TEMPORAL-DIFFERENCE ESTIMATION OF DYNAMIC DISCRETE CHOICE MODELS

KARUN ADUSUMILLI[†] AND DITA ECKARDT[*]

Abstract. We study the use of Temporal-Difference learning for estimating the structural parameters in dynamic discrete choice models. Our algorithms are based on the conditional choice probability approach but use functional approximations to estimate various terms in the pseudo-log-likelihood function. We suggest two approaches: The first—linear semi-gradient—provides approximations to the recursive terms using basis functions. The second—Approximate Value Iteration—builds a sequence of approximations to the recursive terms by solving non-parametric estimation problems. Our approaches are fast and naturally allow for continuous and/or high-dimensional state spaces. Furthermore, they do not require specification of transition densities. In dynamic games, they avoid integrating over other players' actions, further heightening the computational advantage. Our proposals can be paired with popular existing methods such as pseudo-maximum-likelihood, and we propose locally robust corrections for the latter to achieve parametric rates of convergence. Monte Carlo simulations confirm the properties of our algorithms in practice.

1. Introduction

Recent years have seen a number of important developments in the field of Reinforcement Learning (RL) for computation of value functions. The goal of this paper is to study the use of a popular RL technique, Temporal-Difference (TD) learning, for estimation and inference in Dynamic Discrete Choice (DDC) models.

DDC models are frequently used to describe the inter-temporal choices of forward-looking individuals in a variety of contexts. In these models, agents maximize their expected future payoff through repeated choice amongst a set of discrete alternatives. Based on a revealed preference argument, structural estimation proceeds by using

Key words and phrases. Dynamic discrete choice models, Dynamic discrete games, Temporal-Difference learning, Reinforcement Learning.
This version: August 28, 2025.
[†]Department of Economics, University of Pennsylvania; akarun@sas.upenn.edu.
[*]Department of Economics, University of Warwick; dita.eckardt@warwick.ac.uk.
We would like to thank the editor and three anonymous referees for valuable comments that substantially improved the paper. Thanks also to Xiaohong Chen, Frank Diebold, Aviv Nevo, Whitney Newey and Frank Schorfheide for helpful discussions.

microdata on choices and outcomes to recover the underlying model parameters.[1] A key challenge in this literature is the complexity of estimation. Uncovering the structural parameters originally required an explicit solution to a dynamic programming problem in addition to the optimization of an estimation criterion. A key advance has been Hotz and Miller's (1993) Conditional Choice Probability (CCP) algorithm which avoids the repeated solution of the inter-temporal optimization problem by taking advantage of a mapping between value function differences and conditional choice probabilities.

Unfortunately, the standard CCP algorithm of Hotz and Miller (1993) is computationally infeasible when the underlying states are continuous and/or the state space is high-dimensional. Such state spaces are common in applications. One approach to tackle continuous state spaces is through state space discretization, e.g., Kalouptsidi (2014) and Almagro and Domínguez-Iino (2025) use aggregation and clustering methods to do this. However, it is not always clear how to perform such a discretization in practice, and moreover, it introduces bias into the parameter estimates. An alternative is to employ functional approximations for the value functions. For instance, Barwick and Pathak (2015) and Kalouptsidi (2018) use estimated transition densities and numerical/analytical integration to approximate the value functions using linear sieves and LASSO, respectively. However, the theoretical properties of these methods when using machine learning methods (such as LASSO) are as yet unknown, and they still require estimation of transition densities, which is not straightforward, along with numerical integration, which can be computationally expensive.[2]

The aim of this paper is to develop tractable algorithms for CCP estimation when the state variables are continuous and/or the state space is large. Such algorithms should possess three properties: First, they should be fast to compute even under high-dimensional state spaces. Second, they should avoid state space discretization, and instead rely on functional approximation of value functions. Third, they should avoid estimation of transition densities which are difficult to parameterize and estimate under continuous states. If the DDC model in question satisfies either finite dependence or a terminal state property there already exist algorithms possessing these properties, see, e.g., Ackerberg et al. (2014) and Chernozhukov et al. (2022). Our interest here is in developing general purpose algorithms that do not require these assumptions.

[1]See Aguirregabiria and Mira (2010) for a survey of the literature on the estimation of DDC models.
[2]Yet another alternative is to use forward Monte Carlo simulations (Bajari et al., 2007, Hotz et al., 1994), but this again becomes very involved as the number of continuous state variables or players increases. The use of a finite number of Monte Carlo simulations also adds bias to the estimates.

We suggest two methods, based on TD learning, that satisfy all the above properties. The methods involve two different techniques for estimating recursive terms (which are akin to value functions) that arise in CCP estimation. The first approach, the linear semi-gradient method, provides functional approximations to the recursive terms using basis functions. This simply involves inverting a matrix whose dimension is the number of basis functions, so the computational cost is generally trivial. Furthermore, it only requires the observed sequences of current and future state-action pairs as input and estimation of transition densities is not needed. The second approach, Approximate Value Iteration (AVI), builds a sequence of approximations to the value terms by solving a non-parametric estimation problem in each step. Almost any machine learning (ML) method for prediction can be used for the approximation, including (but not limited to) LASSO, Random Forests and Neural Networks. To our knowledge, the AVI method is the first estimator for general DDC models that can be applied with any ML method that achieves suitable rates of convergence. Hence, it naturally allows for very high-dimensional state spaces. Again, no estimation of transition densities is required. We derive the non-parametric rates of convergence for estimation of the value terms under both methods. Using the estimates of these functions, estimation of the structural parameters can proceed with standard methods such as pseudo-maximum-likelihood estimation (PMLE, Aguirregabiria and Mira (2002)) or minimum distance estimation.

The focus of this paper is on the estimation of structural parameters. To this end, our procedures avoid modeling state transitions. Performing counterfactual analysis may still require estimating the transition density, but we argue that our techniques remain useful, even for this purpose, for two reasons: First, counterfactuals often involve transition densities which are different from the ones that enter the estimation of the structural parameters, see e.g., Kalouptsidi (2018). Our methods thus avoid estimation of the original transition densities. Second, with continuous states, decoupling the estimation of structural parameters and transition densities may be beneficial for robustness and efficiency. For instance, it is common to employ AR (e.g., Aguirregabiria and Mira, 2007; Kalouptsidi, 2014) or VAR (Barwick and Pathak, 2015) specifications for transition densities. However, these specifications involve a number of choices (e.g., dimension of VARs, distribution of error terms etc.), which the structural parameter estimates may not be robust to. Importantly, even when non-parametric estimates of transition densities are available, plugging them into the second-stage PMLE criterion would seriously degrade

the rate of convergence of structural parameters. One would need to adjust the PMLE to account for the non-parametric first stage, but the form of this adjustment is not known. By contrast, our proposals use non-parametric estimates of value functions, and as described below, we derive the necessary adjustments to account for this. To perform counterfactual analysis, we suggest combining our estimates of the structural parameters - which do not rely on non-parametric estimates of transition densities and are robust to mis-specification - with non-parametric estimates of the transition densities.

The previous discussion highlights that in continuous state spaces, estimation of structural parameters is inherently a problem of semi-parametric estimation. In fact, even under discrete states, estimation of transition densities affects the variance of the structural parameter estimates, see Aguirregabiria and Mira (2002). If the state variables are continuous, existing two-step CCP methods such as the PMLE are no longer $\sqrt{n}$-consistent. We therefore derive a locally robust estimator by adding a correction term to the PMLE criterion function that accounts for the non-parametric estimation of value function terms using either of our TD methods. This construction is novel and does not directly follow from existing results, e.g., in Chernozhukov et al. (2022).[3] The resulting estimator converges at parametric rates under continuous states and unrestricted transition densities.

Our TD estimators are thus consistent, converge at parametric rates, and provide a feasible estimation method when the states are continuous and/or the state space is large. The latter is particularly important for the estimation of dynamic discrete games. Existing methods for the estimation of dynamic games (Bajari et al., 2007; Aguirregabiria and Mira, 2007; Pesendorfer and Schmidt-Dengler, 2008) require integrating out other players' actions, which can get quite cumbersome with many players, or under continuous states. By contrast, our procedure works directly with the joint empirical distribution of the states and their sample successors. Thus the 'integrating out' is done implicitly within the sample expectations.

[3]Independently of our work and around the release of our initial draft, Chernozhukov et al. (2019) derived orthogonal moment conditions for weighted averages of value functions of the form $\theta_0 = \mathbb{E}[w(x)V(x)]$, where $V(\cdot)$ is an estimated value function and $w(\cdot)$ is a known weight function. Their approach to deriving the orthogonal moment is different from ours as it is based on the methods by Ichimura and Newey (2022), but similarly to us, it results in a debiasing function involving backward projections, where the current state is regressed on a future one. Subsequent versions of Chernozhukov et al. (2019), released after the first revision of our paper, extend their framework to estimate parameters of the form $\theta = \mathbb{E}[m(x, V)]$, where $m(x, \cdot)$ is a nonlinear functional of V. In this generalized setting, their construction of the locally robust correction term involves iterating the local-Riesz representer $\alpha(\cdot)$ for $\mathbb{E}[m(x, V)]$ backwards in what the authors call a 'dynamic dual representation'. This approach closely parallels the methodology of our paper.

Finally, we also incorporate permanent unobserved heterogeneity into our methods by combining the TD estimation with an Expectation-Maximization (EM) algorithm.

A range of Monte Carlo studies confirm the workings of our algorithms. First, we present simulations based on the dynamic firm entry problem described in Aguirregabiria and Magesan (2018). The model has seven structural parameters and five continuous state variables. Existing methods often struggle at this dimensionality; certainly, state space discretization would not work too well. We provide simulations for the linear semi-gradient and the AVI method with Random Forests, with and without locally robust corrections. Our estimators perform very well in this setting and they outperform CCP estimators that employ discretization, leading to a 10-fold reduction in average mean squared error across the structural parameters. They also perform similar to or outperform alternative methods such as the 2-step Euler-Equation (EE) approach of Aguirregabiria and Magesan (2018), even though the latter only applies to a more restricted class of models. Our linear semi-gradient method is even three times faster to compute.

Second, we test our algorithms for dynamic discrete games based on a firm entry game similar to that outlined in Aguirregabiria and Mira (2007). We use the linear semi-gradient method here and, as before, our estimates are closely centered around the true parameters. Since this approach requires the selection of a set of basis functions for the functional approximations, we present results for different sets (a second, third and fourth order polynomial) in this model. Our findings suggest that the choice of basis functions has only a small effect on the performance of the estimator. Moreover, a simple cross-validation procedure may be used to select the preferred set of functions.

1.1. **Related literature.** Rust (1987) is the seminal work in the literature of DDC models. Motivated by computational considerations, Hotz and Miller (1993) propose the CCP algorithm. The CCP idea has subsequently been refined by Hotz et al. (1994) who suggest a simulation-based method, and Aguirregabiria and Mira (2002) who develop a pseudo-likelihood estimator. Arcidiacono and Miller (2011) introduce and exploit the property of finite dependence to speed up CCP estimation. Despite these advances, the estimation of DDC models remains constrained by its computational complexity, particularly in the large class of models where finite dependence does not hold. Estimation of dynamic discrete games is particularly affected by these issues as the strategic interaction of agents means that the state space increases exponentially with the number of players. It is also uncommon for finite dependence to hold in dynamic games.

The standard CCP algorithm is a two-step method, and is known to suffer from severe bias in finite samples. Aguirregabiria and Mira (2002; 2007) address this issue by presenting a recursive CCP estimator, the nested pseudo-likelihood (NPL), that is equivalent to the nested-fixed-point estimator (NFXP, Rust, 1987). Both are in turn equivalent to partial-MLE, which employs a plug-in estimate of the transition density in the MLE criterion, but they are not fully efficient (they are not equivalent to full-MLE). In fact, with continuous states, estimation of the transition density introduces bias that is the dominant term in determining the rate of convergence. This motivates the construction of our locally robust estimator which gets rid of this bias and restores parametric rates. In fact, our proposal can be more efficient than these methods even with a parametric model for the transition density, see Section 4.2.2.

Ackerberg et al. (2014) and Chernozhukov et al. (2022) consider semi-parametric estimation using ML methods when either finite dependence or a 'terminal action' property holds (Hotz and Miller, 1993).[4] Chernozhukov et al. (2022) also derive locally robust corrections for this setting. Under finite dependence the PMLE criterion can be written as a function of choice probabilities only (transition densities are not required); the authors employ non-parametric estimates for choice probabilities and correct for this estimation in the second stage. Computation and estimation is thus relatively simpler under finite dependence. By contrast, our methods are applicable to the more general and difficult setting where finite dependence may not apply. Nevertheless, the computational speed of our linear semi-gradient procedure is comparable to methods that exploit finite dependence. For dynamic games, Semenova (2018) allows for high-dimensional state spaces, but the approach it is based on, due to Bajari et al. (2007), is not efficient, e.g., it may only partially identify parameters even if the model is fully identified. On the other hand, it allows for continuous actions, unlike our method.

In making use of TD learning, our methods relate to the literature on RL, particularly batch RL. Batch RL describes learning about how to map states into actions to maximize an expected payoff, using a fixed set of data (a so-called batch); see Lange et al. (2012) for a survey.[5] A key step in RL, including batch RL, is the estimation of value functions. TD learning methods, first formulated by Sutton (1988), are the most commonly used set of algorithms for this purpose. We study non-parametric estimation

[4]In a different application of ML methods in this context, Norets (2012) suggests combining Neural Networks with a Bayesian MCMC approach.
[5]See Sutton and Barto (2018) for a detailed treatment of RL in genreal.

of value functions using two TD methods: semi-gradients and AVI. Our analysis builds on the techniques developed by Tsitsiklis and Van Roy (1997) for linear semi-gradients, and Munos and Szepesvári (2008) for AVI. While Tsitsiklis and Van Roy (1997) focus on online learning (i.e., where collection of data and estimation of value functions is conducted simultaneously), we translate their methods to batch learning.[6] With regards to Munos and Szepesvári (2008), we differ in employing assumptions that are more common to econometrics and our characterization of the rates is also different (compare Theorem 2 in their paper with our Theorem 3).

TD methods are distinct from other value function approximation methods developed in economics, e.g., parametric policy iteration (Benítez-Silva et al., 2000), simulation and interpolation (Keane and Wolpin 1994), and sieve value function iteration (Arcidiacono et al., 2013). The last of these is similar in spirit to AVI with linear functional approximations. However, our semi-gradient method provides a linear approximation in a single step without any need for iterations, and we analyze AVI under generic machine learning methods. Our approximation results, and the technical arguments leading to them, are thus different from Arcidiacono et al. (2013); in fact, their setting is different too as the authors focus on estimating the 'optimal' value function, while the recursive terms in our setting are more akin to a value function under a fixed policy.

2. Setup

We start with an infinite horizon single-agent DDC model in discrete time, where observations consist of $i = 1, \ldots, n$ agents. We assume that the agents are homogeneous, relegating extensions to unobserved heterogeneity to Online Appendix C. In each period, an agent chooses among A mutually exclusive actions, denoted by a. Choosing a when the current state is x gives the agent an instantaneous utility of $z(a,x)^\intercal\theta^* + e$, where $z(a,x)$ is a known vector-valued function of a, x and e is an idiosyncratic error term. We denote the realized state of an agent i at time t by x_{it}, and her corresponding action and error by a_{it} and e_{it}. We assume that e_{it} is an iid draw from some known distribution $g_e(\cdot)$. Let (a', x') denote the one-period ahead random variables following the actions and states (a, x), where $x' \sim K(\cdot|a,x)$, with $K(\cdot|a,x)$ denoting the transition density given a, x (more precisely, it is the Markov kernel). We do not make any parametric assumptions about $K(\cdot|a,x)$. The utility from future periods is discounted by β.

[6] See also Chen and Qi (2022) for related results on Q-learning under series approximations.

Agent i chooses actions $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots)$ to sequentially maximize the discounted sum of payoffs

$$E\left[\sum_{t=1}^{\infty} \beta^t \left\{z(x_{it}, a_{it})^\intercal \theta^* + e_{it}\right\}\right].$$

The econometrician observes a panel consisting of state-action pairs for all individuals, $(\mathbf{x}_i, \mathbf{a}_i) = \{(x_{i1}, a_{i1}), \dots, (x_{iT}, a_{iT})\}$, for T periods (note, however, that the agent maximizes an infinite horizon objective, not a fixed T one). Typically $T \ll n$ in applications, so we work within an asymptotic regime where $n \to \infty$ but T is fixed. Using this data, the econometrician aims to recover the structural parameters θ^*.

In this paper, we study the CCP approach for estimating θ^* (Hotz and Miller, 1993). CCP methods are based on the conditional choice probabilities of choosing action a given state x. We denote these by $P_t(a|x)$ for a given period t but henceforth drop the subscript t with the idea that it can be made a part of the state variable x, if needed (we should also add that some of our theoretical results are based on assuming stationarity, i.e., $P_t(a|x)$ is independent of t). Denote $e(a, x)$ as expected value of the idiosyncratic error term e given that action a was chosen. Hotz and Miller (1993) show that if the distribution of e follows a Generalized Extreme Value (GEV) distribution, it is possible to express $e(a, x)$ as a function of the choice probabilities $P(a|x)$, i.e., $e(a, x) = \mathcal{G}(P(a|x))$. We assume that e follows a Type I Extreme Value distribution, which is perhaps the most common choice in applications. In this case $e(a, x) = \gamma - \ln P(a|x)$, where γ is the Euler constant.

Using the standard CCP approach, under the given distributional assumptions, the parameters are obtained as the maximizers of the pseudo-log-likelihood function

$$Q(\theta) = \sum_{i=1}^{n} \sum_{t=1}^{T-1} \ln \frac{\exp\left\{h(a_{it}, x_{it})^\intercal \theta + g(a_{it}, x_{it})\right\}}{\sum_a \exp\left\{h(a, x_{it})^\intercal \theta + g(a, x_{it})\right\}}, \tag{2.1}$$

where for any (a, x), $h(\cdot)$ and $g(\cdot)$ solve the following recursive expressions:[7]

$$\begin{aligned} h(a, x) &= z(a, x) + \beta \mathbb{E}\left[h(a', x')|a, x\right], \\ g(a, x) &= \beta \mathbb{E}\left[e(a', x') + g(a', x')|a, x\right]. \end{aligned} \tag{2.2}$$

Here, $\mathbb{E}[\cdot|a, x]$ denotes the expectation over the distribution of (a', x') conditional on (a, x); it is a function of $K(\cdot|a, x), P(\cdot|x)$. Both $h(a, x)$ and $g(a, x)$ have a 'value-function' form, which turns out to be useful for our approach.

[7]Note that $h(a_{it}, x_{it}) = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \beta^{(\tau-t)} z(a_{i\tau}, x_{i\tau}) \middle| a_{it}, x_{it}\right]$, i.e., we can interpret $h(a_{it}, x_{it})^\intercal \theta$ as the expected discounted utility (excluding the error term) given the current state a_{it}, x_{it}. A similar interpretation holds for $g(\cdot)$. See Aguirregabiria and Mira (2010) for a further description.

Clearly, $h(\cdot)$ and $g(\cdot)$ are functions of $K(\cdot|\cdot)$ and $P(\cdot|\cdot)$. Since the latter are unknown, current literature generally proceeds by first estimating these as $(\hat{K}, \hat{P})$. Typically, $\hat{K}$ is obtained by MLE based on a parametric form of $K(x'|a, x; \theta_f)$, while $\hat{P}$ is estimated non-parametrically using either a blocking scheme or kernel regression. Then, given $(\hat{K}, \hat{P})$, $h(\cdot)$ and $g(\cdot)$ are estimated by solving the recursive equations (2.2). In the next section, we propose an alternative algorithm for maximizing $Q(\theta)$ that directly estimates $h(\cdot)$ and $g(\cdot)$ in a single step without requiring any knowledge about or estimation of $K(\cdot|\cdot)$.

Notation. We assume that the distribution of $(a_{it}, x_{it}, a_{it+1}, x_{it+1})$ is time stationary. This greatly simplifies our notation. It is not necessary for our results on the approximation properties of our TD methods, see Appendix A, but we do require it for deriving a locally robust estimator. Since the transition density and choice probabilities are time independent (the latter due to Blackwell's theorem), the stationarity assumption is equivalent to supposing that $\{a_{it}, x_{it}, a_{it+1}, x_{it+1}\}_i$ are random draws from the ergodic, i.e., long-run distribution of (a, x, a', x').[8] Let $\mathbb{P}$ denote such a distribution over (a, x, a', x'), and $\mathbb{E}[\cdot]$ the corresponding expectation over $\mathbb{P}$. Define $\mathbb{E}_n[\cdot]$ as the expectation over the empirical distribution, $\mathbb{P}_n$, of (a, x, a', x'). In particular, $\mathbb{E}_n[f(a, x, a', x')] := (n(T-1))^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T-1} f(a_{it}, x_{it}, a_{it+1}, x_{it+1})$, i.e., we always drop the last time period in the summation index even if $f(\cdot)$ does not depend on a', x'.

Let $\mathcal{H}$ denote the space of all square integrable functions over the domain $\mathcal{A} \times \mathcal{X}$ of (a, x). Define the pseudo-norm $\|\cdot\|_2$ over $\mathcal{H}$ as $\|f\|_2 := \mathbb{E}[|f(a, x)|^2]^{1/2}$ for all $f \in \mathcal{H}$. We use $|\cdot|$ to denote the usual Euclidean norm on a Euclidean space.

3. Temporal-difference estimation

This section presents our TD estimation of $h(\cdot)$ and $g(\cdot)$. Note that $h(\cdot)$ is a vector of the same dimension as θ^*. Our methods provide functional approximations separately for each component $h^{(j)}$ of h. To simplify notation, we drop the superscript j indexing the elements of $h(\cdot)$ and proceed as if the latter, and therefore θ^*, is a scalar. However, all our results hold for general $h(\cdot)$, as long as each of its elements is treated separately.

[8] This is a slightly stronger requirement than the one imposed by Aguirregabiria and Mira (2002), who assume only that $\{a_{it}, x_{it}, a_{it+1}\}_i$ are i.i.d. draws from a distribution $\breve{\mathbb{P}}$ satisfying $\breve{\mathbb{P}}(x_{it} = x) > 0$ for all $x \in \text{support}(X)$. In the discrete case, $\breve{\mathbb{P}}$ and the ergodic distribution $\mathbb{P}$ are mutually absolutely continuous, with $d\breve{\mathbb{P}}/d\mathbb{P} \leq C < \infty$. As a result, replacing $\mathbb{P}$ with $\breve{\mathbb{P}}$ would only introduce a constant multiplicative factor in our results, without altering the convergence rates. Moreover, as Aguirregabiria and Mira (2002) note, the i.i.d. assumption is often used as a convenient approximation for time-series dynamics. But the equivalence between time-series and cross-sectional analysis holds only under the ergodic distribution, which guarantees that cross-sectional expectations coincide with long-run time averages.

For any candidate function, $f(a, x)$, for $h(a, x)$, denote the TD error by

$$\delta_z(a, x, a', x'; f) := z(a, x) + \beta f(a', x') - f(a, x),$$

and the dynamic programming operator by

$$\Gamma_z[f](a, x) := z(a, x) + \beta \mathbb{E}[f(a', x')|a, x].$$

Clearly, $h(a, x)$ is the unique fixed point of $\Gamma_z[\cdot]$. TD estimation involves approximating $h(a, x)$ using a functional class $\mathcal{F}$, where each element $h(\cdot; \omega)$ of $\mathcal{F}$ is indexed by a finite-dimensional vector ω. The aim is to ostensibly minimize the mean-squared TD error

$$\text{TDE}(\omega) := \mathbb{E}\left[\|z(a, x) + \beta h(a', x'; \omega) - h(a, x; \omega)\|^2\right].$$

However, this minimization problem is neither computationally feasible nor is it proven to converge when the true $h \notin \mathcal{F}$. Instead, two approaches are commonly used.

The first approach, the semi-gradient method, involves updating ω as

$$\omega_{j+1} = \omega_j + \alpha \mathbb{E}\left[\{z(a, x) + \beta h(a', x'; \omega_j) - h(a, x; \omega_j)\} \nabla_\omega h(a, x; \omega_j)\right] \tag{3.1}$$

for some small value of α. As the name suggests, the above is not a complete gradient as the derivative does not take into account how ω affects the 'target', i.e., the future value $h(a', x'; \omega)$. Nevertheless, for linear functional classes $\mathcal{F}$, it is possible to explicitly characterize the limit point of the updates, ω^*, and compute it directly. Section 3.1 describes this in greater detail. In the RL literature, it is common to employ semi-gradients with Neural Networks as the functional class $\mathcal{F}$, but it appears difficult to extend our theoretical analysis to this setting (we can, however, use Neural Networks with our AVI procedure described below).

The second approach, Approximate Value Iteration (AVI; Munos and Szepesvári, 2008), employs the idea of 'target networks'. Here, the parameters in the future value of h are fixed at the current ω, and the functional parameters iteratively updated as

$$\omega_{j+1} = \arg\min_{\omega} \mathbb{E}\left[\|z(a, x) + \beta h(a', x'; \omega_j) - h(a, x; \omega)\|^2\right]. \tag{3.2}$$

Clearly, the semi-gradient method and AVI are closely related: if one were to solve the problem in (3.2) using gradient descent, the updates within each iteration would look similar to (3.1) except for fixing the value of ω in $h(a', x'; \omega)$ at the past values. After the updates converge, i.e., at the end of the iteration, $h(a', x'; \omega)$ is revised with the new

ω. The semi-gradient approach can thus be considered a one-step variant of AVI. Section 3.2 describes AVI in more detail. We characterize the theoretical properties of AVI under general functional classes $\mathcal{F}$ including Neural Networks, Random Forests, LASSO etc.

The approximation to g follows similarly after replacing $\delta_z(\cdot; f), \Gamma_z[\cdot]$ by

$$\delta_e(a, x, a', x'; f) := \beta e(a', x') + \beta f(a', x') - f(a, x),$$
$$\Gamma_e[f](a, x) := \beta \mathbb{E}[e(a', x') + f(a', x')|a, x].$$

3.1. **Semi-gradients.** Let $\phi(a, x)$ consist of a set of basis functions over the domain (a, x). Then the linear approximation class is $\mathcal{F} \equiv \{\phi(a, x)^\intercal \omega : \omega \in \mathbb{R}^{k_\phi}\}$, where $k_\phi = \dim(\phi)$. Denote the projection operator onto $\mathcal{F}$ by P_ϕ:

$$P_\phi[f](a, x) := \phi(a, x)^\intercal \mathbb{E}[\phi(a, x)\phi(a, x)^\intercal]^{-1} \mathbb{E}[\phi(a, x) f(a, x)].$$

For linear basis functions, it can be shown, e.g., Tsitsiklis and Van Roy (1997), that the sequence of functional approximations $h(a, x; \omega_j) := \phi(a, x)^\intercal \omega_j$ converges to $h^* := \phi(a, x)^\intercal \omega^*$, defined as the fixed point of the projected dynamic programming operator $P_\phi \Gamma_z[\cdot]$ (i.e., $P_\phi \Gamma_z[h^*] = h^*$). Based on this characterization, we show in Lemma 1 (Online Appendix B.2) that $h^*(a, x) = \phi(a, x)^\intercal \omega^*$, where

$$\omega^* = \mathbb{E}\left[\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^\intercal\right]^{-1} \mathbb{E}\left[\phi(a, x) z(a, x)\right]. \tag{3.3}$$

Lemma 2 in Online Appendix B.2 assures that $\mathbb{E}\left[\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^\intercal\right]$ is indeed non-singular as long as $\beta < 1$ and $\mathbb{E}[\phi(a, x)\phi(a, x)^\intercal]$ is non-singular. While ω^* cannot be computed directly, we can obtain an estimator, $\hat{\omega}$, by replacing $\mathbb{E}[\cdot]$ with the sample expectation $\mathbb{E}_n[\cdot]$:

$$\hat{\omega} = \mathbb{E}_n\left[\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\phi(a, x) z(a, x)\right]. \tag{3.4}$$

Using $\hat{\omega}$, we estimate $h(\cdot)$ as $\hat{h}(a, x) = \phi(a, x)^\intercal \hat{\omega}$.

We now turn to the estimation of $g(\cdot)$. As with $h(\cdot)$, we approximate $g(\cdot)$ using basis functions $r(a, x)$, which may generally be different from $\phi(a, x)$. Let P_r denote the projection operator onto the space $\{r(a, x)^\intercal \xi : \xi \in \mathbb{R}^{k_r}\}$, where $k_r = \dim(r)$. The limit of the semi-gradient iterations is $g^*(a, x) := r(a, x)^\intercal \xi^*$, defined as the fixed point of $P_r \Gamma_e[\cdot]$. We thus obtain the following characterization of ξ^* in analogy with (3.3):

$$\xi^* = \mathbb{E}\left[r(a, x)\left(r(a, x) - \beta r(a', x')\right)^\intercal\right]^{-1} \mathbb{E}\left[\beta r(a, x) e(a', x')\right]. \tag{3.5}$$

In the above, $e(a,x) = \gamma - \ln P(a|x)$ is a function of unknown choice probabilities. Denote $\eta(a,x) := P(a|x)$. Suppose that we have access to a non-parametric estimator $\hat{\eta}$ of η, e.g., through series or kernel regression. We can then use this estimate to obtain $e(a,x;\hat{\eta}) := \gamma - \ln \hat{\eta}(a,x)$. This in turn enables us to estimate ξ^* using $\hat{\xi}$, computed as

$$\hat{\xi} = \mathbb{E}_n\left[r(a,x)\left(r(a,x) - \beta r(a',x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\beta r(a,x) e(a',x';\hat{\eta})\right]. \tag{3.6}$$

Using the above, we estimate $g(\cdot)$ as $\hat{g}(a,x) = r(a,x)^\intercal \hat{\xi}$. Algorithm 3 in Online Appendix D describes the estimation steps for both $\hat{\omega}$ and $\hat{\xi}$.

Interestingly, estimation of ξ^* is unaffected to a first order by the estimation of $\hat{\eta}$, even though the latter converges to the true η at non-parametric rates (see Section 4 for a formal statement). This is because of an orthogonality property for the estimation of ξ:

$$\partial_\eta \mathbb{E}\left[\beta r(a,x) e(a',x';\eta)\right] = 0, \tag{3.7}$$

where $\partial_\eta \cdot$ denotes the Fréchet derivative with respect to η. To show (3.7), expand

$$\begin{aligned}\mathbb{E}\left[\beta r(a,x) e(a',x';\eta)\right] &= \mathbb{E}\left[\beta r(a,x) \mathbb{E}\left[e(a',x';\eta)\middle| x'\right]\right] \\ &= \mathbb{E}\left[\beta r(a,x) \mathbb{E}\left[\gamma - \ln \eta(a',x')\middle| x'\right]\right], \end{aligned} \tag{3.8}$$

where the first equality follows from the Markov property. Consider the functional $M(\tilde{\eta}) := \mathbb{E}\left[\ln \tilde{\eta}(a',x')\middle| x'\right]$ at different candidate values $\tilde{\eta}(\cdot,\cdot)$. At the true conditional choice probability, η, $M(\tilde{\eta})$ becomes the conditional entropy of $P(a|x')$ and attains its maximum. Hence, $\partial_\eta \mathbb{E}\left[\ln \eta(a',x')\middle| x'\right] = 0$ and (3.7) follows from (3.8). Consequently, $\hat{\xi}$ is a locally robust estimator for ξ.

Computation of $\hat{\omega}$ and $\hat{\xi}$ is very cheap as it only involves solving linear equations of dimension $\dim(\phi)$ and $\dim(r)$, respectively. Using $\hat{h}(a,x)$ and $\hat{g}(a,x)$, we can in turn estimate θ^* in many different ways. For instance, we can use the PMLE estimator

$$\tilde{\theta} = \arg\max_\theta \hat{Q}(\theta); \quad \hat{Q}(\theta) := \sum_{i=1}^n \sum_{t=1}^{T-1} \ln \frac{\exp\left\{\hat{h}(a_{it},x_{it})\theta + \hat{g}(a_{it},x_{it})\right\}}{\sum_a \exp\left\{\hat{h}(a,x_{it})\theta + \hat{g}(a,x_{it})\right\}}. \tag{3.9}$$

However, such plug-in estimates are sub-optimal. In Section 4.2, we suggest a locally robust version of (3.9).

Suppose that the underlying states and actions are discrete, and that our algorithm uses the set of all discrete elements of x, a as basis functions. We show in Online Appendix

B.1 that the resulting estimate of $h(a,x)$ is identical to that obtained from the standard CCP estimators, if the choice and transition probabilities were estimated using cell values.

A limitation of the linear semi-gradient method is that it requires one to choose a series basis and also does not allow for high-dimensional state spaces (*i.e.*, $\dim(x) \propto n$). The AVI method, described below, does not share this limitation.

3.2. **Approximate Value Iteration (AVI).** For a feasible estimation procedure using AVI, we can replace $\mathbb{E}[\cdot]$ by $\mathbb{E}_n[\cdot]$ in (3.2). The procedure builds a sequence of approximations $\{\hat{h}_j; j = 1, \ldots, J\}$ for h, where

$$\hat{h}_{j+1} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_n \left[\left\| z(a,x) + \beta \hat{h}_j(a',x') - f(a,x) \right\|^2 \right]. \tag{3.10}$$

The process can be started with an arbitrary initialization, e.g., $\hat{h}_1(a,x) = z(a,x)$. The maximum number of iterations, J, is only limited by computational feasibility.[9]

The minimization problem (3.10) is equivalent to a prediction problem using the functional class $\mathcal{F}$, where the outcomes are $z(a,x) + \beta \hat{h}_j(a',x')$. Hence, the estimation target for $\hat{h}_{j+1}$ is the conditional expectation $\mathbb{E}[z(a,x) + \beta \hat{h}_j(a',x')|a,x] \equiv \Gamma_z[\hat{h}_j](a,x)$, i.e., each $\hat{h}_{j+1}$ is a non-parametric approximation to $\Gamma_z[\hat{h}_j]$, and in this manner AVI builds a series of approximate value function iterations.

The interpretation of (3.10) as a prediction problem enables us to employ any machine learning method devised for prediction, including (but not limited to) LASSO, Random Forests and Neural Networks. Our theoretical results show that it is possible to estimate h at suitably fast rates under very weak assumptions on the non-parametric estimation rates of machine learning methods.

The estimation procedure for $g(\cdot)$ is similar: we construct a sequence of approximations $\{\hat{g}_j, j = 1, \ldots, J\}$ for g as

$$\hat{g}_{j+1} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_n \left[\left\| \beta e(a',x';\hat{\eta}) + \beta \hat{g}_j(a',x') - f(a,x) \right\|^2 \right]. \tag{3.11}$$

As in Section 3.1, it will be shown that the estimation error of η is first-order ignorable for the estimation of g. Using $\hat{h}(a,x)$ and $\hat{g}(a,x)$, we can, as before, estimate θ^* in many different ways, including the PMLE estimator (3.9).

Compared to the semi-gradient approach, AVI is computationally more expensive as it requires solving J prediction problems (in Section 4.1 we show that in the worst case

[9] In practice, we suggest monitoring $\varepsilon_j^2 := \mathbb{E}_n\left[\|\hat{h}_{j+1} - \hat{h}_j\|^2\right] / \mathbb{E}_n\left[\|\hat{h}_j - \mathbb{E}_n[\hat{h}_j]\|^2\right]$, the L_2 distance between successive iterations scaled by the variance of $\hat{h}_j$. We could keep increasing J until ε_J goes below a pre-determined threshold, say 0.01.

$J \approx \ln n$, but this can be substantially reduced through good initializations). However, semi-gradient methods require differentiable classes of functions (e.g., Random Forests are not allowed) and it appears difficult to characterize their theoretical properties beyond the case of linear basis functions.

A note on implementing (3.10): Since $z(a,x)$ is known, we recommend running a non-parametric regression of only $\hat{h}_j(a',x')$ on (a,x) at each step. We can then multiply the resulting non-parametric estimator by β and add back $z(a',x')$ to obtain the next estimate $\hat{h}_{j+1}(\cdot)$. A similar comment applies to (3.11). Algorithm 1 describes the estimation steps.

Algorithm 1 AVI using Random Forest

Require: Non-parametric estimate $\hat{\eta}$; initial values $\hat{h}_1, \hat{g}_1$; J (# iterations)

1: **for** $j = 1, 2, \ldots, J-1$: **do**

2: Predict $\beta\hat{h}_j(a',x')$ and $\beta e(a',x',\hat{\eta}) + \beta\hat{g}_j(a',x')$ using (a,x) with Random Forest, obtain prediction functions $\tilde{h}_{j+1}(\cdot)$ and $\hat{g}_{j+1}(\cdot)$

3: $\hat{h}_{j+1}(\cdot) \leftarrow \tilde{h}_{j+1}(\cdot) + z(\cdot)$

4: **end for**

5: Return $\hat{h}(a,x) = \tilde{h}_J(a,x) + z(a,x)$ and $\hat{g}(a,x) = \hat{g}_J(a,x)$

Notes: We recommend estimating η with a logit model using a 2nd or 3rd order polynomial in x, and setting $\hat{h}_1 = (1-\beta)^{-1}\mathbb{E}_n[z(a,x)]$, $\hat{g}_1 = \beta(1-\beta)^{-1}\mathbb{E}_n[e(a',x';\hat{\eta})]$ and $J = 20$ as defaults (or else, employ the procedure set out in Footnote 9 for J). The Random Forest tuning parameters *ntree* and *mtry* can be kept at default values, but we suggest checking whether the results change meaningfully if *mtry* varies by ± 1 (if they do, a cross-validation function can be used to determine *mtry*, e.g., $rfcv$ in R).

3.3. **Tuning parameters.** Both the semi-gradient and AVI methods require choosing tuning parameters. For AVI this is straightforward: as each iteration is a non-parametric estimation problem, the tuning parameters can be chosen in the usual manner, e.g., through cross-validation. In the case of linear semi-gradient methods, the tuning parameters are the dimensions $k_\phi = \dim(\phi)$ and $k_r = \dim(r)$ of the basis functions. In analogy with AVI, we propose selecting both through a procedure akin to cross-validation. The value of ω is estimated using a training sample and its performance evaluated on a hold-out or test sample, where the performance is measured in terms of the empirical mean-squared TD error $\mathbb{E}_{n,\text{test}}[\delta_z^2(a,x,a',x';\hat{h})]$ on the test dataset. The values of k_ϕ, k_r are chosen to minimize the mean squared TD error (see Section 6.2.1 for an example).

3.4. **Unobserved heterogeneity.** In Online Appendix C, we incorporate permanent unobserved heterogeneity by pairing our TD methods with the sequential Expectation-Maximization (EM) algorithm (Arcidiacono and Jones, 2003). This algorithm can handle

discrete heterogeneity in both individual utilities and transition densities. Monte Carlo evidence suggests that the algorithm works well in practice (see Online Appendix E.2).

4. Theoretical Properties of TD estimators

4.1. **Estimation of non-parametric terms.** We characterize rates of convergence for estimation of $h(\cdot)$ and $g(\cdot)$ under both semi-gradients and AVI.

4.1.1. *Linear semi-gradients.* We impose the following assumptions for estimation of $h(\cdot)$.

Assumption 1. *(i) The basis vector $\phi(a,x)$ is linearly independent (i.e., $\phi(a,x)^\intercal\omega = 0$ for all (a,x) if and only if $\omega = 0$). Additionally, the eigenvalues of $\mathbb{E}[\phi(a,x)\,\phi(a,x)^\intercal]$ are uniformly bounded away from zero for all $k_\phi := dim(\phi)$.*

(ii) $|\phi(a,x)|_\infty \leq M$ for some $M < \infty$.

(iii) There exists $C < \infty$ and $\alpha > 0$ such that $\|h - P_\phi[h]\|_2 \leq Ck_\phi^{-\alpha}$.

(iv) The domain of (a,x) is a compact set, and $|z(a,x)|_\infty \leq L$ for some $L < \infty$.

(v) $k_\phi \to \infty$ and $k_\phi^2/n \to 0$ as $n \to \infty$.

Assumption 1(i) rules out multi-collinearity in the basis functions. This is easily satisfied. Assumption 1(ii) ensures that the basis functions are bounded. This is again a mild requirement and is easily satisfied if either the domain of (a,x) is compact, or the basis functions are chosen appropriately (e.g., a Fourier basis). Assumption 1(iii) is a standard condition on the rate of approximation of $h(a,x)$ using a basis approximation. The value of α is related to the smoothness of $h(\cdot)$. Newey (1997) shows that for splines and power series, $\alpha = r/d$, where r is the number of continuous derivatives of $h(a,\cdot)$ and d is the dimension of x. Similar results can also be derived for other approximating functions such as Fourier series, wavelets and Bernstein polynomials. The smoothness properties of $h(a,\cdot)$ are discussed in Online Appendix B.3.2, where we provide primitive conditions on $z(a,x), K(x'|a,x)$ that ensure existence of r continuous derivatives of $h(a,\cdot)$ for each $a \in \mathcal{A}$. Assumption 1(iv) requires $z(a,x)$ to be bounded. Finally, Assumption 1(v) specifies the rate at which the dimension of the basis functions is allowed to grow. The rate requirements are mild, and are the same as those employed for standard series estimation. For the theoretical properties, the exact rate of k_ϕ is not relevant up to a first order since we propose estimators of θ^* that are locally robust to estimation of $h(\cdot)$.

We then have the following theorem on the estimation of $h(a,x)$:

Theorem 1. *Under Assumption 1, the following holds:*

(i) Both ω^* and $\hat{\omega}$ exist, the latter with probability approaching one.

(ii) $\|h(a,x) - \phi(a,x)^\intercal \omega^*\|_2 \leq (1-\beta)^{-1} \|h - P_\phi[h]\|_2 \leq C(1-\beta)^{-1} k_\phi^{-\alpha}$.

(iii) The L^2 error for the difference between $h(a,x)$ and $\phi(a,x)^\intercal \hat{\omega}$ is bounded as

$$\|h(a,x) - \phi(a,x)^\intercal \hat{\omega}\|_2 = O_p\left((1-\beta)^{-1}\left\{\frac{k_\phi}{\sqrt{n}} + k_\phi^{-\alpha}\right\}\right).$$

We prove Theorem 1 in Appendix A.1 by adapting the results of Tsitsiklis and Van Roy (1997). Part (i) ensures that the population and empirical TD fixed points exist. Parts (ii) and (iii) imply that the approximation bias and MSE of linear semi-gradients are analogous to those of standard series estimation apart from a $(1-\beta)^{-1}$ factor.

For the estimation of $\hat{\xi}$ we make use of cross-fitting as a technical device to obtain easy-to-verify assumptions on the estimation of η. This entails the following: we randomly partition the data into two folds. We estimate $\hat{\xi}$ separately for each fold using $\hat{\eta}$ estimated from the opposite fold. The final estimate of ξ^* is the weighted average of $\hat{\xi}$ from both the folds. For specific estimation methods, e.g., series estimation, it is possible to derive our theoretical results without cross-fitting; the latter may then be unnecessary in practice.

We impose the following assumptions for the estimation of $g(a,x)$.

Assumption 2. *(i) The basis vector $r(a,x)$ is linearly independent, and the eigenvalues of $\mathbb{E}[r(a,x)r(a,x)^\intercal]$ are uniformly bounded away from zero for all $k_r := dim(r)$.*

(ii) $|r(a,x)|_\infty \leq M$ *for some* $M < \infty$.

(iii) There exists $C < \infty$ *and* $\alpha > 0$ *such that* $\|g - P_r[g]\|_2 \leq C k_r^{-\alpha}$.

(iv) The domain of (a,x) is a compact set, and $|e(a,x)|_\infty \leq L < \infty$.

(v) $k_r \to \infty$ *and* $k_r^2/n \to 0$ *as* $n \to \infty$.

(vi) $\hat{\xi}$ is estimated from a cross-fitting procedure described above. The conditional choice probability function satisfies $\eta(a,x) > \delta > 0$, *where δ is independent of a, x. Additionally,* $\|\eta - \hat{\eta}\|_\infty = o_p(1)$ *and* $\|\eta - \hat{\eta}\|_2^2 = o_p(n^{-1/2})$.

Assumption 2 is a direct analogue of Assumption 1, except for the last part which provides regularity conditions when $\eta(\cdot)$ is estimated. These conditions are typical for locally robust estimates and only require the non-parametric function $\eta(a,x)$ to be estimable at faster than $n^{-1/4}$ rates. This is easily verified for most non-parametric estimation methods such as kernel or series regression. Under these assumptions, we have the following analogue of Theorem 1, which we prove in Appendix A.2 .

Theorem 2. *Under Assumption 2, the following holds:*

(i) Both ξ^* and $\hat{\xi}$ exist, the latter with probability approaching one.

(ii) $\|g(a,x) - r(a,x)^\intercal \xi^*\|_2 \leq (1-\beta)^{-1} \|g - P_r[g]\|_2 \leq C(1-\beta)^{-1} k_r^{-\alpha}$.

(iii) The L^2 error for the difference between $g(a,x)$ and $r(a,x)^\intercal\hat{\xi}$ is bounded as

$$\left\|g(a,x) - r(a,x)^\intercal \hat{\xi}\right\|_2 = O_p\left((1-\beta)^{-1}\left\{\frac{k_r}{\sqrt{n}} + k_r^{-\alpha}\right\}\right).$$

4.1.2. *Approximate Value Iteration.* We can expand the estimation error $\left\|h - \hat{h}_J\right\|_2$ in terms of the non-parametric estimation errors $\left\|\Gamma_z[\hat{h}_{j-1}] - h_j\right\|_2$ for $j = 1, \ldots, J$. In particular, since $\Gamma_z[h] = h$ and $\Gamma_z[\cdot]$ is a β-contraction, we have

$$\begin{aligned}\left\|h - \hat{h}_j\right\|_2 &\leq \left\|\Gamma_z[h] - \Gamma_z[\hat{h}_{j-1}]\right\|_2 + \left\|\Gamma_z[\hat{h}_{j-1}] - \hat{h}_j\right\|_2 \\ &\leq \beta\left\|h - \hat{h}_{j-1}\right\|_2 + \left\|\Gamma_z[\hat{h}_{j-1}] - \hat{h}_j\right\|_2.\end{aligned}$$

Iterating the above gives

$$\left\|h - \hat{h}_J\right\|_2 \leq \beta^{J-1}\left\|h - \hat{h}_1\right\|_2 + \sum_{j=2}^{J} \beta^{J-j}\left\|\Gamma_z[\hat{h}_{j-1}] - \hat{h}_j\right\|_2. \tag{4.1}$$

Equation (4.1) is a special case of error propagation (Munos and Szepesvári, 2008).

Recall that $\hat{h}_1$ is an arbitrary initialization. It is thus straightforward to provide conditions under which $\left\|h - \hat{h}_1\right\|_2$ is bounded by some constant M_1. As for the second term in (4.1), recall from the discussion in Section 3.2 that the minimization problem (3.10) corresponds to non-parametric estimation of $\Gamma_z[\hat{h}_{j-1}]$. Most machine learning methods come with guarantees on the non-parametric estimation rate $\left\|\Gamma_z[\hat{h}_{j-1}] - \hat{h}_j\right\|_2$.

We now describe our assumptions for AVI. Let $\mathcal{X}$ denote the d-dimensional space of x, and define $\mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$ as the Hölder ball with smoothness parameter γ:

$$\mathcal{W}_M^{\gamma,\infty}(\mathcal{X}) := \left\{f : \max_{0<|p|\leq\gamma} \sup_{x\in\mathcal{X}} |D^p f| < M\right\}.$$

Assumption 3. *There exist $M_0, M < \infty$ such that:*

(i) The domain, $\mathcal{X}$, of x is compact, $|h|_\infty \leq M_0$ and $h(a,\cdot) \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$ for each a.

(ii) $\left|\hat{h}_1\right|_\infty \leq M_0$ *and* $\left\|h - \hat{h}_1\right\|_2 \leq M_1$ *for some* $M_1 < \infty$.

(iii) $|\Gamma_z[f]|_\infty \leq M_0$ *and* $\Gamma_z[f](a,\cdot) \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$ *for all* $a \in \mathcal{A}$ *and* $\{f : |f|_\infty \leq M_0\}$.

(iv) The candidate class of functions $\mathcal{F}$ is such that $|f|_\infty \leq M_0$ for all $f \in \mathcal{F}$. Additionally, consider the non-parametric estimation problem (with i.i.d. observations $i = 1, \ldots, n$ and T fixed): $\hat{f} = \arg\min_{\tilde{f}\in\mathcal{F}} \sum_{i=1}^n \sum_{t=1}^{T-1} (y_{it} - \tilde{f}(a_{it}, x_{it}))^2$, where y_{it} is compactly supported and $\mathbb{E}[y_{it}|a_{it}, x_{it}] = f(a_{it}, x_{it})$ for some $f \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$. Then, uniformly over all

$f \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$, $\mathbb{E}\left[\left\|f - \hat{f}\right\|_2\right] \leq Cn^{-c}$ *for constants* $C < \infty$, $c > 0$ *independent of* n, *but* C *may depend on* M, M_0, γ *and* c *on* γ.

Assumption 3(i) is not needed to obtain a convergence rate for the AVI estimator, but we state it here as it is useful for subsequent results. The assumption of γ-Hölder continuity is taken from Farrell et al. (2021). Assumption 3(ii) is a mild condition on the initialization $\hat{h}_1$. Assumption 3(iii), which is novel to this paper, is a crucial smoothness condition requiring the operator $\Gamma_z[\cdot](a,\cdot)$ to map all bounded f onto $\mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$. In Online Appendix B.3.2, we show that both requirements in Assumption 3(iii) are satisfied if $z(a,\cdot)$ and $K(x'|a,\cdot)$ are γ-Hölder continuous.

Assumption 3(iv) is a high-level condition on the machine learning (ML) method $\mathcal{F}$. The requirement of bounded f implies that the ML method cannot diverge in the l_∞ sense, see Farrell et al. (2021) for a discussion of this in the context of multi-layer perceptrons (MLPs). The second part of Assumption 3(iv) implies that the ML method is able to non-parametrically approximate all functions in $\mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$ at the rate of at least n^{-c}. Most ML methods are proven to satisfy this. Consider, for instance, the class $\mathcal{F}$ of MLPs of width W and depth L; MLPs and, more generally, Neural Networks are widely used in RL. Farrell et al. (2021) show that for $W \asymp n^{\frac{d}{2(\gamma+d)}} \ln^2 n$ and $L \asymp \ln n$,

$$\sup_{f \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})} \mathbb{E}\left[\left\|f - \hat{f}\right\|_2\right] \leq C\left\{n^{-\frac{\gamma}{2(\gamma+d)}} \ln^4 n + \sqrt{\frac{\ln \ln n}{n}}\right\}.$$

Thus, Assumption 3(iv) is satisfied for MLPs. See Biau (2012) for related results on Random Forests. Note that Assumption 3(iv) is also the only way in which the dimension of x enters our estimation. Through a suitable choice of the ML method, e.g., Random Forests or LASSO, we can allow $\dim(x)$ to be proportional to, or even bigger than n.

Assumptions 3(iii) and 3(iv) imply that one can estimate $\Gamma[f]$ for any $|f|_\infty \leq M_0$ at the n^{-c} rate, i.e., $\sup_j \mathbb{E}\left[\left\|\Gamma_z[\hat{h}_{j-1}] - \hat{h}_j\right\|_2\right] \leq Cn^{-c}$. Combined with (4.1), this proves:

Theorem 3. *Suppose Assumptions 3(ii) to 3(iv) hold. Then, for all* n *large enough,*

$$\mathbb{E}\left[\left\|h - \hat{h}_J\right\|_2\right] \leq \frac{C(1-\beta^{J-1})}{1-\beta} n^{-c} + M_1 \beta^{J-1}.$$

See Online Appendix B.3.3 for a formal proof of Theorem 3. The first term in the expression for $\mathbb{E}\left[\left\|h - \hat{h}_J\right\|_2\right]$ from Theorem 3 is the statistical rate of estimation of h. The second term is the numerical error, which is seen to decline exponentially with the number of iterations J. Setting $J \asymp \ln n$ will ensure the numerical error is smaller than

the statistical rate of convergence. The number of iterations can be further reduced using a good initialization, $\hat{h}_1$, that makes M_1 small. For instance, initializing using the linear semi-gradient estimator, which is fast to compute, ensures $M_1 = o_p(1)$. Incidentally, Theorem 3 justifies the use of Neural Networks for batch RL; to the best of our knowledge this appears to be new even in the RL literature.

Turning to estimation of $\hat{g}$, we again assume cross-fitting is employed as in Theorem 2, i.e., $\hat{\eta}$ is computed from one half of the data, and $\hat{g}$ is computed using AVI on the other half, taking $\hat{\eta}$ as given. Define $\Gamma_{e,\tilde{\eta}}[f](a,x) := \beta\mathbb{E}[e(a',x';\tilde{\eta}) + f(a',x')|a,x]$, where $\tilde{\eta}$ is any candidate function for η.

Assumption 4. *(i) Let $\tilde{\eta}(a,x) \in [0,1]$ be any function such that $\inf_{a,x} \tilde{\eta}(a,x) > \delta > 0$. Then, there exist $M_1, M, C < \infty$, that may depend on δ but are otherwise independent of $\tilde{\eta}(\cdot)$, such that Assumptions 3(i) - 3(iv) hold after replacing $(h, \hat{h}_1, \Gamma_z[\cdot])$ with $(g, \hat{g}_1, \Gamma_{e,\tilde{\eta}}[\cdot])$.*

(ii) $\hat{g}$ is estimated from a cross-fitting procedure. The true conditional choice probability function satisfies $\inf_{a,x} \eta(a,x) > \delta > 0$. Additionally, $\|\eta - \hat{\eta}\|_2^2 = o_p(n^{-1/2})$ and with probability approaching one, $\inf_{a,x} \hat{\eta}(a,x) > \delta > 0$.

Assumption 4(i) requires analogues of Assumption 3 to hold. In Online Appendix B.3.2, we show that analogues of Assumptions 3(i) and 3(iii) are satisfied as long as $K(x'|a,\cdot)$ is γ-Hölder continuous (the other assumptions simply place restrictions on the initial value and the ML method used). Assumption 4(ii) is similar to Assumption 2(vi).

Theorem 4. *Suppose Assumption 4 holds. Then, with probability approaching one,*

$$\|g - \hat{g}_J\|_2 \leq \frac{C(1-\beta^{J-1})}{1-\beta} n^{-c} + M_1 \beta^{J-1} + o(n^{-1/2}).$$

See Online Appendix B.3.4 for a formal proof of Theorem 4.

4.2. **Estimation of structural parameters.** Estimation of $h(a,x)$ and $g(a,x)$ is inherently non-parametric because these functions depend on two non-parametric terms: the choice probabilities $\eta(a,x)$, and the transition densities $K(x'|a,x)$. The TD estimators implicitly take both into account. Under the PMLE criterion, the estimates for $K(x'|a,x)$ and θ^* are not orthogonal to each other and this extends to the lack of orthogonality between the estimates $\hat{h}$, $\hat{g}$ and θ^*.[10] We allow θ^* to be vector-valued for the remainder of this section.

[10]For discrete states, this lack of orthogonality implies an additional variance term for the structural parameter estimates, though the rates of convergence are still parametric. With continuous states,

We can recover $\sqrt{n}$-consistent estimation by adjusting the PMLE criterion to account for the first-stage estimation of h and g. Denote $(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}) := (a, x, a', x')$ and $m(a, x; \theta, h, g) := \partial_\theta \ln \pi(a, x; \theta, h, g)$, where

$$\pi(a, x; \theta, h, g) := \frac{\exp\{h(a, x)^\intercal \theta + g(a, x)\}}{\sum_{\breve{a}} \exp\{h(\breve{a}, x)^\intercal \theta + g(\breve{a}, x)\}}.$$

The PMLE estimator with plug-in estimates solves $\mathbb{E}_n[m(a, x; \theta, \hat{h}, \hat{g})] = 0$, but this is not robust to estimation of h, g. Let $V(a, x; \theta, h, g) := h(a, x)^\intercal \theta + g(a, x)$ denote the continuation value given (a, x). Also, define $\lambda(a, x; \theta)$ as the fixed point of the 'backward' dynamic programming operator

$$\Gamma^\dagger[f](a, x) := \psi(a, x; \theta, h, g) + \beta \mathbb{E}\left[f(a^{-\prime}, x^{-\prime}) | a, x\right], \tag{4.2}$$

where $(a^{-\prime}, x^{-\prime})$ denotes the past actions and states preceding (a, x), and

$$\psi(a, x; \theta, h, g) := -m(a, x; \theta, h, g) = -\sum_{\breve{a} \in \mathcal{A}} \{\mathbb{I}(a = \breve{a}) - \pi(\breve{a}, x; \theta, h, g)\} h(\breve{a}, x). \tag{4.3}$$

In Online Appendix B.4, we show that the locally robust moment corresponding to $m(a, x; \theta, h, g)$ is given by

$$\begin{aligned}\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, h, g, \eta, \lambda, \theta^*) := m(a, x; \theta, h, g) + \lambda(a, x; \theta^*) \{z(a, x)^\intercal \theta^* + \beta e(a', x'; \eta) \\ + \beta V(a', x'; \theta^*, h, g) - V(a, x; \theta^*, h, g)\}.\end{aligned} \tag{4.4}$$

Crucially, the correction term is not required to be a function of θ (though if we replaced θ^* in (4.4) with θ, that would be a valid correction term too).

The construction of the locally robust moment (4.4) is new. But it is infeasible since $\theta^*, \lambda(\cdot), h(\cdot), g(\cdot)$ and $\eta(\cdot)$ are unknown. However, we can replace these quantities with consistent estimates. We have already described how to estimate $\eta(\cdot), h(\cdot), g(\cdot)$. Recall that $\tilde{\theta}$ denotes the plug-in estimator of θ^* using (3.9); note that $\tilde{\theta}$ consistently estimates θ^* but is not efficient. An estimator, $\hat{\lambda}(\cdot)$, of $\lambda(\cdot)$ can then be obtained by applying either of our TD estimation methods on (4.2), with $\tilde{\theta}, \hat{h}, \hat{g}, \hat{\eta}$ replacing θ^*, h, g, η. For instance, using AVI, we could obtain iterative approximations $\{\hat{\lambda}^{(j)}, j = 1, \ldots, J\}$ for $\lambda(\cdot)$ using

$$\hat{\lambda}_{j+1} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_n \left[\left\|\psi(a, x; \tilde{\theta}, \hat{h}, \hat{g}) + \beta \hat{\lambda}_j(a^{-\prime}, x^{-\prime}) - f(a, x)\right\|^2\right]. \tag{4.5}$$

however, the PMLE estimator with plug-in values of $\hat{h}$ and $\hat{g}$ will converge at slower than parametric rates.

Plugging in $\hat{\lambda}(\cdot), \hat{h}, \hat{g}, \hat{\eta}$ into (4.4), we obtain the feasible locally robust moment

$$\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}, \hat{g}, \hat{\eta}, \hat{\lambda}, \tilde{\theta}) := m(a, x; \theta, \hat{h}, \hat{g}) + \hat{\lambda}(a, x; \tilde{\theta}) \Big\{ z(a, x)^{\intercal} \tilde{\theta} + \beta e(a', x'; \hat{\eta}) + \beta V(a', x'; \tilde{\theta}, \hat{h}, \hat{g}) - V(a, x; \tilde{\theta}, \hat{h}, \hat{g}) \Big\}. \tag{4.6}$$

Using the above, we can obtain a locally robust estimator, $\hat{\theta}$, as the solution to $\mathbb{E}_n[\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}, \hat{g}, \hat{\eta}, \hat{\lambda}, \tilde{\theta})] = 0$. We recommend obtaining this using cross-fitting, see Section 4.2.1 for details. Compared to the plug-in estimate (3.9), our locally robust estimator requires computation of $\lambda(\cdot)$, but when linear semi-gradients are used to estimate h, g, we can even derive a closed-form expression for $\lambda(\cdot)$, see Online Appendix B.5. Solving $\mathbb{E}_n[\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}, \hat{g}, \hat{\eta}, \hat{\lambda}, \tilde{\theta})] = 0$ is also computationally easy; the correction term is a constant, and $\nabla_\theta \zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}, \hat{g}, \hat{\eta}, \hat{\lambda}, \tilde{\theta}) = \nabla_\theta m(a, x; \theta, \hat{h}, \hat{g})$ is negative definite (as the PMLE criterion is concave), so solving this is no harder than solving the original moment condition without a correction term.

4.2.1. *$\sqrt{n}$-consistent estimation.* We focus on the general construction of the locally robust estimator, $\hat{\theta}$, using (4.6). As mentioned in the previous sub-section, we advocate cross-fitting to obtain this estimator. Algorithm 2 describes the estimation steps.

Algorithm 2 Structural parameter estimation

Require: Non-parametric estimate $\hat{\eta}$; initial values $\hat{h}_1, \hat{g}_1$; J (# iterations)

1: Split the data into two equal folds $\mathcal{N}_1, \mathcal{N}_2$
2: **for** each $\mathcal{N}_k, k = \{1, 2\}$: **do**
3: Run Algorithm 1 to obtain $\hat{h}^{(k)}, \hat{g}^{(k)}$
4: Obtain preliminary estimates $\tilde{\theta}^{(k)} := \arg\max_{\theta^{(k)}} \hat{Q}(\theta^{(k)})$ as in (3.9)
5: Run Algorithm 5 (Online Appendix D) with $\hat{h}^{(k)}, \hat{g}^{(k)}$ and $\tilde{\theta}^{(k)}$ as inputs to obtain $\hat{\lambda}^{(k)}$
6: Using plug-in quantities $\tilde{\theta}^{(-k)}, \hat{h}^{(-k)}, \hat{g}^{(-k)}, \hat{\eta}^{(-k)}, \hat{\lambda}^{(-k)}$ from the other fold $\mathcal{N}_{-k}$, obtain $\hat{\theta}^{(k)}$ by solving $\mathbb{E}_n^{(k)}[\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}^{(-k)}, \hat{g}^{(-k)}, \hat{\eta}^{(-k)}, \hat{\lambda}^{(-k)}, \tilde{\theta}^{(-k)})] = 0$, as in (4.6), where $\mathbb{E}_n^{(k)}[\cdot]$ denotes the empirical expectation using only observations from $\mathcal{N}_k$
7: **end for**
8: Obtain the final estimate $\hat{\theta} = (\hat{\theta}^{(1)} + \hat{\theta}^{(2)})/2$

Notes: We recommend estimating η with a logit model using a 2nd or 3rd order polynomial in x, and setting $\hat{h}_1 = (1-\beta)^{-1}\mathbb{E}_n[z(a, x)]$, $\hat{g}_1 = \beta(1-\beta)^{-1}\mathbb{E}_n[e(a', x'; \hat{\eta})]$ and $J = 20$ as defaults (or else, employ the procedure set out in Footnote 9 for J). The Random Forest tuning parameters *ntree* and *mtry* can be kept at default values, but we suggest checking whether the results change meaningfully if *mtry* varies by ± 1 (if they do, a cross-validation function can be used to determine *mtry*, e.g., $rfcv$ in R).

Following the analysis of Chernozhukov et al. (2022), it can be shown that this estimator has the same limiting distribution as the one based on (4.4). In particular, it achieves parametric rates of convergence. We state the regularity conditions below:

Assumption 5. *(i)* $\theta^* \in \Theta$*, a compact set, and* $\mathbb{E}\left[m\left(a, x; \theta, h, g\right)\right] = 0 \iff \theta = \theta^*$.

(ii) There exists a neighborhood, $\mathcal{N}$*, of* θ^* *such that uniformly over* $\theta \in \mathcal{N}$ *and for* $\| \tilde{h} - h \|$, $\| \tilde{g} - g \|$ *sufficiently small,* $\left\|\nabla_\theta m(a, x; \theta, \tilde{h}, \tilde{g}) - \nabla_\theta m(a, x; \theta^*, \tilde{h}, \tilde{g})\right\| \leq d(a, x) \|\theta - \theta^*\|$*, where* $\mathbb{E}[d(a, x)] < \infty$*. Furthermore,* $G := \mathbb{E}\left[\nabla_\theta m(a, x; \theta^*, h, g)\right]$ *is invertible.*

(iii) $\left\|\hat{h} - h\right\|_2 = o_p(n^{-1/4})$, $\|\hat{g} - g\|_2 = o_p(n^{-1/4})$ *and* $\|\hat{\eta} - \eta\|_2 = o_p(n^{-1/4})$*. Furthermore,* h, g *are continuous,* $\left\|\hat{h}\right\|_\infty, \|\hat{g}\|_\infty \leq M < \infty$ *and there exists* $\delta > 0$ *such that* $\inf_{a,x} \eta(a, x) > \delta$ *and* $\inf_{a,x} \hat{\eta}(a, x) > \delta$ *with probability approaching one.*

(iv) $\left\|\hat{\lambda}(\cdot, \cdot; \tilde{\theta}) - \lambda(\cdot, \cdot; \theta^*)\right\|_2 = o_p(n^{-1/4})$.

Assumption 5(i) implies θ^* is identified. When h, g are exactly known, $m(\cdot)$ is just the derivative of the pseudo-log-likelihood (2.1); the latter is always concave. Assumption 5(i) is satisfied if $\mathbb{E}\left[\nabla_\theta m\left(a, x; \theta, h, g\right)\right]$ is strictly positive-definite at θ^*. This is essentially equivalent to the requirement for identification in the discrete state space regime, which previous work e.g., Aguirregabiria and Mira (2002), has also assumed.[11] For instance, when the action space is binary ($a \in \{0, 1\}$), direct computation shows that a sufficient condition for Assumption 5(i) is:

$$\mathbb{E}\left[\eta(1, x)\eta(0, x)\left\{h(1, x) - h(0, x)\right\}\left\{h(1, x) - h(0, x)\right\}^\intercal\right] \succ 0,$$

where $\eta(a, x) := \pi(a, x; \theta^*, h, g)$ is the true conditional choice probability, and '$\succ 0$' indicates that the matrix in question is strictly positive-definite. In fact, under our assumptions (continuity of h, g and compactness of x), $\eta(a, x) > \delta > 0$ independently of a, x, so we can further rewrite the sufficient condition as $\Omega := \mathbb{E}\left[\left\{h(1, x) - h(0, x)\right\}\left\{h(1, x) - h(0, x)\right\}^\intercal\right] \succ 0$. This holds as long as $h(1, x) - h(0, x)$ is not linearly dependent at (almost surely) every x. For $\beta = 0$, it is equivalent to linear independence of $z(1, x) - z(0, x)$.[12]

Assumption 5(ii) is a mild regularity condition that is similar to Assumption 4 in Chernozhukov et al. (2022). The first part of Assumption 5(iii) follows from Theorems 1-4 under suitable conditions on the degree of smoothness of h, g. For instance, it is satisfied for AVI with Neural Networks if $\gamma \geq d$. The second part of Assumption 5(iii) is mild, and is satisfied as long as $\hat{h}, \hat{g}, \hat{\eta}$ are continuous (given that the support of x

[11] When the error distribution is unspecified, Buchholz et al. (2021) show that identification of θ^* is feasible only when β is sufficiently smaller than 1. Their findings do not directly apply to our setting as we assume the errors follow a Type I Extreme Value distribution; however, we do also require $\beta < 1$.

[12] More generally, under the ergodic distribution, $\Omega = (1 - \beta^2)^{-1}\{\Psi_0 + \sum_{j=0}^{\infty} \beta^{2j}(\Psi_j + \Psi_j^\intercal)\}$, where $\Psi_j := \mathbb{E}\left[\{z(1, x_t) - z(0, x_t)\}\{z(1, x_{t+j}) - z(0, x_{t+j})\}^\intercal\right]$. One can then posit various conditions on β and $\{\Psi_j\}_j$ such that $\Omega \succ 0$. For instance, if $\Psi_j + \Psi_j^\intercal$ is positive semi-definite and $\Psi_0 \succ 0$, we have $\Omega \succ 0$ for any $\beta < 1$. However, $\beta = 1$ is never possible. We leave open the discussion of alternative conditions.

is compact). In fact, we also directly impose this restriction in the context of the AVI estimator. Importantly Assumption 5(iii) only requires L_2-convergence of $\hat{h}, \hat{g}, \hat{\eta}$, and not uniform convergence. This is due to the use of a locally robust moment together with cross-fitting; see Chernozhukov et al. (2022) for a discussion of how they enable very mild assumptions on the convergence rates of ML estimators.[13]

Assumption 5(iv) requires $\lambda(\cdot, \cdot; \theta^*)$ to be estimable at faster than $n^{-1/4}$ rates as well. If h, g are known, it is straightforward to derive $n^{-1/4}$ rates as in Theorems 1-4. For plug-in estimation, we would need additional assumptions. For instance, we could employ three-way sample splitting as in Chernozhukov et al. (2018) where the first third of the sample is used to compute $\hat{h}, \hat{g}, \hat{\eta}, \tilde{\theta}$, and these estimates are then plugged into the second third of the sample to estimate λ. Lemma 8 in Online Appendix B.6 then shows that Assumption 5(iv) holds under the previous assumptions and some mild conditions on the AVI estimator (4.5) as long as $\hat{h}(a, \cdot), \hat{g}(a, \cdot) \in \mathcal{W}_M^{\gamma,\infty}(\mathcal{X})$ for some $M < \infty$ at each $a \in \mathcal{A}$ and γ is sufficiently large. It is possible to verify Assumption 5(iv) without three-way sample splitting as well, but this requires much stronger regularity conditions.

We are now ready to state the main result of this section.

Theorem 5. *Suppose that either Assumptions 1, 2 & 5 (for linear semi-gradients) or 3-5 (for AVI) hold. Then the estimator, $\hat{\theta}$ of θ^*, based on (4.6) is $\sqrt{n}$-consistent, and satisfies*

$$\sqrt{n}(\hat{\theta} - \theta^*) \Longrightarrow N(0, V),$$

where $V = (G^\intercal \Omega^{-1} G)^{-1}$, with $\Omega := \mathbb{E}\left[\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta^*, h, g, \eta, \lambda, \theta^*)\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta^*, h, g, \eta, \lambda, \theta^*)^\intercal\right]$.

The proof of the above theorem follows by verifying the regularity conditions of Chernozhukov et al. (2022, Theorem 9), see Appendix A.3 for the details. For inference on $\hat{\theta}$, the covariance matrix V can be estimated as $\hat{V} = (\hat{V}_1 + \hat{V}_2)/2$, where $\hat{V}_1 = \left(\hat{G}_1^\intercal \hat{\Omega}_1^{-1} \hat{G}_1\right)^{-1}$ with (a similar expression holds for $\hat{V}_2$)

$$\hat{G}_1 = \mathbb{E}_n^{(1)}\left[\left.\frac{\partial \zeta\left(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{h}^{(2)}, \hat{g}^{(2)}, \hat{\eta}^{(2)}, \hat{\lambda}^{(2)}, \tilde{\theta}^{(2)}\right)}{\partial \theta^\intercal}\right|_{\theta=\hat{\theta}^{(2)}}\right], \quad \text{and}$$

$$\hat{\Omega}_1 = \mathbb{E}_n^{(1)}\left[\zeta\left(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \hat{\theta}^{(2)}, \hat{h}^{(2)}, \hat{g}^{(2)}, \hat{\eta}^{(2)}, \hat{\lambda}^{(2)}, \tilde{\theta}^{(2)}\right)\zeta\left(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \hat{\theta}^{(2)}, \hat{h}^{(2)}, \hat{g}^{(2)}, \hat{\eta}^{(2)}, \hat{\lambda}^{(2)}, \tilde{\theta}^{(2)}\right)^\intercal\right].$$

In Online Appendix B.10, we show that $\hat{V}$ is consistent for V under our stated assumptions.

[13]The intuitive reason for this is that cross-fitting ensures only the prediction properties of the non-parametric estimator are relevant.

4.2.2. *On the relative efficiency of TD estimation.* The TD estimator from (4.6) is robust to non-parametric estimation of transition densities. When the transition density has a parametric form, it is less efficient than the full-MLE estimator that jointly estimates the structural and transition density parameters. However full-MLE is seldom, if ever, used. Standard approaches such as NFXP and NPL are equivalent to partial-MLE, which employs a plug-in estimate of the transition density. For this reason, neither NFXP nor NLP are fully efficient: if we make the model for the transition density richer, while still keeping it parametric, the performance of NFXP and NLP will start to degrade and become worse than TD estimation. In the non-parametric regime, these methods lose $\sqrt{n}$-consistency. On the other hand, when the transition density is fully known, NFXP and NPL are equivalent to full-MLE, and therefore more efficient than TD estimation. Between these two extremes, whether or not TD estimation is more efficient than NFXP will depend on the statistical complexity of the model used for the transition density.

An interesting open question is whether our estimator attains the semi-parametric efficiency bound when the transition density is unknown. The GMM formulation of the problem in (B.9)-(B.10) suggests this may be the case, but we leave this as a conjecture.

5. Estimation of dynamic discrete games

Our setup for dynamic games is based on Aguirregabiria and Mira (2010). We assume a single Markov-Perfect-Equilibrium setup where multiple players $i = 1, 2, \ldots, n$ play against each other in M different markets. Each player chooses among A mutually exclusive actions to maximize an infinite horizon objective. We observe the state of play for T time periods, where both T and the number of players n are fixed, while $M \to \infty$. Utility of the players in any time period is affected by the actions of all the others, and a set of states x that are observed by all players. The per-period utility is denoted by $z_i(a_i, a_{-i}, x)^\intercal \theta^* + e_i$ for each player i, for some finite-dimensional parameter θ^*, where a_i denotes player i's action, a_{-i} denotes the actions of all other players and e_i is an idiosyncratic error term. As in Section 3, we take θ^* to be scalar to simplify the notation; all our results continue to hold for vector-valued θ^*, as long as each dimension is treated separately. Evolution of the states in the next period is determined by the transition density $K(x'|a, x)$ where $\mathbf{a} := (a_1, \ldots, a_n)$ denotes the actions of all the players. We denote by x_{tm} the state at market m in time period t, by $\mathbf{a}_{tm}$ the vector of actions by all players at time t in market m, and by a_{itm} the action of player i at time t in market m.

We also let $P_i(a_i|x_t)$ denote the choice probability of player i taking action a_i when the state is x_t , and define $e_i(a_i, x) := \gamma - \ln P_i(a_i|x)$.

As in the single-agent case, the parameter θ^* can be obtained as solutions to the pseudo-log-likelihood function:

$$Q(\theta) = \sum_{i=1}^{n} \sum_{m=1}^{M} \sum_{t=1}^{T-1} \ln \frac{\exp\{h_i(a_{itm}, x_{tm})\theta + g_i(a_{itm}, x_{tm})\}}{\sum_a \exp\{h_i(a, x_{tm})\theta + g_i(a, x_{tm})\}}, \tag{5.1}$$

where $h_i(.)$ and $g_i(.)$ are now player-specific, and given by

$$\begin{aligned} h_i(a_i, x) &= \mathbb{E}[z_i(a, x)|a_i, x] + \beta \mathbb{E}\left[h_i(a', x')|a_i, x\right], \\ g_i(a_i, x) &= \mathbb{E}\left[e_i(a', x') + \beta g_i(a', x')|a_i, x\right]. \end{aligned} \tag{5.2}$$

In contrast to (2.2), the expectation averages over the actions of the other players as well.

Previous literature estimates θ^* using a two-step procedure: In the first step, the conditional choice probabilities $P_i(a_i|x_t)$ are calculated non-parametrically. These, along with estimates of $K(.)$ are then used to recursively solve for $h_i(.)$ and $g_i(.)$ using equation (5.2). This step requires integrating over the actions of all the other players. Finally, given the estimated values of $h_i(.)$ and $g_i(.)$, the parameter θ^* is estimated through either pseudo-maximum-likelihood (PML; Aguirregabiria and Mira, 2007), minimum distance estimation (MDE; Pesendorfer and Schmidt-Dengler, 2008) or iterative versions of these (Bugni and Bunting, 2021). By contrast, our algorithm is a straightforward extension of those suggested in earlier sections for single-agent models. Let $\hat{\eta}_i(a_i, x)$ denote a non-parametric estimate of the choice probabilities for player i and denote $e(a_i, x; \hat{\eta}_i) = \gamma - \ln \hat{\eta}_i(a_i, x)$. We apply our TD methods on the recursion (5.2), separately for each player. The linear semi-gradient estimates are given by $\hat{h}_i(a_i, x) = \phi(a_i, x)^\intercal \hat{\omega}_i$ and $\hat{g}_i(a_i, x) = r(a_i, x)^\intercal \hat{\xi}_i$, where

$$\begin{aligned} \hat{\omega}_i &= \mathbb{E}_n\left[\phi(a_i, x)\left(\phi(a_i, x) - \beta\phi(a_i', x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\phi(a_i, x) z_i(a_i, a_{-i}, x)\right], \\ \hat{\xi}_i &= \mathbb{E}_n\left[r(a_i, x)\left(r(a_i, x) - \beta r(a_i', x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\beta r(a_i, x) e(a_i', x'; \hat{\eta}_i)\right], \end{aligned} \tag{5.3}$$

and for any function $f(\cdot)$, we define

$$\mathbb{E}_n[f(\mathbf{a}, x, \mathbf{a}', x')] := \frac{1}{M(T-1)} \sum_{m=1}^{M} \sum_{t=1}^{T-1} f(\mathbf{a}_{tm}, x_{tm}, \mathbf{a}_{t+1m}, x_{t+1m}). \tag{5.4}$$

Similarly, the AVI iterations for $h_i(\cdot), g_i(\cdot)$ are given by

$$\hat{h}_i^{(j+1)} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_n \left[\left\| z_i(a_i, a_{-i}, x) + \beta \hat{h}_i^{(j)}(a_i', x') - f(a_i, x) \right\|^2 \right],$$
$$\hat{g}_i^{(j+1)} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_n \left[\left\| \beta e(a_i', x'; \hat{\eta}_i) + \beta \hat{g}_i^{(j)}(a_i', x') - f(a_i, x) \right\|^2 \right]. \tag{5.5}$$

If the players are symmetric ($z_i(a_i, a_{-i}, x)$ does not depend on player i) we can obtain computationally faster and more precise estimates by pooling across players.

Importantly, neither of the estimation strategies (5.3) nor (5.5) require partialling out other players' actions, leading to a tremendous reduction of computation. The non-parametric estimates $\hat{h}_i, \hat{g}_i$ can be plugged into the PMLE criterion (5.1) to obtain an estimate for θ^* as

$$\tilde{\theta} = \arg\max_{\theta} \sum_i \hat{Q}_i(\theta), \text{ where}$$
$$\hat{Q}_i(\theta) := \sum_{m=1}^{M} \sum_{t=1}^{T-1} \ln \frac{\exp\left\{ \hat{h}_i(a_{itm}, x_{tm})\theta + \hat{g}_i(a_{itm}, x_{tm}) \right\}}{\sum_a \exp\left\{ \hat{h}_i(a, x_{tm})\theta + \hat{g}_i(a, x_{tm}) \right\}}. \tag{5.6}$$

It is straightforward to construct a locally robust estimator for θ^* in analogy with that for single-agent models. We describe this in Online Appendix B.7. The convergence properties of the locally robust estimators for games are also similar to those for single-agent models; a formal statement is provided in Online Appendix B.8.

The PMLE criterion (5.6) with plug-in estimates for $h_i(.)$ and $g_i(.)$ is not efficient even with discrete states, as discussed by Aguirregabiria and Mira (2007). However the values of $h_i(.)$ and $g_i(.)$ can be plugged into other, more efficient objectives, such as the MDE criterion with an efficient weighting matrix; Bugni and Bunting (2021) show that the latter is more efficient than even iterated PMLE estimation. With continuous states, however, one would need to employ locally robust corrections even for MDE to recover parametric rates of convergence for estimation of θ^*. The locally robust correction term can be constructed in a similar way as that for the PMLE criterion.

6. Simulations

In this section, we run two Monte Carlo simulations to test our methods, and compare them to alternative approaches. Our first simulation is based on the firm entry problem in Aguirregabiria and Magesan (2018). In the second set of Monte Carlo simulations, we

test our estimation method for dynamic discrete games. The latter simulations are based on the dynamic firm entry game used in Aguirregabiria and Mira (2007).[14]

Online Appendices E.2 and E.3 report additional simulations based on the famous Rust (1987) bus engine replacement problem. Using this model, we provide results for a case with permanent unobserved heterogeneity and also compare our methods to the estimator proposed by Chernozhukov et al. (2018) for DDC models with finite dependence.

6.1. **Firm entry problem.** Consider the following dynamic firm entry problem described in Aguirregabiria and Magesan (2018). A firm decides whether to enter ($a_t = 1$) or not enter ($a_t = 0$) in a market for $t = 1, ..., T$ time periods. The payoff when entering is given by $\Pi_t = VP_t - FC_t - EC_t + \varepsilon_t$, where VP_t, FC_t and EC_t denote the firm's variable profit, fixed cost and entry cost, and ε_t is a transitory shock that follows a logistic distribution. Variable profit is given by $VP_t = (\theta_0^{VP} + \theta_1^{VP} z_{1t} + \theta_2^{VP} z_{2t}) \exp(\omega_t)$, where ω_t denotes the firm's productivity shock, and z_{1t}, z_{2t} are exogenous state variables affecting the price-cost margin in the market. The fixed cost is given by $FC_t = \theta_0^{FC} + \theta_1^{FC} z_{3t}$, and the entry cost is given by $EC_t = (\theta_0^{EC} + \theta_1^{EC} z_{4t})(1 - a_{t-1})$, where z_{3t}, z_{4t} are further exogenous state variables, and a_{t-1} denotes the entry decision in period $t-1$ which is an endogenous state variable. The payoff of not entering is normalized to zero. The parameters $\theta^* \equiv \left\{\theta_0^{VP}, \theta_1^{VP}, \theta_2^{VP}, \theta_0^{FC}, \theta_1^{FC}, \theta_0^{EC}, \theta_1^{EC}\right\}$ are the structural parameters of interest. The exogenous state variables z_{jt}, $j \in \{1, 2, 3, 4\}$, and ω_t are continuous and follow AR(1) processes, where $z_{jt} = \gamma_0^j + \gamma_1^j z_{jt-1} + e_{jt}$, and $\omega_t = \gamma_0^\omega + \gamma_1^\omega \omega_{t-1} + e_{\omega t}$. The error terms $e_{jt}, e_{\omega t}$ follow normal $N(0, 1)$ distributions. The discount factor β is 0.95.

To carry out the simulations, we choose values for the structural parameters θ^* ($\theta_0^{VP} = 0.5$, $\theta_1^{VP} = 1.0$, $\theta_2^{VP} = -1.0$, $\theta_0^{FC} = 1.5$, $\theta_1^{FC} = 1.0$, $\theta_0^{EC} = 1.0$, $\theta_1^{EC} = 1.0$) and for the autoregressive processes of z_{jt} and ω_t ($\gamma_0^j = 0.0$, $\gamma_1^j = 0.6$, $\gamma_0^\omega = 0.2$, $\gamma_1^\omega = 0.6$), and discretize the exogenous state variables to obtain a transition matrix with a 6-point support following Tauchen (1986). The resulting dimension of the state space is $2 \times 6^5 = 15,552$. The discretization of the support is for simulations only; our methods treat these variables as continuous and do not require any prior knowledge of how they evolve (the knowledge of AR(1) dynamics is also not used). We iterate on the value function to obtain the vector of choice probabilities for each combination of the states, and use these to derive the ergodic, i.e., steady-state distribution of the state variables. Using this distribution, we generate data for 3000 firms, with $T = 2$ time periods.

[14] R code for all simulations is made available as part of the replication package.

6.1.1. *Simulation results - firm entry problem.* Table 1 shows the results based on 1000 simulations using the linear semi-gradient and AVI methods. For the linear semi-gradient method, we parameterize $h(a, x)$ and $g(a, x)$ using a first order polynomial in the state variables.[15] For AVI, we approximate $h(a, x)$ and $g(a, x)$ using a Random Forest, and iterate the AVI procedure 20 times for each round of the simulations. For both the linear semi-gradient and the AVI methods, we estimate the choice probabilities η that enter $e(a', x'; \eta)$ using a logit model where the explanatory variables are the state variables, their squares and interactions up to the second order.

We present results generated with and without the locally robust correction. For the results without correction, we obtain estimates for θ^* using (3.9). To generate the locally robust estimates, we use moment equation (B.12) for the linear semi-gradient method, and moment equation (4.6) for the AVI method where we employ a Random Forest to derive an estimate for the $\lambda(a, x, \tilde{\theta})$ term contained in the locally robust moment. As before, the AVI method for estimation of $\lambda(\cdot)$ is iterated 20 times. We also use the sample splitting method described in Section 4.2.1 for the locally robust estimators, and we obtain the final $\hat{\theta}$ as weighted average of the θ^* estimates from the two samples.

Both the linear semi-gradient and AVI estimates are closely centered around the true values, but the latter is clearly preferable in terms of mean squared error (MSE). While the locally robust estimator should in theory be preferable, we find that it produces results which are similar and if anything have slightly higher MSE than the non-robust versions. In fact, we find that there is very little bias to begin with, and the distribution of the estimates under the non-robust versions are already very close to normal, see Online Appendix E.1 for the plots of the finite sample distributions. The lower bias may be due to the specific nature of the example, which falls under a special class of DDC models called 'dynamic-logit models' (see Section 6.1.2). On the flip side, the locally robust methods are associated with higher variance due to cross-fitting. So, overall, there appears to be no gain from using the locally robust method in this example. Presumably, the variability of locally robust estimators can be lowered by using more folds in the cross-fitting procedure (we use two folds in all our examples); this would, however, come at the expense of slower computation times.

[15]For the ω's relating to parameters $\theta_0^{VP}, \theta_1^{VP}, \theta_2^{VP}, \theta_0^{FC}, \theta_1^{FC}$, and for ξ, the terms include a constant, the exogenous state variables, the player's binary choice a_t and the interactions of a_t with all terms in the exogenous states. Given the set-up of the model, we also include the interactions $z_{1t}\exp(\omega_t)$ and $z_{2t}\exp(\omega_t)$ as state variables. In addition to the terms included above, the ω's relating to parameters θ_0^{EC} and θ_1^{EC} also contain the terms $(1 - a_{t-1})$ and $(1 - a_{t-1})z_{4t}$, respectively. The total number of terms included is 16 (17 for θ_0^{EC} and θ_1^{EC}).

6.1.2. *Comparison with existing methods.* Table 1 compares our methods to the two-step Euler Equation (EE) approach of Aguirregabiria and Magesan (2018). As given in Aguirregabiria and Magesan (2018, eq. 30), the EE estimator is not universally applicable; it can only be employed on the restricted class of 'dynamic-logit models' where the only endogenous variable is the past action and all the other variables are exogenous.[16] Our estimation strategy, unlike EE, does not exploit this special feature of the model (which is satisfied by the simulation study but is otherwise restrictive). Nevertheless, the linear semi-gradient method without locally robust corrections is three times faster than EE, albeit at the expense of a somewhat higher MSE.

On the other hand, the MSE of AVI is slightly lower than that of EE, but it is also much slower. However, we think raw computation times do not paint the full picture here. The reason AVI is slower is because we employ Random Forest (RF). The computational time can be made an order of magnitude smaller using other ML techniques such as series estimation, Ridge, LASSO or MARS, but that is not necessarily a reason to choose these over RF. An analogy can be drawn here with prediction: despite being slower, RF is often used in moderate and high-dimensional prediction problems as its predictive performance is superior, and more importantly, it is less sensitive to how the state variables are transformed. By contrast, both linear semi-gradient and EE approaches require choosing a specification; we need to choose the family of basis functions for the former, and the level and type of discretization for the latter. In practice, one typically runs specification checks to ensure robustness, but this takes up significantly more computational time, and in truly high-dimensional scenarios (e.g., when $\dim(x) \propto n$), finding the right specification (e.g., the level of discretization) may not even be feasible. A major advantage of RF, then, is that it does not require a specification and is also remarkably robust to tuning parameter choices (Hastie et al., 2009, p.590). In many practical applications, we think this advantage trumps the additional computational time that it involves.

Our locally robust corrections also make computations more time consuming, but are needed to achieve $\sqrt{n}$-consistency under continuous states. The EE estimator is only $\sqrt{n}$-consistent if the states are discrete, but for continuous states, discretization bias would imply a loss of $\sqrt{n}$-consistency. A fair comparison of computational times would thus require comparing the locally robust estimator with a locally robust version of EE, but constructing the latter is beyond the scope of this paper.

[16]While we presume that an EE estimator can also be derived for more general models, the construction and computation of the EE mapping with endogenous state variables is much more involved.

TABLE 1. Simulations: Firm entry problem

		Linear semi-gradient				AVI				2-step EE	
		not locally robust		locally robust		not locally robust		locally robust			
	DGP	TDL	MSE	TDL	MSE	TDL	MSE	TDL	MSE	EE	MSE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
θ_0^{VP}	0.5	0.5028	0.0058	0.5087	0.0068	0.4877	0.0035	0.4844	0.0053	0.5052	0.0031
		(0.0760)		(0.0821)		(0.0582)		(0.0711)		(0.0555)	
θ_1^{VP}	1.0	0.9831	0.0050	1.0049	0.0058	1.0045	0.0034	1.0118	0.0051	1.0105	0.0034
		(0.0689)		(0.0762)		(0.0581)		(0.0704)		(0.0572)	
θ_2^{VP}	-1.0	-0.9839	0.0055	-1.0061	0.0065	-1.0059	0.0037	-1.0136	0.0053	-1.0119	0.0034
		(0.0725)		(0.0805)		(0.0602)		(0.0719)		(0.0574)	
θ_0^{FC}	1.5	1.5066	0.0220	1.5254	0.0257	1.5379	0.0166	1.5433	0.0232	1.5136	0.0150
		(0.1482)		(0.1583)		(0.1231)		(0.1460)		(0.1218)	
θ_1^{FC}	1.0	0.9746	0.0157	0.9916	0.0180	1.0090	0.0101	1.0041	0.0145	1.0044	0.0088
		(0.1228)		(0.1342)		(0.1001)		(0.1206)		(0.0939)	
θ_0^{EC}	1.0	0.9973	0.0101	1.0132	0.0117	0.9864	0.0103	0.9982	0.0145	1.0030	0.0104
		(0.1003)		(0.1076)		(0.1007)		(0.1203)		(0.1018)	
θ_1^{EC}	1.0	0.9948	0.0260	1.0163	0.0303	0.9645	0.0199	1.0082	0.0291	0.9081	0.0240
		(0.1613)		(0.1735)		(0.1365)		(0.1705)		(0.1248)	
Total MSE			0.0901		0.1050		0.0674		0.0970		0.0681
Time per round (in sec)		0.33		3.70		44.69		76.55		0.99	

Notes: The table reports results based on 1000 simulations with 3000 firms. Column (1) shows the true parameter values in the model. Columns (2), (4), (6), (8), (10) report the empirical mean and standard deviation (in parentheses) for the estimated parameters for each of the estimation methods. Columns (3), (5), (7), (9), (11) report the mean squared errors.

For a second comparison, we compare our estimators to a standard CCP estimator where the state variables are discretized and the transition and choice probabilities are estimated using cell values. We discretize the state space by creating dummy variables for each state variable $z_{1t}, z_{2t}, z_{3t}, z_{4t}$ and $\exp(\omega_t)$ based on whether they are above or below their median. However, even this results in empty cell values, so the state space needs to be restricted further. A common approach is to use K-means clustering, but this is not appropriate in the current setting where the state variables are independent by construction. We therefore restrict the state space grid by combining variables z_{1t} and z_{2t} into a binary variable taking value one whenever both individual dummies take value one. The resulting state space consists of four binary variables, implying 16 cells in the exogenous state space grid. We tried alternative feasible ways of discretizing the state space, but found that these do not lead to improvements over the chosen method. We run 1000 simulations, and the results are shown in Table 2.

Compared to the results from Table 1, the discretized CCP estimator leads to substantially larger bias in some of the estimated parameters. Column (4) shows that the corresponding MSEs are large and generally exceed those obtained using our estimators. This is particularly true for parameters $\theta_1^{VP}$and θ_2^{VP}. Overall, the toal MSE increases more than 10-fold from $0.067 - 0.105$ across all parameters in Table 1 to 1.109 in Table 2. At the same time, our linear semi-gradient method is even three times faster computationally than discretization; this may be related to matrix inversion being more ill-conditioned under discretization.

6.2. **Firm entry game.** Consider the following firm market entry game, which is similar to that described in Aguirregabiria and Mira (2007). There are $i = 1, ..., 5$ firms (players), and we observe their decision to enter ($a_{itm} = 1$) or not enter ($a_{itm} = 0$) in $m = 1, ..., M$ different markets for $t = 1, ..., T$ time periods. Denote a firm's action by $j \in \{1, 0\}$. The payoff of each firm i is affected by the decision of all the other firms whether to enter, as well as firm i's previous-period entry decision. Current-period profits when entering are given by

$$\Pi_{itm} = \theta_{RS} \ln(S_{tm}) - \theta_{RN} \ln(1 + \sum_{j \neq i} a_{jtm}) - \theta_{FC} - \theta_{EC}(1 - a_{i(t-1)m}) + \varepsilon_{itm},$$

where $\ln(S_{tm})$ is a measure of consumer market size of market m in period t, and ε_{itm} is a transitory shock that follows a logistic distribution. We assume that $\ln(S_{tm})$ is continuous and follows an AR(1) process, where the parameters are the same across

TABLE 2. Simulations: Firm entry problem - Comparison with standard CCP

	DGP (1)	TDL (2)	bias (3)	MSE (4)
CCP with discretized state variables				
θ_0^{VP}	0.5	0.1391	-0.3609	0.1815
		(0.2266)		
θ_1^{VP}	1.0	0.7968	-0.2032	0.3474
		(0.5535)		
θ_2^{VP}	-1.0	-0.4017	0.5983	0.4154
		(0.2396)		
θ_0^{FC}	1.5	1.3799	-0.1201	0.0313
		(0.1300)		
θ_1^{FC}	1.0	0.8655	-0.1345	0.0374
		(0.1392)		
θ_0^{EC}	1.0	0.7859	-0.2141	0.0538
		(0.0891)		
θ_1^{EC}	1.0	0.9011	-0.0989	0.0425
		(0.1809)		
Total MSE				1.1093
Time per round (in sec)		0.94		

Notes: The table reports results based on 1000 simulations with 3000 firms. Column (1) shows the true parameter values in the model. Column (2) reports the empirical mean and standard deviation (in parentheses) for the estimated parameters. Column (3) reports the average bias in the estimated parameters. The mean squared errors are reported in column (4).

markets: $\ln(S_{tm}) = \alpha + \lambda \ln(S_{(t-1)m}) + u_{tm}$. The error term u_{tm} is assumed to follow a normal $N(0,1)$ distribution. The profit of not entering is normalized to zero, and the discount factor β is 0.95. The parameters $\theta^* \equiv \{\theta_{RS}, \theta_{RN}, \theta_{FC}, \theta_{EC}\}$ are the structural parameters of interest. The state variables in this setting are given by the current market demand variable S_{tm}, as well as the vector of all firms' previous entry decisions $a_{(t-1)m} = \left\{a_{i(t-1)m} : i = 1, ..., 5\right\}$.

To carry out the simulations, we choose values for the structural parameters θ^* ($\theta_{RS} = 1, \theta_{RN} = 1, \theta_{FC} = 1.7, \theta_{EC} = 1$), and for the autoregressive process for log market size ($\alpha = 1.5$, $\lambda = 0.5$). We discretize $\ln(S_{tm})$ and obtain a transition matrix for the discretized variable with a 10-point support following the method by Tauchen (1986). As in the Monte Carlo experiments for the firm entry problem in Section 6.1, the discretization is for simulations of the data only and we treat the state variables as continuous in our

estimations. We then solve for the Markov-Perfect-Equilibrium of the game.[17] Using the equilibrium (i.e., ergodic) distribution, we generate data for 1000 and for 3000 markets, with $T = 2$ time periods.

6.2.1. *Simulation results - firm entry game.* We present the results of 1000 simulations based on the linear semi-gradient method, without employing the locally robust correction. Each round of the simulations begins by generating new data, where the first-period state variables are drawn from the steady-state distribution. In order to assess the sensitivity of our algorithm to different specifications for the basis functions, we parameterize $h(a, x)$ and $g(a, x)$ using different sets of polynomials in the state variables. In particular, we show results where $h(a, x)$ and $g(a, x)$ are approximated using a second, third or fourth order polynomial.[18] For all simulations, the choice probabilities η that enter $e(a', x'; \eta)$ are estimated using individual logit models for each firm, where we use a third order polynomial in the state variables as explanatory variables. We then estimate the parameters ω and ξ using equation (5.3).[19] Finally, we obtain estimates for the θ^* parameters as the solutions to the pseudo-log-likelihood function (5.1).

The results are shown in Table 3. Panels A, B and C present simulations for the same dataset using different basis functions to parameterize the value function terms $h(a, x)$ and $g(a, x)$. Column (2) shows that even with 1000 markets our algorithm produces parameter estimates that are closely centered around the true values. The results are generally similar across Panels A to C, although the bias and MSE tends to be lowest for the second order polynomial, and highest for the fourth order polynomial. This is especially the case for the parameter on the number of market entrants, θ_{RN}. To assess these differences formally for the case with 1000 markets, we use the cross-validation procedure described in Section 3.3. The procedure is applied to ten random samples of market size 1000, and we find that the TD error criterion consistently selects the second order polynomial as the optimal set of basis functions. Thus, the proposed cross-validation method provides useful guidance for choosing the number of basis functions.

[17]This is done by finding the firms' conditional value functions $\nu_j(S_{tm}, a_{(t-1)m})$ for each of the $2^5 \times 10 = 320$ possible combinations of the state variables through repeated iteration, and using these to derive the equilibrium choice probabilities $p(S_{tm}, a_{(t-1)m})$. Based on the equilibrium probabilities, we compute the equilibrium distribution of state variables.

[18]For the ω's relating to parameters $\theta_{RS}, \theta_{RN}, \theta_{FC}$ and for ξ, the terms include a constant, terms up to the second/third/fourth order in the state variables $\ln(S_{tm})$ and $\ln(1 + \sum_{j \neq i} a_{j(t-1)m})$, the firm's binary choice a_{itm} and the interactions of a_{itm} with all terms in the state variables. The total number of terms is 12/20/30. In addition to these terms, the ω's relating to parameter θ_{EC} also contain the term $(1 - a_{i(t-1)m})a_{itm}$.

[19]Given the symmetric set-up of the game, we pool the data across players in this application.

In a similar version of the firm entry game, Aguirregabiria and Mira (2007) use the NPL algorithm and derive results comparable to ours. Note, however, that for a direct comparison of our results with those obtained using the NPL algorithm, one would need to obtain a non-parametric estimate of the transition density when implementing the latter which is not trivial in practice.

As expected, columns (4) and (5) show that increasing the market size generally reduces the small sample bias in the estimated parameters, and leads to a fall in the empirical standard deviations. In addition to being smaller, the MSE across Panels A to C is also more similar across the three sets of basis functions. As before, we employ the cross-validation method described above to compare these specifications more formally for the case with 3000 markets. In line with the estimation results, we find that all three polynomials now produce very similar sets of mean squared TD errors, even though the second order polynomial continues to be the one that is selected by the criterion.[20] While we view this as further evidence that the proposed cross-validation method can provide useful guidance to choose a suitable set of basis functions, more importantly, the small differences in the results across panels A to C also suggest that our methods prove fairly robust to this choice in practice.

7. Conclusions

We propose two new estimators for DDC models which overcome previous computational and statistical limitations by combining traditional CCP estimation approaches with the idea of TD learning from the RL literature. The first approach, linear semi-gradient, makes use of simple matrix inversion techniques, is computationally very cheap and therefore fast. The second approach, Approximate Value Iteration, can be easily combined with any ML method devised for prediction. Unlike previous estimation methods, our methods are able to easily handle continuous and/or high-dimensional state spaces in settings where a finite dependence property does not hold. This is of particular importance for the estimation of dynamic discrete games. We also propose a locally robust estimator to account for the non-parametric estimation in the first stage. We prove the statistical properties of our estimator and show that it is consistent and converges at parametric rates. A range of Monte Carlo simulations using a dynamic firm entry problem, a dynamic firm entry game and two versions of the famous Rust (1987) engine replacement problem show that the proposed algorithms work well in practice.

[20] As before, we compute the TD criterion for ten random samples.

TABLE 3. Simulations: Firm entry game - Linear semi-gradient

	DGP (1)	TDL (2)	MSE (3)	TDL (4)	MSE (5)
A. 2nd order polynomial		*1000 markets*		*3000 markets*	
θ_{RS} (market size)	1.0	0.9715	0.0264	0.9845	0.0083
		(0.1601)		(0.0897)	
θ_{RN} (n. of entrants)	1.0	0.8956	0.2924	0.9581	0.0860
		(0.5309)		(0.2904)	
θ_{FC} (fixed cost)	1.7	1.7221	0.0867	1.6919	0.0262
		(0.2938)		(0.1617)	
θ_{EC} (entry cost)	1.0	1.0189	0.0042	1.0225	0.0018
		(0.0621)		(0.0353)	
Total MSE			0.4098		0.1222
Time per round (in sec)		0.32		0.93	
B. 3rd order polynomial		*1000 markets*		*3000 markets*	
θ_{RS} (market size)	1.0	0.9145	0.0289	0.9647	0.0088
		(0.1470)		(0.0869)	
θ_{RN} (n. of entrants)	1.0	0.6898	0.3264	0.8857	0.0912
		(0.4800)		(0.2797)	
θ_{FC} (fixed cost)	1.7	1.7811	0.0804	1.7134	0.0249
		(0.2718)		(0.1573)	
θ_{EC} (entry cost)	1.0	1.0172	0.0042	1.0219	0.0017
		(0.0622)		(0.0353)	
Total MSE			0.4398		0.1266
Time per round (in sec)		0.50		1.54	
C. 4th order polynomial		*1000 markets*		*3000 markets*	
θ_{RS} (market size)	1.0	0.8638	0.0360	0.9455	0.0101
		(0.1321)		(0.0846)	
θ_{RN} (n. of entrants)	1.0	0.5067	0.4222	0.8163	0.1070
		(0.4231)		(0.2707)	
θ_{FC} (fixed cost)	1.7	1.8335	0.0808	1.7337	0.0245
		(0.2510)		(0.1530)	
θ_{EC} (entry cost)	1.0	1.0158	0.0041	1.0212	0.0017
		(0.0620)		(0.0352)	
Total MSE			0.5431		0.1433
Time per round (in sec)		0.84		2.64	

Notes: The table reports results for 1000 simulations. Panels A, B and C use different sets of basis functions to parameterize $h(a, x)$ and $g(a, x)$. Column (1) shows the true parameter values in the model. Columns (2) and (4) report the empirical mean and standard deviation (in parentheses) for the estimated parameters, based on a sample of 1000 and 3000 markets, respectively. The mean squared errors are reported in columns (3) and (5). All results are based on the estimation method without correction function.

Data availability statement. The data and code underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.16184776.

References

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-Step GMM," *Review of Economic Studies*, 81, 919–943.

Aguirregabiria, V., and A. Magesan (2018): "Solution and Estimation of Dynamic Discrete Choice Structural Models Using Euler Equations," *Working paper*.

Aguirregabiria, V., and P. Mira (2002): "Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70, 1519–1543.

——— (2007): "Sequential Estimation of Dynamic Discrete Games," *Econometrica*, 75, 1–53.

——— (2010): "Dynamic Discrete Choice Structural Models: A Survey," *Journal of Econometrics*, 156, 38–67.

Almagro, M., and T. Domínguez-Iino (2025): "Location Sorting and Endogenous Amenities: Evidence from Amsterdam," *Econometrica*, 93, 1031–1071.

Arcidiacono, P., P. Bayer, F. A. Bugni, and J. James (2013): "Approximating High-Dimensional Dynamic Models: Sieve Value Function Iteration," in *Structural Econometric Models*: Emerald Group Publishing Limited, 45–95.

Arcidiacono, P., and J. B. Jones (2003): "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71, 933–946.

Arcidiacono, P., and R. A. Miller (2011): "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity," *Econometrica*, 79, 1823–1867.

Bajari, P., C. L. Bankard, and J. Levin (2007): "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75, 1331–1370.

Barwick, P. J., and P. A. Pathak (2015): "The Costs of Free Entry: An Empirical Study of Real Estate Agents in Greater Boston," *The RAND Journal of Economics*, 46, 103–145.

Benítez-Silva, H., G. Hall, G. J. Hitsch, G. Pauletto, and J. Rust (2000): "A Comparison of Discrete and Parametric Approximation Methods for Continuous-State Dynamic Programming Problems," *Working paper*.

Biau, G. (2012): "Analysis of a Random Forests Model," *Journal of Machine Learning Research*, 13, 1063–1095.

Buchholz, N., M. Shum, and H. Xu (2021): "Semiparametric Estimation of Dynamic Discrete Choice Models," *Journal of Econometrics*, 223, 312–327.

Bugni, F. A., and J. Bunting (2021): "On the Iterated Estimation of Dynamic Discrete Choice Games," *Review of Economic Studies*, 88, 1031–1073.

Chen, X., and Z. Qi (2022): "On Well-Posedness and Minimax Optimal Rates of Nonparametric Q-function Estimation in Off-Policy Evaluation," in *International Conference on Machine Learning*, 3558–3582, PMLR.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022): "Locally Robust Semiparametric Estimation," *Econometrica*, 90, 1501–1535.

Chernozhukov, V., W. K. Newey, and V. Semenova (2019): "Welfare Analysis in Dynamic Models," *arXiv preprint arXiv:1908.09173*.

Chernozhukov, V., W. K. Newey, and R. Singh (2018): "Learning L2-Continuous Regression Functionals via Regularized Riesz Representers," *arXiv preprint arXiv:1809.05224*.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer, 2nd edition.

Hotz, V. J., and R. A. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–529.

Hotz, V. J., R. A. Miller, S. Sanders, and J. Smith (1994): "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economic Studies*, 61, 265–289.

Ichimura, H., and W. K. Newey (2022): "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, 13, 29–61.

Kalouptsidi, M. (2014): "Time to Build and Fluctuations in Bulk Shipping," *American Economic Review*, 104, 564–608.

——— (2018): "Detection and Impact of Industrial Subsidies: The Case of Chinese Shipbuilding," *Review of Economic Studies*, 85, 1111–1158.

Keane, M. P., and K. I. Wolpin (1994): "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *The Review of Economics and Statistics*, 648–672.

Lange, S., T. Gabel, and M. Riedmiller (2012): "Batch Reinforcement Learning," in *Reinforcement Learning*: Springer, 45–73.

Munos, R., and C. Szepesvári (2008): "Finite-Time Bounds for Fitted Value Iteration," *Journal of Machine Learning Research*, 9, 815–857.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics* Volume 4: Elsevier, 2111–2245.

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

Pesendorfer, M., and P. Schmidt-Dengler (2008): "Asymptotic Least Squares Estimators for Dynamic Games," *Review of Economic Studies*, 75, 901–928.

Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 55, 999–1033.

Semenova, V. (2018): "Machine Learning for Dynamic Models of Imperfect Information and Semiparametric Moment Inequalities," *arXiv preprint arXiv:1808.02569*.

Sutton, R. S. (1988): "Learning to Predict by the Methods of Temporal Differences," *Machine Learning*, 3, 9–44.

Sutton, R. S., and A. G. Barto (2018): *Reinforcement Learning: An Introduction*: MIT Press, Cambridge, MA, 2nd edition.

Tauchen, G. (1986): "Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions," *Economics Letters*, 20, 177–181.

Tsitsiklis, J. N., and B. Van Roy (1997): "An Analysis of Temporal-Difference Learning with Function Approximation," *IEEE Transactions on Automatic Control*, 42, 674–690.

Appendix A. Proofs of main results

For the proofs of Theorems 1-2, we work within a more general setting than in the main text, by letting the distribution of (a_{it}, x_{it}) be time-varying. Let P_t denote the population distribution of (a, x) at time t. Also, let P denote the probability distribution of the

process $\{(a_1, x_1), \ldots, (a_T, x_T)\}$. Note that $P \equiv P_1 \times \cdots \times P_T$. Denote the expectation over P by $E[\cdot]$. We use the $o_p(\cdot)$ and $O_p(\cdot)$ notations to denote convergence in probability, and bounded in probability, respectively, under the probability distribution P.

We also need to extend the definitions of $\mathbb{P}$ and $\mathbb{E}[\cdot]$: Let $\mathbb{P}$ denote the relative frequency of occurrence of (a, x, a', x') in the data as $n \to \infty$, and $\mathbb{E}[\cdot]$ the corresponding expectation over $\mathbb{P}$. Note that P is different from $\mathbb{P}$ as the latter is the distribution of (a, x, a, x') after dropping the time index. However, the two are related as for any function f, we have $\mathbb{E}[f(a, x, a', x')] = (T-1)^{-1} \sum_{t=1}^{T-1} E[f(a_{it}, x_{it}, a_{it+1}, x_{it+1})]$. These updated definitions of $\mathbb{P}$ and $\mathbb{E}[\cdot]$ are applicable whenever we use these notations in the main text.

Note that due to the Markov process assumption, the conditional distribution $P(a_{t+1}, x_{t+1}|a_t, x_t)$ is always independent of t (indeed, one could always include t in x). Hence, $\mathbb{P}(a', x'|a, x) \equiv P(a_{t+1}, x_{t+1}|a_t, x_t)$ and $\mathbb{E}[f(a', x')|a, x] \equiv E[f(a_{t+1}, x_{t+1})|a_t, x_t]$ for all t. Also note that time stationarity of (a_{it}, x_{it}), if it holds, implies $P_t \equiv \mathbb{P}$ and $E_t[\cdot] \equiv \mathbb{E}[\cdot]$ for all t.

A.1. **Proof of Theorem 1.** Lemma 1 in Online Appendix B.2 implies ω^* exists. To prove that $\hat{\omega}$ exists, it suffices to show that $\hat{A} := \mathbb{E}_n\left[\phi\left(\beta\phi' - \phi\right)^\intercal\right]$ is invertible with probability approaching one. Recall that using our notation, $\hat{A} = (n(T-1))^{-1} \sum_i \sum_{t=1}^{T-1} \phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$, while $A = (T-1)^{-1} \sum_{t=1}^{T-1} E[\phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$. We can thus write $\left|\hat{A} - A\right| \leq (T-1)^{-1} \sum_{t=1}^{T-1} \left|\hat{A}_t - A_t\right|$, where $\hat{A}_t := n^{-1} \sum_i \phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$ and $A_t := E[\phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal]$. By Assumption 1(ii), $|\phi(a, x)|_\infty \leq M$ independent of k_ϕ, so

$$
\begin{aligned}
E\left|\hat{A}_t - A_t\right|^2 &= E\left|\frac{1}{n}\sum_i \phi_{it}\left(\beta\phi_{it+1} - \phi_{it}\right)^\intercal - E\left[\phi_{it}\left(\beta\phi_{it+1} - \phi_{it}\right)^\intercal\right]\right|^2 \\
&\leq \frac{1}{n}\sum_i E\left|\phi_{it}\left(\beta\phi_{it+1} - \phi_{it}\right)^\intercal\right|^2 \leq \frac{k_\phi^2 M^4}{n}.
\end{aligned}
$$

This proves $\left|\hat{A}_t - A_t\right| = O_p(k_\phi/\sqrt{n})$. But T is fixed, which implies that $\left|\hat{A} - A\right| = O_p(k_\phi/\sqrt{n})$ as well. We thus obtain $\bar{\lambda}(\hat{A}) \leq \bar{\lambda}(A) + \left|\hat{A} - A\right| \leq \bar{\lambda}(A) + o_p(1)$. Since $\bar{\lambda}(A) < 0$, this proves that $\bar{\lambda}(\hat{A}) < 0$ with probability approaching one, and subsequently, that $\hat{A}$ is invertible. This completes the proof of the first claim.

The second claim follows from Lemma 3 in Online Appendix B.2 and Assumption 1(iii).

To prove the last claim, we first show that with probability approaching one,

$$
|\hat{\omega} - \omega^*| \leq C(1-\beta)^{-1}\sqrt{\frac{k_\phi}{n}}, \tag{A.1}
$$

for some $C < \infty$. Define $b = \mathbb{E}[\phi z]$ and $\hat{b} = \mathbb{E}_n[\phi z]$. We then have $A\omega^* = b$ and $\hat{A}\hat{\omega} = \hat{b}$. We can combine the two equations to get

$$\hat{A}(\hat{\omega} - \omega^*) = (\hat{b} - b) + (A - \hat{A})\omega^*.$$

The above implies

$$(\hat{\omega} - \omega^*)^\intercal(-\hat{A})(\hat{\omega} - \omega^*) = (\hat{\omega} - \omega^*)^\intercal(b - \hat{b}) + (\hat{\omega} - \omega^*)^\intercal(\hat{A} - A)\omega^*. \tag{A.2}$$

We earlier showed $\left|\hat{A} - A\right| = O_p(k_\phi/\sqrt{n})$. Hence, $\underline{\lambda}(-\hat{A}) \geq \underline{\lambda}(-A) + o_p(1)$, so

$$(\hat{\omega} - \omega^*)^\intercal(-\hat{A})(\hat{\omega} - \omega^*) \geq c(1-\beta)\underline{\lambda}(\mathbb{E}[\phi\phi^\intercal])\left|\hat{\omega} - \omega^*\right|^2, \tag{A.3}$$

with probability approaching one, for any constant $c \in (0,1)$. Given (A.2) and (A.3),

$$|\hat{\omega} - \omega^*| \leq \frac{1}{c(1-\beta)\underline{\lambda}(\mathbb{E}[\phi\phi^\intercal])}\left(\left|\hat{b} - b\right| + \left|\hat{A}\omega^* - A\omega^*\right|\right),$$

with probability approaching one.

It remains to bound $\left|\hat{b} - b\right|$ and $\left|\hat{A}\omega^* - A\omega^*\right|$. As before, we can define $\hat{b}_t = n^{-1}\sum_i \phi_{it}z_{it}$ and $b_t = E[\phi_{it}z_{it}]$ to obtain

$$E\left|\hat{b}_t - b_t\right|^2 = E\left|\frac{1}{n}\sum_i \left\{\phi_{it}z_{it} - E\left[\phi_{it}z_{it}\right]\right\}\right|^2 \leq \frac{1}{n}E\left|\phi_{it}z_{it}\right|^2.$$

This proves

$$E\left|\hat{b} - b\right|^2 \leq \frac{1}{T-1}\sum_{t=1}^{T-1} E\left|\hat{b}_t - b_t\right|^2 \leq \frac{1}{n}\mathbb{E}\left[\left|\phi z\right|^2\right] \leq \frac{k_\phi L^2 M^2}{n} = O_p(k_\phi/n).$$

In a similar vein,

$$\begin{aligned} E\left|\hat{A}\omega^* - A\omega^*\right|^2 &= E\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_i \left\{\phi_{it}\left(\beta\phi_{it+1} - \phi_{it}\right)^\intercal\omega^* - E\left[\phi_{it}\left(\beta\phi_{it+1} - \phi_{it}\right)^\intercal\omega^*\right]\right\}\right|^2 \\ &= O_p\left(k_\phi/n\right), \end{aligned}$$

as long as $\mathbb{E}\left[\left|\phi\left(\beta\phi - \phi\right)^\intercal\omega^*\right|^2\right] = O(k_\phi)$. The latter holds assuming 1(ii)-(iv) since

$$\mathbb{E}\left[\left|\phi\left(\beta\phi^\intercal\omega^* - \phi^\intercal\omega^*\right)\right|^2\right] \leq k_\phi M^2(2 + 2\beta^2)\mathbb{E}\left[\left|\phi^\intercal\omega^*\right|^2\right],$$

$$\mathbb{E}\left[\left|\phi^\intercal\omega^*\right|^2\right]^{1/2} \leq \left\|\phi^\intercal\omega^* - h\right\|_2 + \left\|h\right\|_2 \leq O(k_\phi^{-\alpha}) + (1-\beta)^{-1}L < \infty,$$

where the second inequality uses $\left\|\phi^\intercal\omega^* - h\right\|_2 = O(k_\phi^{-\alpha})$ (as shown above), and $|h(\cdot,\cdot)|_\infty \leq (1-\beta)^{-1}|z(\cdot,\cdot)|_\infty < (1-\beta)^{-1}L$ (which can be easily verified using (2.2) and Assumption

1(iv)). Combining the above, there exists $C < \infty$ such that $|\hat{\omega} - \omega^*| \leq C\sqrt{k_\phi/n}$, with probability approaching one. We have thus shown (A.1).

Now,

$$\begin{aligned}\|\phi^\intercal\hat{\omega} - h\|_2^2 &\leq 2\|\phi^\intercal\hat{\omega} - \phi^\intercal\omega^*\|_2^2 + 2\|\phi^\intercal\omega^* - h\|_2^2 \\ &= 2(\hat{\omega} - \omega^*)^\intercal\mathbb{E}[\phi\phi^\intercal](\hat{\omega} - \omega^*) + 2\|\phi^\intercal\omega^* - h\|_2^2 \\ &\leq 2\bar{\lambda}(\mathbb{E}[\phi\phi^\intercal])O_p\left(\frac{k_\phi}{n}\right) + O_p(k_\phi^{-2\alpha}),\end{aligned}$$

where the final inequality follows from the second claim of this theorem and (A.1). The last claim then follows from the above along with the fact that, by Assumption 1(iv), $\bar{\lambda}(\mathbb{E}[\phi\phi^\intercal]) \leq \|\phi\|_2^2 \leq M^2 k_\phi$.

A.2. **Proof of Theorem 2.** The first two claims follow from steps analogous to those in Theorem 1. We thus need to show that with probability approaching one,

$$\left|\hat{\xi} - \xi^*\right| \leq C(1-\beta)^{-1}\sqrt{k_r/n}, \tag{A.4}$$

for some $C < \infty$. The third claim is a straightforward consequence of this.

Recall that we use cross-fitting to estimate ξ^*. Let n_1, n_2 denote the sample sizes, and $\hat{\eta}_1, \hat{\xi}_1$ and $\hat{\eta}_2, \hat{\xi}_2$ the estimates of η and ξ^* in the two folds. We shall show that $|\hat{\xi}_1 - \xi^*| = O_p(\sqrt{k_r/n})$ (and similarly $|\hat{\xi}_2 - \xi^*| = O_p(\sqrt{k_r/n})$), and therefore $|\hat{\xi} - \xi^*| = O_p(\sqrt{k_r/n})$. To this end, let $A_r := \mathbb{E}[rr^\intercal]$, $b_r := \mathbb{E}[r(a,x)e(a',x';\eta)]$, $\hat{A}_r^{(1)} := \mathbb{E}_n^{(1)}[rr^\intercal]$ and $\hat{b}_r^{(1)} := \mathbb{E}_n^{(1)}[r(a,x)e(a',x';\hat{\eta}_2)]$, where $\mathbb{E}_n^{(1)}[\cdot]$ denotes the empirical expectation using only the first fold. Let $\varsigma(a,x,a',x';\eta) := r(a,x)e(a',x';\eta)$ and $\varsigma_{it}(\eta) := r(a_{it},x_{it})e(a_{it+1},x_{it+1};\eta)$.

Based on the above definitions, we have $\hat{A}_r^{(1)}\hat{\xi}_1 = \hat{b}_r^{(1)}$, and $A_r\xi^* = b_r$. Comparing with the proof of Theorem 1, the only difference is in the treatment of $|\hat{b}_r^{(1)} - b_r|$. As before, define $\hat{b}_{rt}^{(1)} := n^{-1}\sum_i \varsigma_{it}(\hat{\eta}_2)$ and $b_{rt} := E[\varsigma_{it}(\eta)]$. We then have $|\hat{b}_r^{(1)} - b_r| = (T-1)^{-1}\sum_{t=1}^{T-1}|\hat{b}_{rt}^{(1)} - b_{rt}|$. Since T is finite, it suffices to bound $|\hat{b}_{rt}^{(1)} - b_{rt}|$ for some arbitrary t. Now, by similar arguments as in the proof of Theorem 1, we have

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\{\varsigma_{it}(\eta) - E[\varsigma_{it}(\eta)]\} = O_p\left(\sqrt{k_r/n}\right).$$

Hence (A.4) follows once we show

$$\hat{b}_{rt}^{(1)} - b_{rt} = \frac{1}{n_1}\sum_{i=1}^{n_1}\{\varsigma_{it}(\eta) - E[\varsigma_{it}(\eta)]\} + o_p\left(\sqrt{k_r/n}\right). \tag{A.5}$$

We now prove (A.5). Denoting the set of observations in the second fold by $\mathcal{N}_2$:

$$\hat{b}_{rt}^{(1)} - b_{rt} - \frac{1}{n_1}\sum_{i=1}^{n_1}\{\varsigma_{it}(\eta) - E[\varsigma_{it}(\eta)]\}$$
$$= \underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1}\{(\varsigma_{it}(\hat{\eta}_2) - \varsigma_{it}(\eta)) - (E[\varsigma_{it}(\hat{\eta}_2)|\mathcal{N}_2] - E[\varsigma_{it}(\eta)])\}}_{:=R_{1nt}} + \underbrace{\{E[\varsigma_{it}(\hat{\eta}_2)|\mathcal{N}_2] - E[\varsigma_{it}(\eta)]\}}_{:=R_{2nt}}.$$

First consider the term R_{1nt}. Define

$$\delta_{it} := (\varsigma_{it}(\hat{\eta}_2) - \varsigma_{it}(\eta)) - (E[\varsigma_{it}(\hat{\eta}_2)|\mathcal{N}_2] - E[\varsigma_{it}(\eta)]).$$

Clearly, $E[\delta_{it}|\mathcal{N}_2] = 0$. We then have

$$E\left[|R_{1nt}|^2\middle|\mathcal{N}_2\right] = \frac{1}{n_1}E\left[|\delta_{it}|^2\middle|\mathcal{N}_2\right] = \frac{1}{n_1}E\left[|\varsigma_{it}(\hat{\eta}_2) - \varsigma_{it}(\eta)|^2\middle|\mathcal{N}_2\right]. \tag{A.6}$$

Now for any (a, x, a', x'), from the definition of $\varsigma(\cdot)$, with probability approaching one,

$$\begin{aligned}|\varsigma(a,x,a',x';\hat{\eta}_2) - \varsigma(a,x,a',x';\eta)| &\leq |r(a,x)||\ln\hat{\eta}_2 - \ln\eta| \\ &\leq M\sqrt{k_r}|\ln\hat{\eta}_2 - \ln\eta| \leq 2M\sqrt{k_r}\delta^{-1}|\hat{\eta}_2 - \eta|, \end{aligned}\tag{A.7}$$

where the second inequality follows from Assumption 2(ii), and the third follows from 2(v).[21] In view of (A.6) and (A.7), there exists $C < \infty$ such that

$$\begin{aligned}E\left[|R_{1nt}|^2\middle|\mathcal{N}_2\right] &\leq \frac{Ck_r}{n_1}E\left[|\hat{\eta}_2(a_{it+1}, x_{it+1}) - \eta(a_{it+1}, x_{it+1})|^2\middle|\mathcal{N}_2\right] \\ &\leq \frac{Ck_rT}{n_1}\|\hat{\eta}_2 - \eta\|_2^2 = o_p(k_r/n),\end{aligned}$$

where the last equality follows by Assumption 2(v). This proves

$$|R_{1nt}| = o_p(\sqrt{k_r/n}). \tag{A.8}$$

Next consider the term R_{2nt}. Note that $E[\varsigma_{it}(\eta)]$ is twice Fréchet differentiable. Indeed, in the main text we have shown that $\partial_\eta E[\varsigma_{it}(\eta)] = 0$, where $\partial_\eta\cdot$ denotes the Fréchet differential of $E[\varsigma_{it}(\eta)]$ (cf. equation (3.7)). Furthermore, $\ln\eta$ is an infinitely differentiable function of $\eta(a,x)$ with second derivatives bounded by δ^{-2} (since $1 > \eta(a,x), \hat{\eta}(a,x) > \delta$ under Assumption 2(vi)). This implies $\varsigma(a,x,a',x';\eta)$ is also infinitely differentiable with respect to $\eta(a,x)$, with second derivatives bounded by $\delta^{-2}|r(a,x)| \lesssim \delta^{-2}\sqrt{k_r}$, where the '$\lesssim$' is due to Assumption 1(ii) which implies $|r(\cdot)|_\infty$ is bounded. The above facts imply,

[21] In particular, we have used the fact $\hat{\eta}_2 > \delta + o_p(1)$ which follows from $\eta > \delta$ and $|\hat{\eta}_2 - \eta| = o_p(1)$.

through a second order Taylor expansion, that

$$|E\left[\varsigma_{it}(\hat{\eta}_2) - \varsigma_{it}(\eta)|\mathcal{N}_2\right]| \leq C_1 k_r E\left[\left|\hat{\eta}_2(a_{it+1}, x_{it+1}) - \eta(a_{it+1}, x_{it+1})\right|^2 \middle| \mathcal{N}_2\right],$$

for some $C_1 < \infty$. Hence,

$$\begin{aligned} E\left[|R_{2nt}|^2 \middle| \mathcal{N}_2\right] &\leq C_1 k_r E\left[\left|\hat{\eta}_2(a_{it+1}, x_{it+1}) - \eta(a_{it+1}, x_{it+1})\right|^2 \middle| \mathcal{N}_2\right] \\ &= C_1 T k_r \left\|\hat{\eta}_2 - \eta\right\|_2^2 = o_p(k_r/n). \end{aligned} \tag{A.9}$$

(A.8) and (A.9) imply (A.5), leading to the desired claim (A.4).

A.3. **Proof of Theorem 5.** Define

$$\phi(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; h, g, \eta, \lambda, \theta) = \lambda(a, x; \theta)\left\{z(a, x)^\intercal \theta + \beta e(a', x'; \eta) + \beta V(a', x'; \theta, h, g) - V(a, x; \theta, h, g)\right\}.$$

Denote by $\tilde{\theta}^{(l)}$ the preliminary estimator of θ^* from the l-th data fold under cross-fitting.

By similar arguments as in Newey and McFadden (1994), $\tilde{\theta}^{(l)}$ is consistent for θ^* under Assumptions 1-5. We now prove the stronger statement that

$$\tilde{\theta}^{(l)} - \theta^* = o_p(n^{-1/4}). \tag{A.10}$$

Without loss of generality, take $l = 1$. By a first order Taylor expansion using the definition of $\tilde{\theta}^{(1)}$,

$$\tilde{\theta}^{(1)} - \theta^* = \mathbb{E}_n^{(1)}\left[\nabla_\theta m(a, x; \breve{\theta}^{(1)}, \hat{h}^{(1)}, \hat{g}^{(1)})\right]^{-1} \mathbb{E}_n^{(1)}\left[m(a, x; \theta^*, \hat{h}^{(1)}, \hat{g}^{(1)})\right],$$

for some $\breve{\theta}^{(1)}$ such that $\left\|\breve{\theta}^{(1)} - \theta^*\right\| \leq \left\|\tilde{\theta}^{(1)} - \theta^*\right\|$. Now, $\mathbb{E}_n^{(1)}\left[\nabla_\theta m(a, x; \breve{\theta}^{(1)}, \hat{h}^{(1)}, \hat{g}^{(1)})\right]^{-1} = O_p(1)$ by Assumption 5(ii). Furthermore,

$$\begin{aligned} \mathbb{E}_n^{(1)}\left[m(a, x; \theta^*, \hat{h}^{(1)}, \hat{g}^{(1)})\right] &= \mathbb{E}_n^{(1)}\left[m(a, x; \theta^*, h, g)\right] + \mathbb{E}_n^{(1)}\left[m(a, x; \theta^*, \hat{h}^{(1)}, \hat{g}^{(1)}) - m(a, x; \theta^*, h, g)\right] \\ &= O_p(n^{-1/2}) + \underbrace{\mathbb{E}_n^{(1)}\left[m(a, x; \theta^*, \hat{h}^{(1)}, \hat{g}^{(1)}) - m(a, x; \theta^*, h, g)\right]}_{:=R_1}, \end{aligned}$$

where the second equality follows from Chebyshev's inequality as $m(a, x; \theta^*, h, g)$ is uniformly bounded under the stated assumptions (continuity of h, g; compactness of the support $\mathcal{X}$ of x). It remains to bound R_1. To this end, we use (B.5) in Online Appendix B.4. Note that the Riesz representers $\psi_h(\cdot, \cdot; \theta^*, h, g), \psi_g(\cdot, \cdot; \theta^*, h, g)$ (defined in B.4.2) are uniformly bounded under compactness of $\mathcal{X}$ and smoothness of h. It therefore follows that

$$R_1 \lesssim \left\|\hat{h}^{(1)} - h\right\|_2 + \left\|\hat{g}^{(1)} - g\right\|_2 = o_p(n^{-1/4}), \tag{A.11}$$

where the last equality uses Assumption 5(iii).

To complete the proof of the theorem, it suffices to verify Assumptions 1-3 and 5 in Chernozhukov et al. (2022).

Assumption 1 of Chernozhukov et al. (2022) requires

$$\mathbb{E}\left[\left\|\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta^*, h, g, \eta, \lambda, \theta^*)\right\|^2\right] < \infty, \tag{A.12}$$

$$\left\|m(\cdot, \cdot; \theta^*, \hat{h}, \hat{g}) - m(\cdot, \cdot; \theta^*, h, g)\right\|_2^2 = o_p(1), \tag{A.13}$$

$$\left\|\phi(\cdot, \cdot; \hat{h}, \hat{g}, \hat{\eta}, \lambda, \theta^*) - \phi(\cdot, \cdot; h, g, \eta, \lambda, \theta^*)\right\|_2^2 = o_p(1), \text{ and} \tag{A.14}$$

$$\left\|\phi(\cdot, \cdot; h, g, \eta, \hat{\lambda}, \tilde{\theta}) - \phi(\cdot, \cdot; h, g, \eta, \lambda, \theta^*)\right\|_2^2 = o_p(1). \tag{A.15}$$

We first note that the smoothness conditions on h, g, η (imposed in Assumptions 1-3) along with the compactness of the domain $\mathcal{X}$, ensure $z(\cdot), e(\cdot, \cdot; \eta), h(\cdot), g(\cdot), m(a, x; \theta^*, h, g)$ and $\psi(a, x; \theta^*, h, g)$ are all uniformly bounded in (a, x). By standard dynamic programming arguments, they also imply $\lambda(a, x; \theta^*)$ is uniformly bounded in (a, x).[22] These bounds lead to $\|\zeta(\cdot, \cdot; \theta^*, h, g, \eta, \lambda, \theta^*)\|_\infty \leq M < \infty$. This shows that the first requirement (A.12) holds. Next, we show (A.13): by the form of $m(a, x; \theta, h, g)$ and the fact that

$$\left|\pi(\breve{a}, x; \theta^*, \hat{h}, \hat{g}) - \pi(\breve{a}, x; \theta^*, h, g)\right| \leq \sum_{\bar{a} \in \mathcal{A}} \left\{\left|\theta^{*\intercal}\hat{h}(\bar{a}, x) - \theta^{*\intercal}h(\bar{a}, x)\right| + \left|\hat{g}(\bar{a}, x) - g(\bar{a}, x)\right|\right\}$$

for all $\breve{a}, x$, some straightforward algebra gives

$$\begin{aligned}\left|m(\cdot, \cdot; \theta^*, \hat{h}, \hat{g}) - m(\cdot, \cdot; \theta^*, h, g)\right| \leq & \sum_{\breve{a} \in \mathcal{A}} \sum_{\bar{a} \in \mathcal{A}} \left\{\left|\theta^{*\intercal}\hat{h}(\bar{a}, x) - \theta^{*\intercal}h(\bar{a}, x)\right| + \left|\hat{g}(\bar{a}, x) - g(\bar{a}, x)\right|\right\} \left|h(\breve{a}, x)\right| \\ & + \sum_{\breve{a} \in \mathcal{A}} \left|\hat{h}(\breve{a}, x) - h(\breve{a}, x)\right|.\end{aligned}$$

Observe that $|h(\breve{a}, \cdot)|_\infty < \infty$ for each $\breve{a}$ by the assumptions of compact support for x and continuity of h. We thus obtain

$$\begin{aligned}&\left\|m(\cdot, \cdot; \theta^*, \hat{h}, \hat{g}) - m(\cdot, \cdot; \theta^*, h, g)\right\|_2 \\ &\lesssim \mathbb{E}^{1/2}\left[\sum_{\breve{a} \in \mathcal{A}} \left|\hat{h}(\breve{a}, x) - h(\breve{a}, x)\right|^2\right] + \mathbb{E}^{1/2}\left[\sum_{\breve{a} \in \mathcal{A}} \left|\hat{g}(\breve{a}, x) - g(\breve{a}, x)\right|^2\right] \\ &\lesssim \delta^{-1}\mathbb{E}^{1/2}\left[\sum_{\breve{a} \in \mathcal{A}} \pi(\breve{a}, x; \theta^*, h, g) \left|\hat{h}(\breve{a}, x) - h(\breve{a}, x)\right|^2\right] + \delta^{-1}\mathbb{E}^{1/2}\left[\sum_{\breve{a} \in \mathcal{A}} \pi(\breve{a}, x; \theta^*, h, g) \left|\hat{g}(\breve{a}, x) - g(\breve{a}, x)\right|^2\right] \\ &\lesssim \left\|\hat{h} - h\right\|_2 + \left\|\hat{g} - g\right\|_2 = o_p(1),\end{aligned} \tag{A.16}$$

[22] Note that $|\lambda(\cdot, \cdot; \theta^*)|_\infty \leq (1 - \beta)^{-1} |\psi(\cdot, \cdot; \theta^*, h, g)|_\infty$ by the fixed point definition of $\lambda(\cdot)$.

where the second inequality employs $\pi(a,x;\theta^*,h,g) \equiv \eta(a,x) > \delta > 0$, as stated in Assumption 5(iii), and the last equality also follows from the same assumption. This proves (A.13). The third requirement (A.14) follows from the boundedness of $\lambda(\cdot,\cdot;\theta^*)$ together with Assumption 5(iii). Finally, (A.15) follows from the boundedness of $z(\cdot), h, g, \ln\eta^{-1}$ (all of which hold under Assumption 5(iii)), along with the consistency of $\tilde{\theta}$ and Assumption 5(iv).

Assumption 2 of Chernozhukov et al. (2022) requires

$$\begin{aligned}\sqrt{n}\mathbb{E}\left[\tilde{\Delta}(\tilde{\mathbf{a}},\tilde{\mathbf{x}})\right] = o_p(1), \text{ and } \mathbb{E}\left[\left|\tilde{\Delta}(\tilde{\mathbf{a}},\tilde{\mathbf{x}})\right|^2\right] = o_p(1), \text{ where} \qquad (A.17)\\ \tilde{\Delta}(\tilde{\mathbf{a}},\tilde{\mathbf{x}}) := \left\{\phi(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\hat{h},\hat{g},\hat{\eta},\hat{\lambda},\tilde{\theta}) - \phi(\tilde{\mathbf{a}},\tilde{\mathbf{x}};h,g,\eta,\hat{\lambda},\tilde{\theta})\right\}\\ - \left\{\phi(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\hat{h},\hat{g},\hat{\eta},\lambda,\theta^*) - \phi(\tilde{\mathbf{a}},\tilde{\mathbf{x}};h,g,\eta,\lambda,\theta^*)\right\}.\end{aligned}$$

To this end, observe that

$$\begin{aligned}\tilde{\Delta}(\tilde{\mathbf{a}},\tilde{\mathbf{x}}) = &\left(\hat{\lambda}(a,x;\tilde{\theta})\tilde{\theta}^\intercal - \lambda(a,x;\theta^*)\theta^{*\intercal}\right)\left(\beta\hat{\Delta}_h(a',x') - \hat{\Delta}_h(a,x)\right)\\ &+ \left(\hat{\lambda}(a,x;\tilde{\theta}) - \lambda(a,x;\theta^*)\right)\left(\beta\hat{\Delta}_g(a',x') - \hat{\Delta}_g(a,x)\right) +\\ &- \left(\hat{\lambda}(a,x;\tilde{\theta}) - \lambda(a,x;\theta^*)\right)\beta\hat{\Delta}_{\ln\eta}(a',x')\end{aligned}$$

where $\hat{\Delta}_h := \hat{h} - h$, $\hat{\Delta}_g := \hat{g} - g$ and $\hat{\Delta}_{\ln\eta} := \ln\hat{\eta} - \ln\eta$. In view of the $n^{-1/4}$ consistency of $\tilde{\theta}$ along with Assumptions 5(iii)-(iv), straightforward algebra shows $\mathbb{E}\left[\left|\tilde{\Delta}(a,x)\right|\right] = o_p(n^{-1/2})$. This proves the first part of (A.17). The second part follows immediately from the consistency of $\tilde{\theta}$ and Assumption 5(iv) after noting that $\hat{\Delta}_h(\cdot,\cdot)$, $\hat{\Delta}_g(\cdot,\cdot)$ and $\hat{\Delta}_{\ln\eta}(\cdot,\cdot)$ are all uniformly bounded (due to Assumption 5(iii) and the compactness of $\mathcal{X}$).

Assumption 3 of Chernozhukov et al. (2022) requires existence of some $C < \infty$ independent of $\tilde{h}, \tilde{g}$ and $\tilde{\eta}$ such that

$$\begin{aligned}\mathbb{E}\left[\phi(\tilde{\mathbf{a}},\tilde{\mathbf{x}};h,g,\eta,\hat{\lambda},\tilde{\theta})\right] = 0 \text{ and}\\ \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta^*,\tilde{h},\tilde{g},\tilde{\eta},\lambda,\theta^*)] \leq C\left\{\left\|\tilde{h}-h\right\|_2^2 + \|\tilde{g}-g\|_2^2 + \|\tilde{\eta}-\eta\|_2^2\right\}\end{aligned}$$

for all $\left\|\tilde{h}-h\right\|_2, \|\tilde{g}-g\|_2, \|\tilde{\eta}-\eta\|_2$ small enough and where the space of $\tilde{\eta}$ is $\{\tilde{\eta} : \inf_{a,x}\tilde{\eta}(a,x) > \delta > 0\}$.[23] These requirements are verified in Section B.4 in Online Appendix B.

Finally, Assumption 5 of Chernozhukov et al. (2022) is directly equivalent to Assumption 5(ii).

[23]The assumption in Chernozhukov et al. (2022) also requires L_2 consistency of $\hat{h}, \hat{g}, \hat{\eta}$ which is directly stated as Assumption 5(iii).