theta}_n - \theta^*) + \mathcal{R}\right) = 0. \tag{A.40}$$

Re-arranging the terms in (A.40) leads to:

$$\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)\right)\left(\hat{\theta}_n - \theta^*\right) = \nabla f_n(\hat{\theta}_n)^\top W \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n) + \mathcal{R}\right). \tag{A.41}$$

The fact that Λ_n exists means that $\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)$ is invertible. This allows us to simplify (A.41) into the following:

$$\hat{\theta}_n - \theta^* = \Lambda_n \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right) + \Lambda_n \mathcal{R}. \tag{A.42}$$

Finally, having $\|\Lambda_n\|_2 \leq L_0$ for all $n \geq N_0$, we conclude that $\Lambda_n \mathcal{R} = \mathcal{O}(\|\hat{\theta}_n - \theta^*\|_2^2)$. □

A.5 Proof of Lemma 1

Proof. Since the data points are drawn randomly, the quantity δ_n is a random variable, and the goal is to find a probabilistic bound on δ_n. To this end, we start by investigating the probability of $\delta_n > r$ for a given $r > 0$. Note that if $\delta_n > r$, then the definition of δ_n in (4) implies that there exists $\alpha_0 \in \mathbf{P}_\theta$ such that no data point lies in $\mathbf{B}(\alpha_0, r)$ (defined in item 12 of Appendix A.1.2). As a result, we can write the following inequality for the probability of $\delta_n > r$:

$$\mathbb{P}\left[\delta_n > r\right] \leq \mathbb{P}\left[\text{no data point lies in } \mathbf{B}(\alpha_0, r)\right] = \left(1 - \frac{\operatorname{vol}\left(\mathbf{B}(\alpha_0, r) \cap \mathbf{P}_\theta\right)}{\operatorname{vol}(\mathbf{P}_\theta)}\right)^n. \tag{A.43}$$

The leftmost inequality in (A.43) holds due to item 16 of Appendix A.1.3. Note that $r < \delta_n \leq \operatorname{diam}(\mathbf{P}_\theta)$, where $\operatorname{diam}(\cdot)$ is defined in item 13 of Appendix A.1.2. Therefore, due to Lemma A.5, there exists $\xi(\mathbf{P}_\theta) > 0$ such that:

$$\operatorname{vol}\left(\mathbf{B}(\alpha_0, r) \cap \mathbf{P}_\theta\right) \geq \xi(\mathbf{P}_\theta) \cdot \operatorname{vol}\left(\mathbf{B}(\alpha_0, r)\right). \tag{A.44}$$

Next, let $p = \mathbb{P}[\delta_n > r]$. By substituting this together with (A.44) into (A.43), we conclude that:

$$p \leq \left(1 - \frac{\xi(\mathbf{P}_\theta) \cdot \operatorname{vol}(\mathbf{B}(\alpha_0, r))}{\operatorname{vol}(\mathbf{P}_\theta)}\right)^n \quad \Rightarrow \quad \log p \leq n \log\left(1 - c_0\, r^K\right), \tag{A.45}$$

where $c_0 > 0$ is a constant that does not depend on r, n, or p. Next, using the Taylor series expansion of $\log(1/(1-x))$ for $x \in (0, 1)$, the following inequality results from (A.45):

$$\log \frac{1}{p} \geq n \log \frac{1}{1 - c_0\, r^K} = n \left(c_0\, r^K + \sum_{i=2}^{\infty} \frac{(c_0\, r^K)^i}{i}\right) \geq n c_0 r^K. \tag{A.46}$$

Therefore,

$$r \leq c_0^{-\frac{1}{K}} \cdot n^{-\frac{1}{K}} \cdot \log^{\frac{1}{K}}\left(\frac{1}{p}\right). \tag{A.47}$$

Recalling $p = \mathbb{P}\left(\delta_n > r\right)$, we know that with probability $1 - p$, we have $\delta_n \leq r$. This, together with (A.47), leads to the fact that with probability $1 - p$:

$$\delta_n = (-\log p)^{\frac{1}{K}} \cdot \mathcal{O}\left(n^{-\frac{1}{K}}\right). \tag{A.48}$$

□

A.6 Proof of Proposition 2

Proof. We start by fixing some notations. For $1 \leq r \leq M$, let f^r and f_n^r denote the r-th component function of f and f_n, respectively. Moreover, denote the i-th data point by θ_i. Next, we rewrite the kernel smoothing function $f_n(\theta)$ in Equation (6) based on nonzero terms. Recall that $\eta(x) = 0$ for $x \geq 1$. This means $k_n(\theta, \theta_i) = 0$ when $\|\theta - \theta_i\|_2 \geq \lambda_n$. Therefore, f_n consists of a weighted average over the data points lying in $\mathbf{B}(\theta, \lambda_n)$. Let $\mathbf{I}_n(\theta)$ denote these data points, i.e.,

$$\mathbf{I}_n(\theta) = \{1 \leq i \leq n \mid \theta_i \in \mathbf{B}(\theta, \lambda_n)\}. \tag{A.49}$$

For $\theta \in \mathbf{P}_\theta$, observe that $\mathbf{I}_n(\theta) \neq \emptyset$ because otherwise, we would have $\min_{1\leq i\leq n} \|\theta - \theta_i\|_2 \geq \lambda_n > \delta_n$, which contradicts the definition of δ_n in (4). Having $\mathbf{I}_n(\theta) \neq \emptyset$ guarantees that $f_n(\theta)$ remains well-defined. More formally, for all $\theta \in \mathbf{P}_\theta$, we have:

$$\sum_{i=1}^{n} k_n(\theta, \theta_i) = \sum_{i\in\mathbf{I}_n(\theta)} k_n(\theta, \theta_i) > 0. \tag{A.50}$$

Hence, we can rewrite $f_n(\theta)$ as follows:

$$f_n(\theta) = \frac{\sum_{i\in\mathbf{I}_n(\theta)} k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i\in\mathbf{I}_n(\theta)} k_n(\theta, \theta_i)}. \tag{A.51}$$

Suppose $n \in \mathbb{N}$, $1 \leq r \leq M$, and $\theta \in \mathbf{P}_\theta$ are given. We claim that there exist $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n)$ such that:

$$|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)|. \tag{A.52}$$

We prove (A.52) as follows: Based on (A.51), $f_n^r(\theta)$ is a weighted average of the values $f^r(\theta_i)$, and thus it is bounded between the maximum and the minimum of these values:

$$\min_{i\in\mathbf{I}_n(\theta)} f^r(\theta_i) \leq f_n^r(\theta) \leq \max_{i\in\mathbf{I}_n(\theta)} f^r(\theta_i). \tag{A.53}$$

Next, let:

$$\omega_1 = \operatorname*{arg\,min}_{\alpha\in\bar{\mathbf{B}}(\theta,\lambda_n)\cap\mathbf{P}_\theta} f^r(\alpha), \qquad \omega_2 = \operatorname*{arg\,max}_{\alpha\in\bar{\mathbf{B}}(\theta,\lambda_n)\cap\mathbf{P}_\theta} f^r(\alpha). \tag{A.54}$$

Note that ω_1 and ω_2 exist because f^r is continuous and $\bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$ is compact (see items 3 and 4 of Appendix A.1.3). Moreover, if $i \in \mathbf{I}_n(\theta)$, then $\theta_i \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$. Therefore,

$$f^r(\omega_1) = \min_{\alpha\in\bar{\mathbf{B}}(\theta,\lambda_n)\cap\mathbf{P}_\theta} f^r(\alpha) \leq \min_{i\in\mathbf{I}_n(\theta)} f^r(\theta_i), \tag{A.55}$$

$$\max_{i\in\mathbf{I}_n(\theta)} f^r(\theta_i) \leq \max_{\alpha\in\bar{\mathbf{B}}(\theta,\lambda_n)\cap\mathbf{P}_\theta} f^r(\alpha) = f^r(\omega_2). \tag{A.56}$$

Putting (A.53), (A.55), and (A.56) together leads to:

$$f^r(\omega_1) \leq f_n^r(\theta) \leq f^r(\omega_2). \tag{A.57}$$

Moreover, since $\theta \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$, we can write:

$$f^r(\omega_1) \leq f^r(\theta) \leq f^r(\omega_2). \tag{A.58}$$

Due to (A.57) and (A.58), both $f_n^r(\theta)$ and $f^r(\theta)$ lie between $f^r(\omega_1)$ and $f^r(\omega_2)$. Hence, (A.52) holds.

Having established the above results, we present the proofs as follows:

i. Note that as shown above, f_n is well-defined. Moreover, f_n is continuous due to the continuity of $\eta(\cdot)$. Therefore, using Theorem 1, it suffices to prove the uniform convergence $f_n \xrightarrow{\text{unif}} f$. To prove $f_n \xrightarrow{\text{unif}} f$, suppose $\epsilon > 0$ is given. It suffices to show that there exists N such that for $n \geq N$, $\|f_n(\theta) - f(\theta)\|_2 < \epsilon$ holds for all $\theta \in \mathbf{P}_\theta$.

To this end, note that f^r is a continuous function over a compact set $\mathbf{P}_\theta$, and thus is uniformly continuous over $\mathbf{P}_\theta$ (see item 5 of Appendix A.1.3). Hence, there exists $\beta_r > 0$ such that if $\|\alpha_1 - \alpha_2\|_2 < \beta_r$, then $|f^r(\alpha_1) - f^r(\alpha_2)| < \epsilon/\sqrt{M}$. Let $\beta = \min(\beta_1, \ldots, \beta_M)$. Since $\lambda_n \to 0$, there exists N such that for $n \geq N$, we have $\lambda_n < \beta/2$.

Suppose the values $n \geq N$, $1 \leq r \leq M$, and $\theta \in \mathbf{P}_\theta$ are given. Due to (A.52), there exist $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n) \subset \mathbf{B}(\theta, \beta/2)$ such that $|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)|$. Note that $\|\omega_2 - \omega_1\|_2 < \beta$ and thus $|f^r(\omega_2) - f^r(\omega_1)| < \epsilon/\sqrt{M}$ holds due to the definition of β mentioned earlier. Hence, $|f_n^r(\theta) - f^r(\theta)| < \epsilon/\sqrt{M}$. Applying this result for all $1 \leq r \leq M$ together, we conclude that if $n \geq N$, then $\|f_n(\theta) - f(\theta)\|_2 < \epsilon$ holds for all $\theta \in \mathbf{P}_\theta$. This means $f_n \xrightarrow{\text{unif}} f$.

ii. We proceed through the following steps:

Step 1. Showing that ∇f_n exists and is continuous over any open set $\mathbf{O} \subseteq \mathbf{P}_\theta$:

Suppose the open set $\mathbf{O} \subseteq \mathbf{P}_\theta$ is given. Considering f_n from (6), it is straightforward to compute its corresponding $M \times K$ Jacobian matrix evaluated at $\theta \in \mathbf{O}$ (see item 21 of Appendix A.1.2), which is given by:

$$\nabla f_n(\theta) = \frac{1}{\sum_{i=1}^n k_n(\theta, \theta_i)} \left[\sum_{i=1}^n \sum_{j=1}^n \alpha_{ij} f(\theta_i) \nabla k_n(\theta, \theta_j)^\top \right], \tag{A.59}$$

where $f(\theta_i) \in \mathbb{R}^M$ is a column vector (i.e., it is $M \times 1$) and:

$$\alpha_{ij} = \begin{cases} -w_i & \text{if } i \neq j \\ 1 - w_i & \text{if } i = j \end{cases}, \quad w_i = \frac{k_n(\theta, \theta_i)}{\sum_{i=1}^n k_n(\theta, \theta_i)}, \quad \nabla k_n(\theta, \theta_i) = \frac{2(\theta - \theta_i)}{\lambda_n^2} \eta' \left(\frac{\|\theta - \theta_i\|_2^2}{\lambda_n^2} \right). \tag{A.60}$$

Note that in (A.60), $\eta'(\cdot)$ denotes the derivative of $\eta(\cdot)$ and $\nabla k_n(\theta, \theta_i)$ is a row vector (i.e., it is $K \times 1$). Moreover, (A.59) and (A.60) together imply that $\nabla f_n(\theta)$ exists and is continuous over $\mathbf{O}$, due to the existence and continuity of $\eta'(\cdot)$, provided that the denominator $\sum_{i=1}^n k_n(\theta, \theta_i)$ is nonzero for all $\theta \in \mathbf{O}$. This is shown in (A.50).

Step 2. Proving f is Lipschitz continuous around θ^*:

We know that the assumptions of Proposition 1 hold. As a result, there exists $\mathbf{O} \subseteq \mathbf{P}_\theta$ around θ^* such that ∇f is continuous over $\mathbf{O}$. In particular, ∇f is continuous at $\theta^* \in \mathbf{O}$. Therefore, there exists $\beta_0 > 0$ such that for $\theta \in \mathbf{B}(\theta^*, \beta_0) \subseteq \mathbf{O}$, we have $\|\nabla f(\theta) - \nabla f(\theta^*)\|_2 < \frac{1}{2}\|\nabla f(\theta^*)\|_2$, and consequently, $\|\nabla f(\theta)\|_2 < \frac{3}{2}\|\nabla f(\theta^*)\|_2$.

Let $L_1 = \frac{3}{2}\|\nabla f(\theta^*)\|_2$ and define $\mathbf{S} = \mathbf{B}(\theta^*, \beta_0)$. Observe that $\mathbf{S}$ is open and convex in $\mathbb{R}^K$, and $\|\nabla f(\theta)\|_2 < L_1$ holds for $\theta \in \mathbf{S}$. This leads to f being Lipschitz continuous with the Lipschitz factor L_1, due to the Mean Value Inequality (see item 15 of Appendix A.1.3). More formally, $\|f(\alpha_1) - f(\alpha_2)\|_2 \leq L_1 \|\alpha_1 - \alpha_2\|_2$ holds for all $\alpha_1, \alpha_2 \in \mathbf{S}$.

Step 3. Finding the convergence rate of $\|\hat{\theta} - \theta^*\|_2$:

Recall $\mathbf{S} = \mathbf{B}(\theta^*, \beta_0)$ from Step 2. We know that $\lambda_n = \gamma\, \delta_n \to 0$. Therefore, there exists N_1 such that for $n \geq N_1$, we have $\lambda_n < \beta_0/2$. Now, suppose $n \geq N_1$, $\theta \in \mathbf{B}(\theta^*, \beta_0/2)$, and $1 \leq r \leq M$

are given. Then $\bar{\mathbf{B}}(\theta, \lambda_n) \subseteq \mathbf{S}$. Additionally, from (A.52), there exist $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n)$ such that $|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)|$. This, together with the Lipschitz continuity of f on $\mathbf{S}$ from Step 2 and the fact that $\omega_1, \omega_2 \in \mathbf{S}$, leads to the following inequality:

$$|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)| \leq \|f(\omega_1) - f(\omega_2)\|_2 \leq L_1 \|\omega_1 - \omega_2\|_2 \leq 2L_1\lambda_n. \tag{A.61}$$

Note that the rightmost inequality in (A.61) holds because $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n)$. By applying (A.61) to every component for $1 \leq r \leq M$, we conclude that $\|f_n(\theta) - f(\theta)\|_2 \leq 2\sqrt{M}L_1\lambda_n$ holds for $n \geq N_1$ and $\theta \in \mathbf{B}(\theta^*, \beta_0/2) \subseteq \mathbf{S} \subseteq \mathbf{O}$.

Next, having $\hat{\theta}_n \to \theta^*$ from part (i) of the theorem, there exists N_2 such that if $n \geq N_2$, then $\hat{\theta}_n \in \mathbf{B}(\theta^*, \beta_0/2)$. This implies that $\|f_n(\theta) - f(\theta)\|_2 \leq 2\sqrt{M}L_1\lambda_n$ holds when $\theta = \hat{\theta}_n$. Therefore, knowing that the conditions of Proposition 1 are all met (due to Step 1 and the assumptions) and using its notation, we can write the following inequality for $n \geq \max(N_0, N_1, N_2)$:

$$\left\|\Lambda_n \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right)\right\|_2 \leq \|\Lambda_n\|_2 \|f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\|_2 \leq L_0 \cdot 2\sqrt{M}L_1\lambda_n = 2L_0L_1\gamma\sqrt{M}\delta_n. \tag{A.62}$$

Applying Proposition 1 together with (A.62), we then conclude that $\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}(\delta_n)$ as $n \to \infty$.

□

A.7 Proof of Proposition 3 (Informal)

Proof. i. We simply verify the conditions of Theorem 1, where f_n comes from a neural network estimate with the structure of Equation (9). Note that $\phi(\cdot)$ in (9) is a continuous function, and so is f_n. Moreover, due to the result from Cybenko (1989), when the target function f is a continuous function defined over a compact set, the output of an appropriately trained single-layer neural network uniformly converges to the target function. Therefore, the conditions of Theorem 1 are satisfied, and thus $\hat{\theta}_n \to \theta^*$.

ii. Knowing that the assumptions in Proposition 1 hold, we use its corresponding notation. Based on the assumptions of Proposition 1, Step 2 in the proof of Proposition 2 shows that f is Lipschitz continuous on an open ball $\mathbf{B}(\theta^*, \beta_0) \subseteq \mathbf{P}_\theta$, with Lipschitz factor L_1. Moreover, consider $\hat{\mathbf{S}} = \bar{\mathbf{B}}(\theta^*, \beta_0/2) \subset \mathbf{B}(\theta^*, \beta_0)$ and note that for sufficiently large n, $\hat{\theta}_n$ lies in $\hat{\mathbf{S}}$. Now, by applying Proposition 1, we conclude that:

$$\|\hat{\theta}_n - \theta^*\|_2 \leq \left\|\Lambda_n \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right)\right\|_2 + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right) \tag{A.63}$$

$$\leq L_0 \sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2 + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right). \tag{A.64}$$

Next, we investigate the bound on $\sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2$. Let f^r and f_n^r denote the r-th component functions of f and f_n, respectively. According to Barron (1994), the mean integrated squared error of the neural network estimator f_n^r for f^r is $\mathcal{O}(\sqrt{(K/n)\log n})$, provided that the number of nodes satisfies $N = \sqrt{n/(K \log n)}$. More formally, we have:

$$\frac{1}{\text{vol}(\mathbf{P}_\theta)} \int_{\mathbf{P}_\theta} |f^r(\theta) - f_n^r(\theta)|^2 \, d\theta = \mathcal{O}\left(\left(\frac{K \log n}{n}\right)^{\frac{1}{2}}\right). \tag{A.65}$$

The goal is to translate the bound in (A.65) to a bound on $\sup_{\theta\in\hat{\mathbf{S}}}\|f(\theta)-f_n(\theta)\|_2$. To this end, we proceed as follows: The uniform boundedness of $\|\nabla f_n(\theta)\|_2$ across n and θ implies that there exists $L_2>0$ such that every function f_n is Lipschitz continuous with Lipschitz constant L_2 (see item 15 of Appendix A.1.3). Therefore, letting $h_n(\theta)=|f^r(\theta)-f_n^r(\theta)|$, we conclude that $h_n(\theta)$ is Lipschitz continuous on $\hat{\mathbf{S}}$ with Lipschitz constant L_1+L_2. Furthermore, due to the continuity of h_n and the compactness of $\hat{\mathbf{S}}$, there exist $\theta_{\min}=\arg\min_{\theta\in\hat{\mathbf{S}}}h_n(\theta)$ and $\theta_{\max}=\arg\max_{\theta\in\hat{\mathbf{S}}}h_n(\theta)$ (see item 4 of Appendix A.1.3). Consequently, by the Lipschitz continuity of h_n on $\hat{\mathbf{S}}$, we have:

$$h_n(\theta_{\max})\le h_n(\theta_{\min})+(L_1+L_2)\,\|\theta_{\max}-\theta_{\min}\|_2\le h_n(\theta_{\min})+(L_1+L_2)\,\beta_0. \tag{A.66}$$

Recall that β_0 is the diameter of $\hat{\mathbf{S}}$. Moreover, the following inequality holds:

$$h_n^2(\theta_{\min})\le\frac{1}{\mathrm{vol}(\hat{\mathbf{S}})}\int_{\hat{\mathbf{S}}}|f^r(\theta)-f_n^r(\theta)|^2\,d\theta\le\frac{1}{\mathrm{vol}(\hat{\mathbf{S}})}\int_{\mathbf{P}_\theta}|f^r(\theta)-f_n^r(\theta)|^2\,d\theta. \tag{A.67}$$

Now, (A.65) and (A.67) together imply that $h_n(\theta_{\min})=\mathcal{O}(((K/n)\log n)^{1/4})$. Applying this to (A.66) yields the same bound on $h_n(\theta_{\max})$. Hence, $\sup_{\theta\in\hat{\mathbf{S}}}|f^r(\theta)-f_n^r(\theta)|=\mathcal{O}(((K/n)\log n)^{1/4})$. Therefore, if we use M neural networks with $N=\sqrt{n/(K\log n)}$ nodes for each to estimate $f^1,\dots,f^M$, we obtain the following bound for the overall error:

$$\sup_{\theta\in\hat{\mathbf{S}}}\|f(\theta)-f_n(\theta)\|_2\le\sqrt{M}\max_{1\le r\le M}\sup_{\theta\in\hat{\mathbf{S}}}|f^r(\theta)-f_n^r(\theta)|=\mathcal{O}\left(\sqrt{M}\left(\frac{K\log n}{n}\right)^{1/4}\right). \tag{A.68}$$

Applying (A.68) to (A.64) then completes the argument.

□

A.8 Proof of Theorem 2

Proof. i. The convexity of $\mathbf{X}\times\mathbf{P}$ implies the convexity of $\mathbf{X}$, which ensures that $\mathbf{X}$ is Lebesgue measurable (Lang, 1986). Consequently, the restriction of the Lebesgue measure on $\mathbb{R}^d$ to $\mathbf{X}$, denoted by ν_d, is well-defined, and the theorem assumes $\mu_x\ll\nu_d$ (see item 19 of Appendix A.1.2). With this assumption, the measurability of a set or function with respect to μ_x and ν_d coincides.

We prove the result for a more general case by replacing light discontinuity in the assumptions with negligible discontinuity, defined in Definition A.1. First, we show that if $h(x,\theta)$ is lightly discontinuous over θ, then $h(x,\theta)$ is negligibly discontinuous over θ. To this end, suppose $h(x,\theta)$ is lightly discontinuous over θ. Letting $\mathbf{D}$ denote the set of discontinuity points of h, and $\mathbf{D}_{\theta_0}\subseteq\mathbf{X}$ the cross-section of $\mathbf{D}$ with $\theta=\theta_0$, we have:

$$\mathbf{D}_{\theta_0}=\bigcup_{r=1}^{\infty}\mathbf{D}_{\theta_0}^r,\quad\text{where}\quad\mathbf{D}_{\theta_0}^r=\{x\mid x_{\mathbf{I}}=F_r(x_{-\mathbf{I}},\theta_0),\ \text{for some}\ \emptyset\neq\mathbf{I}\subseteq[d]\}. \tag{A.69}$$

Considering the notation $z=x_{\mathbf{I}}$, $y=x_{-\mathbf{I}}$, and $h(\cdot)=F_r(\cdot,\theta_0)$ leads to $z=h(y)$ and the fact that the elements of x are a reordered version of the elements of (y,z). Therefore, $\mathbf{D}_{\theta_0}^r$ is in one-to-one correspondence with $\{(y,z)\mid z=h(y)\}$. This clarifies that $\mathbf{D}_{\theta_0}$ describes the "graph" of a function lying in $\mathbb{R}^d$, and thus it has zero measure with respect to ν_d (see item 19 of Appendix A.1.3),

i.e., $\nu_d(\mathbf{D}_{\theta_0}^r) = 0$, or equivalently, $\mathbf{D}_{\theta_0}^r$ is ν_d-negligible. Therefore, $\mathbf{D}_{\theta_0}$ is a countable union of ν_d-negligible sets and is thus ν_d-negligible (see item 18 of Appendix A.1.3). Hence, $h(x, \theta)$ is negligibly discontinuous over θ. Now observe that the conditions of Lemma A.6 are all met. By applying Lemma A.6, we conclude that $f(\theta)$ is continuous.

ii. This is a direct result of part (i) by observing that $f(\theta) = \mathbb{P}_x[A(x, \theta)] = \mathbb{E}_x[\mathbf{1}_{A(x,\theta)}] = \mathbb{E}_x[h(x, \theta)]$.

□

A.9 Proof of Theorem 3

Proof. Let $\mathbf{X} = \mathbf{P}_a \times \mathbf{P}_z \times \mathbf{P}_\theta$ with $\mathbf{X} \subseteq \mathbb{R}^d$, and define the set-valued function $\mathbf{P}(s, z; \theta) \subseteq \mathbf{P}_a$, whose domain is $\mathbb{R}^d$, such that $\mathbf{P}(s, z; \theta) = \emptyset$ if $(s, z; \theta) \notin \mathbf{X}$, and for $(s, z; \theta) \in \mathbf{X}$:

$$\mathbf{P}(s, z; \theta) = \{a \in \mathbf{P}_a \mid \ M_i(a, s, z; \theta) \leq 0, \ \ 1 \leq i \leq N\}, \tag{A.70}$$

where, in the absence of inequality constraints, we simply have $\mathbf{P}(s, z; \theta) = \mathbf{P}_a$. Our approach is to apply Lemma A.4 through the following steps:

Step 1. Proving the requirements for $\mathbf{P}(s, z; \theta)$:
In this step, our goal is to apply Lemma A.3. To this end, first observe that $\text{int}(\mathbf{P}_a) \neq \emptyset$ because $\mathbf{P}_a$ is assumed to contain an open set (see item 18 of Appendix A.1.2). Moreover, substituting $x = (s, z; \theta)$, $\mathbf{A} = \mathbf{P}_a$, and $g_i(\cdot) = M_i(\cdot)$, note that the conditions of Lemma A.3 are all satisfied. This implies that $\mathbf{P}(\cdot)$ is continuous relative to $\mathbf{X}$ (in the Kuratowski sense), and $\mathbf{P}(x)$ is a nonempty and closed subset of $\mathbf{P}_a$ for every $x \in \mathbf{X}$.

Step 2. Proving the existence of a unique and continuous $V(\cdot)$:
Let $\mathcal{F}$ be the space of continuous real-valued functions over the compact set $\mathbf{X}$, equipped with the supremum norm $\|\cdot\|_\infty$ (defined in item 15 of Appendix A.1.2). As a result, $(\mathcal{F}, \|\cdot\|_\infty)$ is a Banach space (see item 12 of Appendix A.1.3). This is a requirement for the Contraction Mapping Theorem, which will be discussed below.

Next, define the operator $T(\cdot)$, which accepts as input a continuous function $V : \mathbf{X} \to \mathbb{R}$, and returns as output the function $T(V) : \mathbf{X} \to \mathbb{R}$, where the relation between the input and output is described by:

$$T(V(s, z; \theta)) = \sup_{a \in \mathbf{P}(s,z;\theta)} \pi(a, s, z; \theta) + \beta\, \mathbb{E}_{z'}[\, V(a, z'; \theta) \mid z]. \tag{A.71}$$

Note that the domain of $T(\cdot)$ is $\mathcal{F}$. We claim that $T(\mathcal{F}) \subseteq \mathcal{F}$, or equivalently, $T(V) \in \mathcal{F}$ if $V \in \mathcal{F}$. To prove this, suppose $V \in \mathcal{F}$, meaning that V is continuous. Therefore, the mapping $(a, s, z; \theta) \mapsto \pi(a, s, z; \theta) + \beta\, \mathbb{E}_{z'}[\, V(a, z'; \theta) \mid z]$ is well-defined and continuous. To see why it is so, observe that the continuity of V over a compact set implies its integrability with respect to the Lebesgue measure (see item 21 of Appendix A.1.3). This, together with the fact that $\mu_{z'|z}$ is dominated by the Lebesgue measure, ensures the preservation of continuity under the expectation $\mathbb{E}_{z'}[\cdot \mid z]$ (see item 20 of Appendix A.1.3).

The continuity of the aforementioned mapping, together with Step 1 and substituting $x = (s, z; \theta)$, $\mathbf{A} = \mathbf{P}_a$, $\mathbf{Q}(\cdot) = \mathbf{P}(\cdot)$, and $f(a, x) = \pi(a, s, z; \theta) + \beta\, \mathbb{E}_{z'}[V(a, z'; \theta) \mid z]$, shows that the conditions of Lemma A.4 are satisfied. As a result, $T(V)$ is a continuous function, and thus $T(V) \in \mathcal{F}$. Hence, the claim holds, i.e., $T(\mathcal{F}) \subseteq \mathcal{F}$.

Moreover, it is straightforward to verify that $T(\cdot)$ satisfies Blackwell's sufficient conditions for a contraction mapping (Stokey et al., 1989, Theorem 3.3). Therefore, $T : \mathcal{F} \to \mathcal{F}$ is a contraction mapping operating over the Banach space $\mathcal{F}$. As a result, the conditions of the Contraction Mapping Theorem (Stokey et al., 1989, Theorem 3.2) are all satisfied, ensuring that T has a unique fixed point. This means there exists a unique $V \in \mathcal{F}$ satisfying $V = T(V)$. Recall that $\mathcal{F}$ is the space of continuous functions over $\mathbf{X}$, and thus V is continuous.

Step 3. Proving the existence of $a^*(\cdot)$ and its continuity:
Our approach is to apply Lemma A.4. To this end, note that within Step 2, it is shown that if $V \in \mathcal{F}$, the conditions of Lemma A.4 are all satisfied. Now, as a final result of Step 2, we know that $V \in \mathcal{F}$. Hence, Lemma A.4 applies, and we conclude that the optimal solution of the maximization problem (13) is attained at some $a^*(s, z; \theta) \in \mathbf{P}(s, z; \theta)$. Furthermore, given the uniqueness of $a^*(s, z; \theta)$ for every $(s, z; \theta) \in \mathbf{X}$, from Lemma A.4 we conclude that $a^*(\cdot)$ is continuous. □

B Corporate finance model: Implementation details

B.1 Construction of moments

We start with a COMPUSTAT extract over 1970-2019. We only keep firms that appear at least twice in the sample. We drop firms in the financial (SIC code 6) or regulated (SIC code 49) sectors. We also drop observations with total assets that are less than 10 million real 1982 dollars, or sales or book assets that grow by more than 200%. This results in a sample of 117,976 firm-year observations and 11,198 unique firms. We compute moments targeted in the baseline estimation in the data as follows: m_1, the average investment to capital ratio, is $\frac{\texttt{capx}}{l.\texttt{at}}$. m_2, the average profit to asset ratio, is $\frac{\texttt{oibdp}}{l.\texttt{at}}$. m_3, the average equity issuance to asset ratio, is computed net of repurchases: $\frac{\texttt{sstk}-\texttt{prstkc}}{l.\texttt{at}}$. m_4, mean net leverage, is $\frac{\texttt{dlc+dltt }-\texttt{che}}{\texttt{at}}$. m_5, the autocorrelation of investment rates, is measured by regressing $\frac{\texttt{capx}}{l.\texttt{at}}$ on its lag, with year fixed-effects. Last, m_6 and m_7, are the sample standard deviations of 1-year and 5 year log sales growth: $\log \texttt{sale} - \log \texttt{l.sale}$ and $\log \texttt{sale} - \log \texttt{l5.sale}$. All ratios are winsorized at the median +/- five times the interquartile range. We also remove firm fixed-effects from all the variables used in the empirical analysis, as the model does not feature any source of fully persistent heterogeneity across firms: for each variable, we subtract the within-firm average and add back the overall sample average. The bold lines in Table B.1, in Column 1, provide the means and standard errors of these moments in our sample.

B.2 Kernel approximation: Implementation

We use a kernel function that maps parameters into moments using the formula (6):

$$f_n(\theta) = \frac{\sum_{i=1}^{n} k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i=1}^{n} k_n(\theta, \theta_i)}, \tag{B.1}$$

where the kernel itself is a Gaussian kernel such that:

$$k_n(\tilde{\theta}, \tilde{\theta}_i) = \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\left(\frac{\theta_k - \theta_{i,k}}{b_k}\right)^2\right)$$

where b_k is the bandwidth corresponding to parameter number k. Following proposition 2, we set the bandwitdh to be:

$$b_k = \gamma\left(\overline{\theta_k} - \underline{\theta_k}\right)\left(\frac{1}{n}\right)^{\frac{1}{K}}$$

where $\left(\overline{\theta_k} - \underline{\theta_k}\right)$ is the range of parameter number k. n is the number of points in the training sample and K the number of parameters. The parameter γ, common to all parameters, is used to fit the kernel approximation. We use $\gamma = 0.5$ in our application provides the best in-sample fit.

B.3 Neural net approximation: Implementation

The neural net is a pyramidal MLP with 5 layers and 512, 256, 128, 64 and 32 nodes. The activation function is $ReLu$ at all layers. The loss function is the mean absolute error (which turns out to give a better in-sample fit than the MSE). The optimization algorithm is ADAM, a modern popular version of

Table B.1: Simulation Moments (Corporate Finance Model)

		(1) Data	(2) True SMM	(3) Approximate	(4) Simulation
m_1	**mean(investment/assets)**	.0760 (.0007)	.0760	.0760	.0759
m_2	**mean(profit/assets)**	.1343 (.0011)	.1343	.1343	.1343
m_3	**mean(equity issuance/assets)**	.0158 (.0006)	.0159	.0158	.0164
m_4	**mean(leverage)**	.1049 (.0029)	.1050	.1049	.1035
m_5	**autocorr(investment/assets)**	.3754 (.0066)	.3758	.3754	.3869
m_6	**std(log sales growth)**	.2270 (.0017)	.2270	.2270	.2259
m_7	**std(log sales growth 5yr)**	.5851 (.0050)	.5851	.5851	.5890
m_8	var(investment/assets)	.0033 (.0001)	.0167	.0176	.0167
m_9	var(equity issuance/assets)	.0071 (.0002)	.0024	.0020	.0025
m_{10}	frequency(equity issuance)	.1178 (.0015)	.1534	.1565	.1576
m_{11}	coeff. regr. investment ratio on market/book	.0122 (.0005)	.3188	.3051	.3074
m_{12}	coeff. regr. net leverage on market/book	-0.0348 (.0018)	-0.0313	-0.0278	-0.0285
m_{13}	coeff. AR(1) regr. of profit/assets	.5210 (.0055)	.5357	.5445	.5433
m_{14}	resid std AR(1) regr. of profit/assets	.0728 (.0006)	.0286	.0285	.0285
m_{15}	var(leverage)	.0266 (.0004)	.0002	.0002	.0002
m_{16}	mean(dividend/assets)	.0267 (.0004)	.0492	.0505	.0498
m_{17}	var(dividend/assets)	.0013 (.0000)	.0054	.0052	.0053

Notes. The 'Data' column reports moments in the data with standard errors in parenthesis. Column 'True SMM' reports simulated moments using the true economic model $f(\theta)$ and parameters estimated using the true SMM. Column 'Approximate' reports moments calculated using the benchmark approximation f_n (neural net) and the parameters estimated using the approximate SMM. Column 'Simulation' reports moments calculated from the true economic model $f(\theta)$ but using parameter estimates from the approximate SMM. Targeted moments in the SMM are shown in bold font.

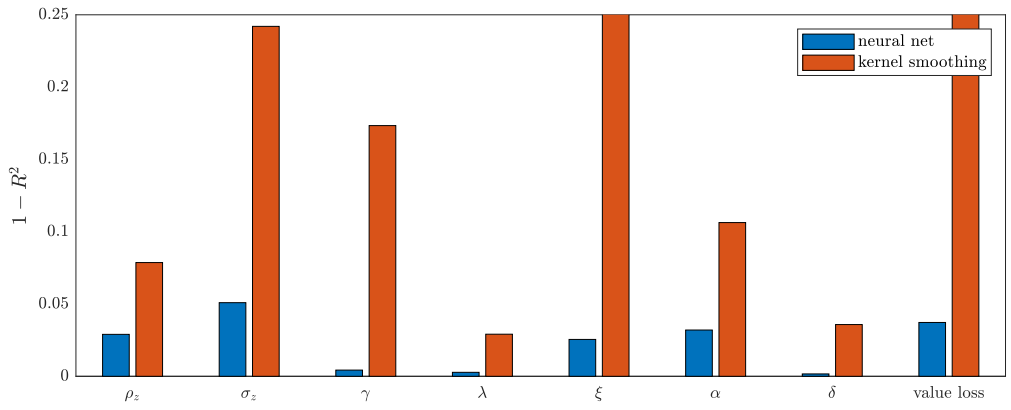
stochastic gradient descent with adaptive learning rate. The maximum number of epochs is 40, batch size is 256 points, initial learning rate is 0.005 (divided by 10 every 10 epochs).

B.4 The superiority of the NN fit

We now compare the fit of neural nets and kernel approximations. Figure B.1 reports the precision of our estimation method on the validation sample, for each parameter separately, and for each parametrization of f_n. We report one minus the R^2 of true parameters regressed on estimated ones. A value of 0 means that estimated parameters are perfectly (and linearly) correlated with true ones. First, Figure B.1 shows that our NN is much more accurate than the kernel. This is not surprising given that NNs converge faster than Kernel especially in high dimensions (see our propositions 2 and 3; see also Farrell et al. (2021) for DNNs). Second, we see that the NN performs very well for all parameters (with an R^2 greater than 95% for all deep parameters). Third, there is, however, some heterogeneity, with some parameters (for instance the debt constraint λ) being better estimated than others (for instance, the cost of equity issuance ξ).

Overall, the kernel-based approach performs much less well, as expected from our results on speeds

Figure B.1: Precision of Approximate Estimates on the Validation Sample



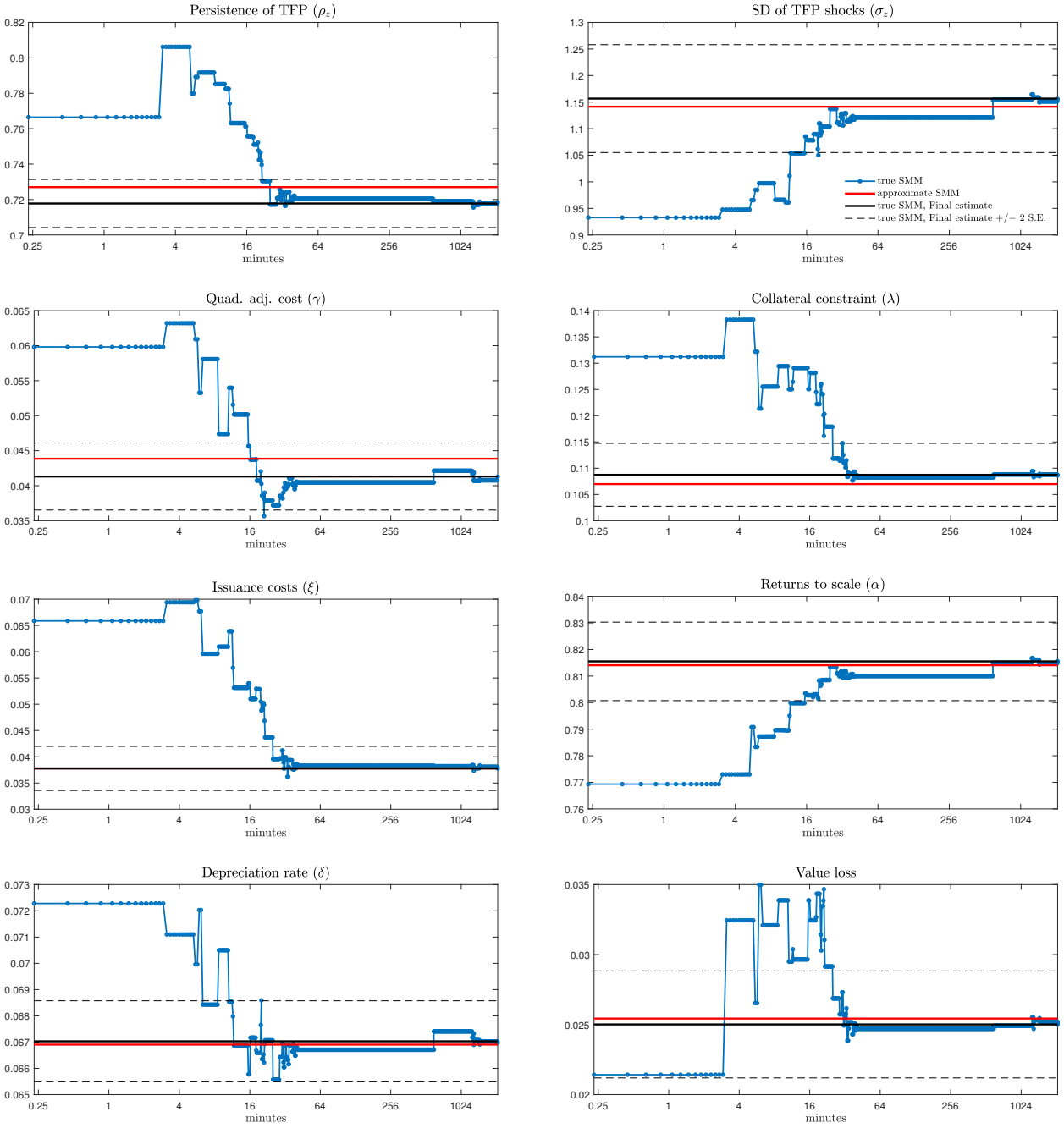
Notes. This figure reports a measure of the estimation error, on the validation sample, from the approximate SMM for two approximations: Kernel and neural net. For each one of the two parametrizations, and each one of the 7 parameters plus the value loss, we report $1 - R^2$, one minus the R^2 of a linear regression of the approximate moment estimate on the true parameter value. If the approximate moment estimate is exactly equal to the true value, this quantity is zero. When it is uncorrelated with the true value, it is 1.

of convergence, and existing results on deep NNs ((Farrell et al., 2021)). This is why the paper focuses on the deep NN specification for f_n.

B.5 Speed gains from the NN approximation

Figure B.2 shows that our approach is faster than the standard SMM by several orders of magnitude. The computing times we report exclude the simulation of the training sample, which is long but required for both estimations. For the true SMM, it takes an additional 17 minutes for the estimation to converge to its final value. In contrast, the approximation-based estimation converges in less than a second (provided the NN has been estimated).

Figure B.2: True SMM Estimates, Convergence Speed of Local Optimization Stage (Corporate Finance Model)



Notes. We report parameter values estimated as a function of time taken via two different algorithms in our numerical setup. The blue line corresponds to the local optimization stage of the true SMM, i.e. the minimization of the distance of empirical moments to the true model $f(\theta)$. The optimization algorithm used in this case is Tiktak, using 50 starting points selected from a training set of 50,000 cases and Nelder-Mead algorithm for optimization per starting point with 200 max function iteration. The red line corresponds to the benchmark approximate SMM, which uses a neural net. The approximate estimation requires .9 seconds — hence the red line jumps to its final value at the origin. The black line corresponds to the true SMM estimate and the dashed lines represent the confidence interval (+/-2 standard errors) around it.

C Dynamic Household Finance Model

C.1 Model, parameters and moments

Macroeconomic environment The stock market log return in year t is $s_t = s_{1,t} + s_{2,t}$ where $s_{2,t}$ is normally distributed with variance $\sigma^2_{s_2}$. s_1 will be correlated with labor income, and follows a normal mixture distribution:

$$s_{1,t} = \begin{cases} s^-_{1,t} \sim \mathcal{N}(\mu^-_s, \sigma^2_{s_1}) & \text{with probability} \quad p_s \\ s^+_{1,t} \sim \mathcal{N}(\mu^+_s, \sigma^2_{s_1}) & \text{with probability} \quad 1-p_s \end{cases} \tag{C.1}$$

where we impose $\mu^-_s < \mu^+_s$ and interpret μ^-_s as the expected log return in stock market crash years, and p_s their frequency. The growth of the log national wage index l_1 is:

$$l_{1,t} - l_{1,t-1} = \mu_l + \lambda_{ls} s_{1,t} + \varepsilon_{l,t}, \tag{C.2}$$

where $\varepsilon_{l,t}$ follows $\mathcal{N}(0, \sigma^2_l)$, μ_l is the average growth rate, and λ_{ls} captures the correlation with stock returns.

Income risk Labor earnings can be decomposed as the product of the wage index and an idiosyncratic component $L_{2,it}$:

$$L_{it} = L_{1,t} \cdot L_{2,it}. \tag{C.3}$$

The idiosyncratic component is further decomposed into a deterministic function of age f_{it}[8], a persistent component z_{it} and a transitory shock η_{it}:

$$L_{2,it} = e^{f_{it} + z_{it} + \eta_{it}}. \tag{C.4}$$

The persistent component follows an AR(1) process, with innovations drawn from a normal mixture. Specifically, the dynamics of z_i are given by

$$z_{it} = \rho_z z_{it-1} + \zeta_{it}, \tag{C.5}$$

where

$$\zeta_{it} = \begin{cases} \zeta^-_{it} \sim \mathcal{N}\left(\mu^-_{z,t}, \sigma^{-2}_z\right) & \text{with probability} \quad p_z \\ \zeta^+_{it} \sim \mathcal{N}\left(\mu^+_{z,t}, \sigma^{+2}_z\right) & \text{with probability} \quad 1-p_z \end{cases} \tag{C.6}$$

The values of p_z, $\mu^-_{z,t}$ and $\mu^+_{z,t}$ control the degree of asymmetry in the distribution of income shocks. To capture the cyclicality of skewness, $\mu^-_{z,t}$ is an affine function of the log growth rate of the wage index:

$$\mu^-_{z,t} = \overline{\mu^-_z} + \lambda_{zl}(l_{1,t} - l_{1,t-1}). \tag{C.7}$$

where $p_z \mu^-_{z,t} + (1-p_z)\mu^+_{z,t} = 0$ and $p_z \leq 0.5$. If $\sigma^-_z \gg \sigma^+_z$, p_z represents the frequency of significant events in a worker's career. Finally, the transitory component of income is also modeled as a mixture of normals whose first and second components always coincide with the first and second components

[8] Specically, we assume f to be a cubic polynomial function of age $\theta_2 \text{age}^3/100 + \theta_1 \text{age}^2/10 + \theta_0$.

of the normal mixture governing the innovations to z_i.

$$\eta_{it} = \begin{cases} \eta_{it}^{-} \sim \mathcal{N}(0, \sigma_{\eta}^{-2}) & \text{if} \quad \zeta_{it} = \zeta_{it}^{-} \\ \eta_{it}^{+} \sim \mathcal{N}(0, \sigma_{\eta}^{+2}) & \text{if} \quad \zeta_{it} = \zeta_{it}^{+} \end{cases} \tag{C.8}$$

Social Security Social Security payroll taxes represent 12.4% of the agent's earnings below the maximum taxable earnings, which represents 2.5 times the national wage index.

$$\mathrm{T}_{it} = .124 \cdot \min\left\{L_{it}, 2.5 \cdot L_{1,t}\right\}. \tag{C.9}$$

Retirement benefits depend on historical taxable earnings, adjusted for the growth in the national wage index. Specifically, the agent's Social Security benefits B are:

$$\frac{B_i}{L_{1,R}} = \begin{cases} .9 \cdot S_{iR} & \text{if } S_{iR} < .2 \\ .116 + .32 \cdot S_{iR} & \text{if } .2 \leq S_{iR} < 1 \\ .286 + .15 \cdot S_{iR} & \text{if } 1 \leq S_{iR}, \end{cases} \tag{C.10}$$

where R is the retirement age and $L_{1,R}$ is the value of the wage index at that age. The variable S_{it} keeps track of a worker's average taxable idiosyncratic earnings:

$$S_{it} = \sum_{k=t_0}^{t} \frac{\min\left\{L_{2,ik}, 2.5\right\}}{t - t_0 + 1}, \tag{C.11}$$

where t_0 denotes his first year of earnings.

The parameters of income, stock market and social security are drawn from Catherine et al. (2022a) and summarized in Table C.1.

Table C.1: Preset Parameters of Life-cycle Model

p_z	.136	r	.02	p_s	.146
ρ_z	.967	θ_1	.1237	μ_s^-	-.245
μ_z^-	-.086	θ_2	-.0125	μ_s^+	.115
λ_{zl}	4.291	θ_0	-3.015	σ_{s1}	.077
σ_z^-	.562	t_0	23	σ_{s2}	.114
σ_z^+	.037	R	65	μ_l	.008
σ_η^-	.895	T	100	λ_{ls}	.161
σ_η^+	.089			σ_l	.017

Notes. This table shows the calibrated parameters used in the estimation of the life-cycle model introduced in Section C.

Finally, wealth evolves as:

$$W_{it+1} = \left[W_{it} + L_{it} + B_{it} - \mathrm{T}_{it} - C_{it} - c_{it}\right] \cdot \left[\pi_{it} e^{s_t} + (1 - \pi_{it}) e^{r}\right], \tag{C.12}$$

where π_{it} is the share of his wealth invested in equity. Owning stocks incurs a cost $c_{it} = \Phi L_{1,t}$ if $\pi_{it} > 0$. Short selling or leveraging are not allowed, such that $0 \leq \pi_{it} \leq 1$.

C.2 Construction of data moments

We compute the three core data moments using the 1989–2016 waves of the triennial Survey of Consumer Finances (SCF). We restrict the sample to households whose head is between age 22 and 99 and have positive net worth. The three "core" moments are estimated as:

- m_1, the mean wealth, is measured using the net worth variable (*networth*) from the SCF summary extract public data; note that to improve comparability across survey years, we scale wealth by the average wage income ($wageinc$) of each survey year
- m_2 is the average participation rate, which is the share of households whose total holdings of stock (*equity*) is strictly positive
- m_3 the mean conditional equity share, which is the total holdings of stock (*equity*) divided by net worth, excluding vehicles (*vehic*), and is only computed for households with strictly positive holdings of stocks.

Table C.2: Estimated Moments

		Data	True SMM	Approximate	Simulation
m_1	**mean(wealth)**	5.633 (0.028)	5.633	5.633	5.620
m_2	**participation rate**	0.557 (0.002)	0.557	0.557	0.528
m_3	**mean(cond. equity share)**	0.345 (0.003)	0.345	0.345	0.340
m_4	median(wealth)	1.996 (0.013)	2.919	2.908	2.936
m_5	median(cond. equity share)	0.224 (0.002)	0.269	0.273	0.259

Notes. This table reports the moments targeted in estimation in bold fonts and untargeted moments representing median of statistics in regular fonts. Column "Data" shows the empirical moments, with standard errors in parenthesis. Column "True SMM" corresponds to the simulated moments for the parameters using the true SMM estimation. Column "Approximate" show the approximate moments at the approximate parameter estimates. Column "Simulation" reports the true simulated moments at the approximate parameter estimates. The moments used in the estimation are defined in Section 5.3.1.

C.3 Building training and validation samples

We restrict the set of parameters values $\mathcal{P}$ for the household finance model to: $\gamma \in [1; 20]$, $\beta \in [.5; 1]$, $\Phi \in [0; .25]$.

To build the training sample, we first draw a Halton sequence of $n = 2,000$ parameters (we use a smaller training sample to explore the effect of sample size). For each draw θ_i, we simulate the model and compute the model-generated moments to be matched. We then remove 8 points which correspond to absurd observations (wealth higher than 1,000 the national income). We end up with a training sample of 1,992 observations.

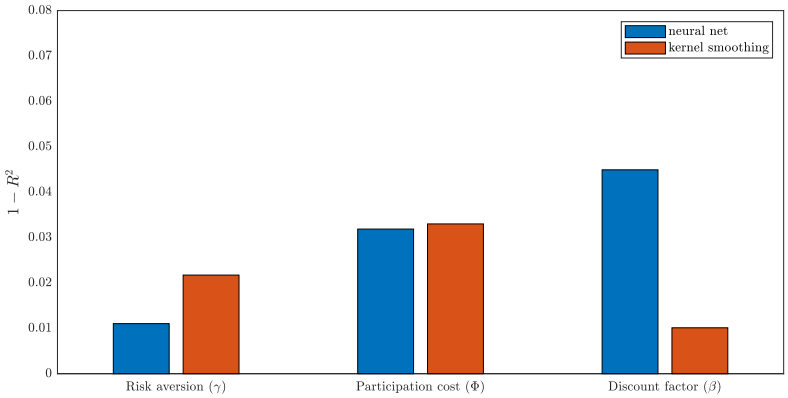
To measure the quality of our approximation, we also build a separate "validation" dataset starting with 200 additional uniform draws of parameters from $\mathcal{P}$ and their corresponding moments. We then

remove 7 absurd draws. Then, like in the corporate finance model, we remove "badly identified" draws. We label a draw as "badly identified" when the s.e. implied by the formula $(G^\top W^{-1} G)^{-1}$[9] is more than 100 times larger than the s.e. of SMM estimate using the data moments. We remove such draws and end up with 106 points in the validation sample.

C.4 Fit of NN versus Kernel smoother

We compare here the quality of our estimation for two types of approximate moment function: the neural net (a MLP with 5 layers and 256, 128, 64, 32, and 16 nodes) and a kernel smoother (). For each specification, we first train it on the training sample, and then use the fitted function f_n to estimate parameters off of the moments in the validation sample. We then report, in Figure C.1, the R^2 between estimated and actual parameters.

Figure C.1: Precision of Approximate Estimates on the Validation Sample



Notes. This figure reports a measure of the estimation error from the approximate SMM for different approximations. For each of the four approximation functions, and each one of the 7 parameters plus the value loss, we report $1 - R^2$, one minus the R^2 of a linear regression of the approximate moment estimate on the true parameter value. Such linear regressions are estimated on the validation sample. If the approximate moment estimate is exactly equal to the true value, this quantity is zero. When it is uncorrelated, it is 1. We consider the following specification for the approximation $f()$: local linear fit, local third-order polynomial fit, a neural net with 5 layers and 10 nodes and kernel smoothing with Gaussian kernel.

First, we see that the R^2 of the estimates using NN approximation is always higher than 95%, highest for risk aversion and lowest for the discount factor. So NNs offer a good approximation of the moment function, even with a smaller number of parameters.

Second, the kernel smoother now does quite well, even better than our deep NN for the discount factor. We conjecture that this is because the HF model has a lower dimensionality than the CF model (3 instead of 7 parameters). At low dimensions, our convergence results suggests that convergence speed is actually higher for the kernel smoother. Another potential explanation for the relatively good performance of the kernel is that the training sample size may be too low to properly fit a deep NN, even though moments are reasonably smooth.

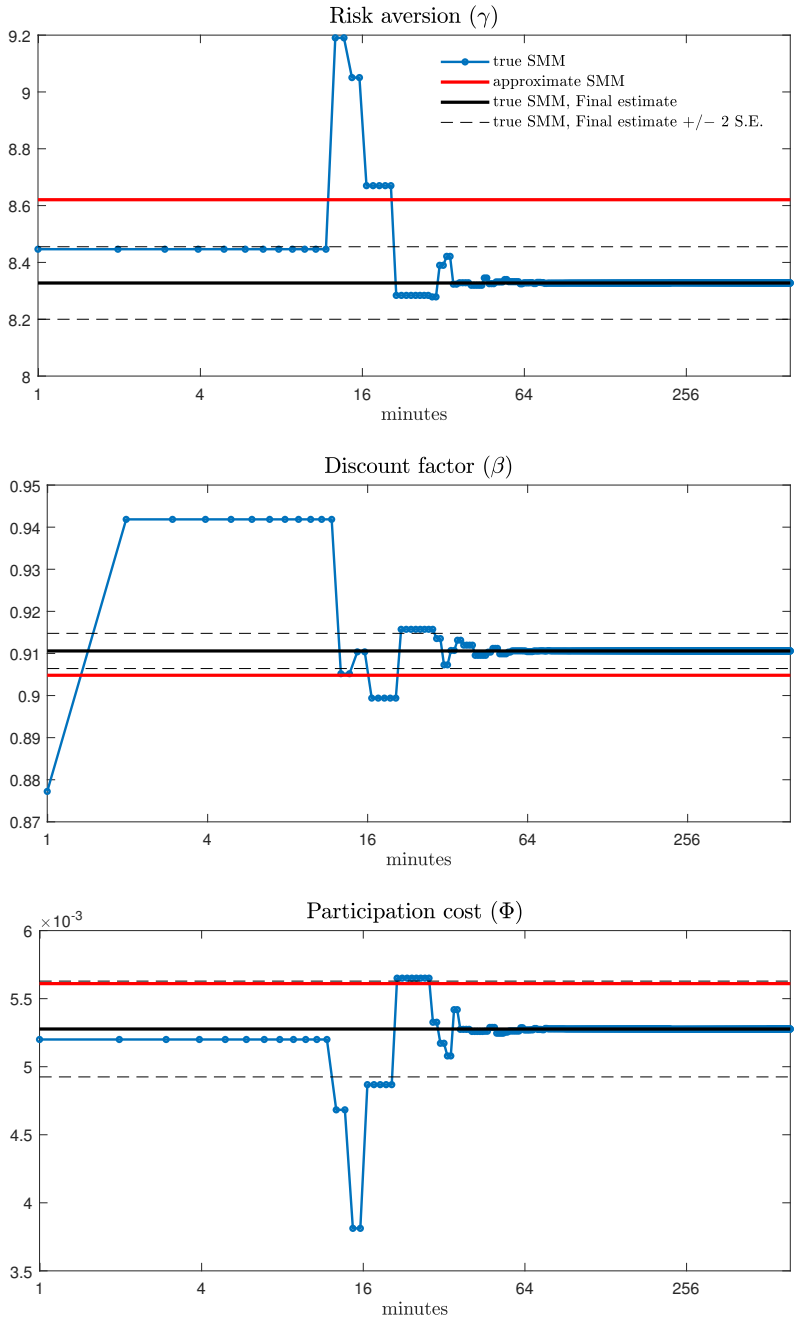
[9] In this formula, W is the variance matrix of the three main moments and G the approximate Jacobian matrix (using f_n). We expect using the true Jacobian not to make much of a difference here.

C.5 Speed gains from the NN approximation

Figure C.2 shows that our approach is faster than the standard SMM. Like for the corporate finance model, the SMM estimation is based on a TikTak algorithm, with a first evaluation of the SMM criterion for 2,000 random points, and then Nelder-Mead optimizations at the best 5 starting points (with at most 200 iterations per optimization). So the compute time we report for our method as well as the SMM abstract from the first 2,000 draws, which both methods have to perform

As shown in Figure C.2, our method is orders of magnitude faster than the SMM. It takes about 30mn for the SMM to converge to the true value, while the approximate SMM does so almost instantaneously (about 0.45s). The slight difference with the corporate finance model is that the approximate estimation somewhat differs from the SMM estimate (significantly only for risk aversion). We attribute this to the fact that our training sample is too small here (since deep NNs are data hungry). With a larger one, we conjecture that the approximate moment method would do better.

Figure C.2: True SMM Estimates, Convergence Speed of Local Optimization Stage (Household Finance Model)



Notes. We report, for the household finance model, parameter values estimated as a function of time taken via two different algorithms in our numerical setup. The blue line corresponds to the local optimization stage of the true SMM, i.e. the minimization of the distance of empirical moments to the true model $f(\theta)$. The optimization algorithm used in this case is Tiktak, using 5 starting points selected from a training set of 2,000 cases and Nelder-Mead algorithm for local optimization per starting point with 200 max function iteration. The red line corresponds to the benchmark approximate SMM. The approximate estimation requires .2 seconds — hence the red line jumps to its final value at the origin. The black line corresponds to the true SMM estimate and the dashed lines represent the confidence interval (+/-2 standard errors) around it.

D Appendix Figures and Tables

Table D.1: Literature

Sub-field	Year	Reference	Comparative statics	Jacobian	AGS	N/A
Investment, capital structure, and financing	1992	Whited JF				✓
	2003	Love RFE				✓
	2005	Hennessy et al. JF				✓
	2007	Hennessy et al. JF		✓		
	2011	DeAngelo et al. JFE	✓			
		Lin et al. JFE				✓
	2013	Matvos RFS				✓
	2014	Nikolov et al. JF	✓			
	2016	Warusawitharana et al. RFS	✓			
		Li et al. RFS				✓
	2017	Bakke et al. JFE	✓			
		Gu JFE			✓	
	2018	Wu RFS		✓	✓	
	2019	Nikolov et al. JFE	✓			
	2021	Frank et al. RFS				✓
		Begenauet al. JFE				✓
	(Forthcoming)	Catherine et al. JF	✓	✓	✓	
Corporate governance	2009	Gayle et al. AER				✓
	2010	Taylor JF				✓
		Kang et al. JFE				✓
	2012	Coles et al. JFE		✓		
	2013	Taylor JFE				✓
	2017	Jung et al. JFE		✓		
	2018	Page JFE			✓	
	2022	Bertomeu JFE				✓
Bankruptcy	2016	Glover JFE				✓
Banking	2014	Schroth et al. JFE		✓		
Corporate control	2014	Dimopoulos et al. JFE	✓			
	2015	Albuquerque et al.JF	✓			
	2018	Li et al. JFE		✓		
		Wang JFE				✓
	2020	Wang JFE	✓			
Entrepreneurship	2020	Jones et al. AER		✓		
	2022	Ewens et al. JFE	✓			
		Catherine JFE	✓			
Household finance	2018	Pagel Econometrica				✓
	2019	Sun et al. JFE	✓			
	2020	Ameriks et al. JPE				✓
	2022	Catherine RFS				✓
Real estate	2015	Corbae et al. JPE				✓
	2017	Landvoigt RFS				✓
	2020	Oh et al. JFE				✓
	2021	Ghent JFE				✓