TEMPORAL-DIFFERENCE ESTIMATION OF DYNAMIC DISCRETE CHOICE MODELS

KARUN ADUSUMILLI[1] AND DITA ECKARDT[2]

Abstract. We propose a new algorithm to estimate the structural parameters in dynamic discrete choice models. The algorithm is based on the conditional choice probability approach, but uses the idea of Temporal-Difference learning from the Reinforcement Learning literature to estimate the various value function like terms in the pseudo-likelihood estimator. In estimating these terms with functional approximations using basis functions, our approach has the advantage of naturally allowing for continuous state spaces. Furthermore, it does not require specification of transition probabilities, and even estimation of choice probabilities can be avoided using a recursive procedure. Computationally, our algorithm only requires solving a low dimensional linear equation. For the estimation of dynamic games, our procedure does not require integrating over the actions of other players, which further heightens the computational advantage. We show that our estimator is consistent, and efficient under discrete state spaces. In settings with continuous states, we propose easy to implement locally robust corrections in order to achieve parametric rates of convergence. Preliminary Monte Carlo simulations confirm the workings of our algorithm.

This version: December 30, 2020
[1]Department of Economics, University of Pennsylvania, akarun@sas.upenn.edu. [2]Institute for Fiscal Studies, dita.eckardt@ifs.org.uk. We would like to thank Xiaohong Chen, Frank Diebold, Aviv Nevo, Whitney Newey and Frank Schorfheide for helpful discussions.

1. Introduction

Dynamic discrete choice (DDC) models are frequently used to describe the intertemporal 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 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 typically requires an explicit solution to the dynamic programming problem in addition to the optimization of an estimation criterion. In a seminal contribution, Rust (1987) develops an iterative solution algorithm, the Nested Fixed Point algorithm, that repeatedly solves the dynamic programming problem and searches for the root of the likelihood equations to update the structural parameters. To ease the computational burden associated with fully solving the dynamic optimization problem in each iteration, alternative methods have been developed. A key advance has been Hotz and Miller's (1993) Conditional Choice Probability (CCP) algorithm which avoids the repeated solution of the intertemporal optimization problem by taking advantage of a mapping between value function differences and conditional choice probabilities. This idea has subsequently been refined by Hotz et al. (1994) who suggest a simulation-based CCP method, and Aguirregabiria and Mira (2002) who develop a more efficient recursive CCP algorithm, the nested pseudo-likelihood (NPL) algorithm. More recently, Arcidiacono and Miller (2011) exploit the property of finite dependence to speed up CCP estimation. This idea has been extended by Chernozhukov et al. (2018) to high dimensional states, also under finite dependence. Separately, Semenova (2018) also allows for high-dimensional states, but the parameters are only partially identified.

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. While the CCP algorithm substantially reduces the computational burden compared to traditional methods in such settings, it becomes computationally infeasible if the number of discrete state variables is large. This problem is even more apparent when the underlying state variables are continuous and the resulting discretization gives rise to a very high-dimensional state space. An application that is particularly affected by this issue is the estimation of dynamic discrete games, where the strategic interaction of agents means that the state space increases exponentially with the number of players. Furthermore, it is uncommon for finite dependence to hold under dynamic games. Existing methods in discrete state space settings such as the pseudo-likelihood estimator proposed by Aguirregabiria and Mira (2007) or the minimum distance estimator suggested by Pesendorfer and Schmidt-Dengler (2008) become computationally difficult when the state space is large. If the states are continuous, discretization may be avoided by using forward Monte Carlo simulations (Bajari et al., 2007), but this may become very involved as the number of continuous state variables or players increases.

To overcome these limitations, we propose a new algorithm for the estimation of DDC models. Our approach is based on traditional CCP methods, but makes use of a Temporal-Difference

[1] See Aguirregabiria and Mira (2010) for a detailed survey of the literature on the estimation of DDC models.

(TD) method from the Reinforcement Learning literature to provide functional approximations for the various value function terms in the pseudo-likelihood estimator.[2] We start by choosing a set of basis functions in actions and observed state variables. We then project the value function operator onto the linear span of these basis functions and compute the resulting fixed point (of the projected value function operator). This fixed point is our functional approximation to the value function. Unlike most existing estimation approaches, our algorithm does not require any specification or estimation of transition probabilities. Estimating the parameters requires a soving a single, low-dimensional linear equation. In the unlikely case where the dimensionality of the state space and therefore matrix makes the inversion computationally difficult, we propose an alternative stochastic gradient procedure to obtain the functional approximations for the terms in the value function. With these at hand, estimation of the structural parameters can proceed with standard methods such as maximum likelihood estimation (MLE) or minimum distance estimation.

As noted earlier, a useful feature of our approach is that we do not need to estimate, or impose any restrictions on the form of transition probabilities. Only an estimate for the conditional choice probabilities is required. Aguirregabiria and Mira (2002) show that, if the state variables are discrete, the error from estimation of choice probabilities does not have a first-order impact on the estimation of structural parameters, but this result does not carry over to estimation of transition probabilities. Therefore, if the state variables are continuous, the pseudo maximum likelihood estimator for the structural parameters will no longer converge at parametric rates. We explain how, following Ackerberg et al. (2014), Newey (1994) and Chernozhukov et al. (2018), our estimation approach for the functional approximations and the structural parameters can be easily adapted to continuous state spaces using a correction term to provide locally robust estimators. The resulting estimator converges at parametric rates under continuous states and unrestricted transition probabilities. We also propose a recursive version of our algorithm, similar to the NPL algorithm by Aguirregabiria and Mira (2002), in which the conditional choice probabilities are updated as part of the estimation of the functional form approximations. Finally, we incorporate permanent unobserved heterogeneity into our methods by combining the TD estimation with an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977).

Our estimator is thus consistent, converges at parametric rates and computationally very cheap, even in models that do not exhibit a finite dependence property. Most importantly, our TD estimator provides a feasible estimation method when the state variables are continuous or the state space is large. This is particularly important for the estimation of dynamic discrete games. Even with discrete states, existing methods for estimation of dynamic games ((Bajari et al., 2007); Aguirregabiria and Mira (2007); Pesendorfer and Schmidt-Dengler (2008)) require integrating out the actions of the other players. With many players, or under continuous states this can get quite cumbersome. 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. Furthermore, the statistical properties of our estimator - consistency, parametric rates of convergence etc. - also carry over to estimation of dynamic

[2] See Sutton and Barto (2018) for details on TD learning.

games with continuous states and unknown transition probabilities. A Monte Carlo study based on the Rust (1987) bus engine replacement problem confirms the workings of our algorithm.

While most of the computational gain is achieved in models with high-dimensional state space, our approach is also as efficient as other methods in models with fewer state variables. In fact, we show that in cases where the underlying states and actions are discrete, the basis function in our functional approximations can be chosen such that our estimate is numerically identical to the one obtained from standard CCP estimators. We therefore view our method as broadening the class of DDC models that can be structurally estimated, while being as efficient as existing estimation approaches for simpler versions of the DDC problem.

In making use of a TD step in the estimation, our method relates to the literature on Reinforcement Learning. Reinforcement Learning is an area of machine learning which describes learning about how to map states into actions so as to maximize an expected payoff.[3] A central component in Reinforcement Learning is the estimation of value functions. Unlike traditional dynamic programming methods, TD learning updates the current value function using sample successors. In contrast to other sample updating methods, it uses an estimate of the return instead of the actual return as target. Finally, it also employs functional approximations to approximate the value functions under continuous states. The combination of functional approximation, sample successors and estimated returns makes TD estimation extremely fast. For this reason TD algorithms are the standard method of choice for approximating value functions in Reinforcement Learning. The idea of TD learning has a long history, but the formulation in its current form is due to Sutton (1988). Tsitsiklis and van Roy (1997) studied the theoretical properties of the algorithm under functional approximation. However these were derived in the setup of online learning, whereas we intend to use our TD algorithm on a given set of observational data, i.e in an offline manner. Consequently we develop the statistical properties of TD estimation using offline data. We find that TD learning behaves very similarly to the usual series approximation in terms of convergence rates. Indeed, due to the similarity in statistical and computational properties, we like to think of TD estimation as the counterpart of least-squares regression, but for approximating value functions.

Our methods also contribute to the literature on approximating value functions. A number of techniques have been proposed for this in Economics, including parametric policy iteration (Hall et al., 2000), simulation and interpolation (Keane and Wolpin 1994), and sieve value function iteration (Arcidiacono et al., 2013). The last of these comes closest in spirit to our own approach, as the authors propose a non-parametric approximation of the value function. The difference, however, is that Arcidiacono et al. (2013) propose minimizing the TD error in the sup norm, while we minimize the projected TD error in expectation. The latter is much easier to compute and we are also able to provide strong statistical guarantees when the choice and transition probabilities are unknown, with rates of approximation that mirror standard series estimation. We also refer to Section 11.4 of Sutton and Barto (2018) for a useful discussion on the differences between minimizing the TD error and the projected TD error.

[3]See Sutton and Barto (2018) for a detailed treatment of Reinforcement Learning.

The remainder of this paper is organized as follows. Section 2 outlines the setup of the DDC model and fixes notation. Section 3 describes our TD estimation method for the functional approximations of the value functions, proves its theoretical properties and describes the second-step estimation of the structural parameters under discrete and continuous state variables. Section 4 describes various extensions including a recursive version of our algorithm which avoids the initial estimation of conditional choice probabilities. Section 5 incorporates permanent unobserved heterogeneity into our algorithm. Section 6 discusses the estimation of dynamic discrete games. Section 7 provides preliminary Monte Carlo simulations for our algorithm using a version of the Rust (1987) bus engine replacement problem. Section 8 concludes.

2. Setup

We start with a single agent DDC model. Our treatment of this uses the same notation as Aguirregabiria and Mira (2010).

We consider a model in discrete time with $t = 1, \dots, T;\ 2 \le T < \infty$ periods and $i = 1, \dots, n$ agents. Typically $T \ll n$ in applications, so we shall always work within an asymptotic regime where $n \to \infty$ but T is fixed. We assume that the individuals are homogeneous, relegating extensions for unobserved heterogeneity to Section 5. In each period, an agent chooses among A mutually exclusive actions, each of which is denoted by a. The payoff from the action depends on the current state x. In particular, choosing action a when the state is x gives the agent an instantaneous utility of $z(a, x)^\intercal \theta + e$, where $z(a, x)$ is some known vector valued function of a, x and e is an idiosyncratic error term. We denote the realization of the state of an individual i at time t by x_{it}, and her corresponding action and error terms by a_{it} and e_{it}. We shall 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 immediately following the actions and states (a, x), where $x' \sim f_X(\cdot|a, x)$. We do not make any assumptions about f_X. The utility from future periods is discounted by β.

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

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

The econometrician observes the state action pairs $(\mathbf{x}_i, \mathbf{a}_i) = \{(x_{i1}, a_{i1}), \dots, (x_{iT}, a_{iT})\}$ for all individuals, but not the idiosyncratic error terms e_{it}. Using this data, the econometrician aims to recover the structural parameters θ^*. By now, a number of different algorithms have been proposed to estimate θ^*. One such algorithm, which is very popular in the literature due to its computational simplicity, is the CCP method due to Hotz and Miller (1993). This has been subsequently refined in many ways by Hotz et al. (1994), Aguirregabiria and Mira (2002), and Arcidiacono and Miller (2011), among others.

CCP methods utilize the knowledge of the conditional choice probabilities of choosing action a given state x. We shall denote these by $P_t(a|x)$ for a given period t but shall 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 here 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 $e(a, x) = \mathcal{G}(P(a|x))$. For concreteness we shall assume in this paper that e follows a Type I Extreme Value distribution, which is perhaps the most common choice in the literature. In this case $e(a, x) = \gamma - \ln P(a|x)$, where γ is the Euler constant.

The standard procedure in the CCP approach is as follows: Under the given distributional assumptions, the parameters are obtained as the maximizers of the pseudo-likelihood function

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

where $h(.)$ and $g(.)$ solve the following recursive expressions:

$$h(a, x) = z(a, x) + \beta \sum_{x'} f_X(x'|a, x) \sum_{a'} P(a'|x) h(a', x'),$$
$$g(a, x) = \beta \sum_{x'} f_X(x'|a, x) \sum_{a'} P(a'|x) \left\{e(a', x') + g(a', x')\right\}.$$

Note that we omit the subscripts in (a, x) to denote the random variables, as opposed to realizations (a_{it}, x_{it}). The above assumes a discrete state space. To obtain more insight, let us convert the above equations to expectations:

$$\begin{aligned} h(a, x) &= z(a, x) + \beta \mathbb{E}\left[h(a', x')|a, x\right], \\ g(a, x) &= \beta \mathbb{E}\left[e(a', x') + g(a', x')|a, x\right], \end{aligned} \tag{2.1}$$

where $\mathbb{E}[.]$ denotes the expectation over the distribution of (a', x') conditional on (a, x). Note that $\mathbb{P}$ is a function of the distribution F of the transition and choice probabilities given by (f_X, P). The above formulation is also valid for continuous state spaces. Both $h(a, x)$ and $g(a, x)$ have a 'value-function' form, which turns out to be useful as there now exist fast algorithms for computing value functions.

Observe that $h(.)$ and $g(.)$ are functions of the probability distributions f_X and $P(.|.)$, which represent the transition and conditional choice probabilities respectively. Since these are typically unknown, one usually proceeds by first estimating these as $(\hat{f}_X, \hat{P})$. Typically, $\hat{f}_X$ is obtained by MLE based on a parametric form of $f_X(x'|a, x; \theta_f)$, while $\hat{P}$ is estimated nonparametrically using either a blocking scheme or kernel regression. Then, given $(\hat{f}_X, \hat{P})$, the values of $h(.)$ and $g(.)$ can be estimated by solving the recursive equation 2.1. This is done by first discretizing the state space, and then solving for $h(.)$, $g(.)$ in terms of $z(.)$, $e(.)$, using either backward induction or matrix inversion.

When the underlying state variables are really continuous, discretization effectively gives rise to a very high-dimensional state space, making estimation of $h(.)$, $g()$ computationally extremely expensive. To ameliorate this issue, Hotz et al. (1994) propose forward simulation based estimators for $h(.)$ and $g(.)$. Nevertheless, the computational requirements remain quite high, given that such a simulation estimate has to be carried out for every possible combination of a and x. Furthermore, the simulation errors create another source of bias in small samples. Additionally, given that all the common CCP-based methods require initial estimators of θ_f and

P, these procedures often suffer from heavy bias in small samples, as θ_f and P are estimated very imprecisely and enter non-linearly in the optimization problem for the structural parameters.

In the next section we propose an alternative algorithm for maximizing $Q(\theta)$ that allows for continuous states and does not require any knowledge about or estimation of $f_X(\cdot)$. In Section 4.2, we go further and show how we can also avoid the estimation of the choice probabilities.

Notation. We shall assume that the distribution of (a_{it}, x_{it}) is time stationary. This greatly simplifies our notation, but is not strictly necessary for many of our results; see Appendix A for extensions. Let $\mathbb{P}$ denote the stationary population (i.e, in the limit as $n \to \infty$) probability distribution of (a, x, a', x'), and $\mathbb{E}[\cdot]$ the corresponding expectation over $\mathbb{P}$. We shall also define $\mathbb{E}_n[\cdot]$ as the expectation over the empirical distribution $\mathbb{P}_n$ of (a, x, a', x'). In particular, we set $\mathbb{E}_n[f(a, x, a', x')] := (n(T-1))^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T-1} f(a_{it}, x_{it}, a_{it+1}, x_{it+1})$, i.e we always drop the last time period in the summation index even if $f(\cdot)$ does not depend on a', x'.

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

Finally, we use $|\cdot|$ to denote the usual Euclidean norm on a Euclidean space.

3. Temporal-difference estimation

This section presents our TD method for estimating $h(.)$ and $g(.)$. Let us first start with the $h(\cdot)$ function. Our method is based on a functional approximation for $h(.)$. To this end, we (approximately) parameterize this as

$$h^{(j)}(a, x) :\approx \phi(a, x)^{\intercal} \omega^{*(j)},$$

where $\phi(a, x)$ consists of a set of basis functions over the domain (a, x), and the superscript j represents the jth dimension of $h()$. Here, $\omega^{*(j)}$ denotes some approximation weights (more on this below). For the remainder of this paper, we shall drop the superscript j indexing the dimension of $h(.)$ and proceed as if the latter, and therefore θ^*, is a scalar. However, it should be taken as implicit that all our results hold for general $h(.)$, as long as each dimension is treated separately.[4] Also, to simplify the notation, we shall denote $\phi_{it} := \phi(a_{it}, x_{it})$ and $z_{it} := z(a_{it}, x_{it})$.

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

$$\delta(a, x; f) := z(a, x) + \beta \mathbb{E}\left[f(a', x')|a, x\right] - f(a, x),$$

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 $\Gamma_z[\cdot]$. However we want to approximate $h(a, x)$ with a function from the linear span, $\mathcal{L}_\phi$, of $\phi(a, x)$. The difficulty with the dynamic programming operator is that in general $\Gamma_z[f] \notin \mathcal{L}_\phi$ even if $f \in \mathcal{L}_\phi$. This suggests that to find a suitable approximation for $h(a, x)$ within $\mathcal{L}_\phi$, we should project the dynamic programming operator

[4] Even with component-wise estimation, the computational difficulty will be quite low since the most burdensome step, involving a matrix inversion, is common to all j.

back into this space. To do so, denote by P_ϕ the projection operator into the linear span of $\mathcal{L}_\phi$:

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

We then obtain our approximation $\phi(a,x)^\intercal\omega^*$ to $h(a,x)$ as the fixed point of the projected dynamic programming operator $P_\phi\Gamma_z[\cdot]$:

$$P_\phi\Gamma_z[\phi(a,x)^\intercal\omega^*] = \phi(a,x)^\intercal\omega^*.$$

In Lemma 1 in the Appendix, we show that this in turn is equivalent to

$$\mathbb{E}\left[\phi(a,x)\left\{z(a,x) + \beta\phi(a',x')^\intercal\omega^* - \phi(a,x)^\intercal\omega^*\right\}\right] = 0, \tag{3.1}$$

which enables us to identify ω^* as

$$\omega^* = \mathbb{E}\left[\phi(a,x)\left(\phi(a,x) - \beta\phi(a',x')\right)^\intercal\right]^{-1}\mathbb{E}\left[\phi(a,x)z(a,x)\right]. \tag{3.2}$$

Lemma 2 in the Appendix 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. We will discuss the properties of $\phi(a,x)^\intercal\omega^*$ in the next sub-section, but let us note here that in general

$$\phi(a,x)^\intercal\omega^* \neq P_\phi[h(a,x)].$$

Thus $\phi(a,x)^\intercal\omega^*$ is not the best linear approximation of $h(a,x)$, although it comes very close, as we will see shortly.

As defined above, ω^* cannot be computed directly, since it is a function of the true expectation $\mathbb{E}[\cdot]$. We can however obtain an estimator, $\hat{\omega}$, after replacing $\mathbb{E}[\cdot]$ with $\mathbb{E}_n[\cdot]$:

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

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

We now turn to the estimation of $g(\cdot)$. We approximate $g(\cdot)$ using basis functions $r(a,x)$:

$$g(a,x) :\approx r(a,x)^\intercal\xi^*.$$

We allow $r(a,x)$ to be generally different from $\phi(a,x)$. As before, denote by $f(a,x)$ any candidate function for $g(a,x)$. Now, $g(a,x)$ is the unique fixed point of the operator $\Gamma_e[\cdot]$, where

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

We then obtain our approximation $r(a,x)^\intercal\xi^*$ to $g(a,x)$ as the fixed point of the projected operator $P_r\Gamma_e[\cdot]$, where P_r is the projection operator into the linear span of r, i.e

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

As before, we may equivalently write

$$\mathbb{E}\left[r(a,x)\left\{\beta e(a',x') + \beta r(a',x')^\intercal\xi^* - r(a,x)^\intercal\xi^*\right\}\right] = 0. \tag{3.4}$$

This allows us to identify ξ^* as

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

Assuming $e(a,x)$ is known, this suggests the following estimator for ξ^*:

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

In general, $e(a,x) = \gamma - \ln P(a|x)$ is a function of choice probabilities, which are unknown. We first need to non-parametrically estimate them. Denote $\eta(a,x) := P(a|x)$. Suppose that we have access to a non-parametric estimator $\hat{\eta}$ of η. This can be obtained in many ways, e.g through series or kernel regression. We can then plug in this estimate to obtain $e(a,x;\hat{\eta}) := \gamma - \ln \hat{\eta}(a,x)$. This in turn enables us to obtain $\hat{\xi}$ 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 obtain an estimate of $g(\cdot)$ as $\hat{g}(a,x) = r(a,x)^\intercal \hat{\xi}$.

In the discrete setting, the estimation error from $\hat{\eta}$ is not first order relevant for the estimation of θ^*, as long as θ^* is estimated using a pseudo-MLE. This was first noted in Aguirregabiria and Mira (2002). In fact, even with continuous states, the estimation of $\hat{\xi}$ is unaffected to a first order by the estimation of $\hat{\eta}$, even though the latter only converges to the true η at non-parametric rates. This is because an orthogonality property holds for the estimation of ξ, in that

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

where $\partial_\eta \cdot$ denotes the Fréchet derivative with respect to η. To show (3.7), let us first expand the term $\mathbb{E}\left[\beta r(a,x) e(a',x';\eta)\right]$ as follows

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

where the second equality follows from the Markov property. Now, it turns out that

$$\partial_\eta \mathbb{E}\left[\ln \eta(a',x') \middle| x'\right] = 0.$$

Indeed, consider the expression $M(\tilde{\eta}) := \mathbb{E}\left[\ln \tilde{\eta}(a',x') \middle| x'\right]$, evaluated at different candidate values $\tilde{\eta}(\cdot,\cdot)$. When evaluated at the true conditional choice probability, i.e when $\tilde{\eta}(\cdot,\cdot) = \eta(\cdot,\cdot)$, $M(\tilde{\eta})$ becomes the conditional entropy and attains its maximum. Consequently, in view of (3.8), it follows that (3.7) holds. Thus, $\hat{\xi}$ is a locally robust estimator for ξ.

Even with a locally robust estimator, the use of a non-parametric estimator may lead to substantial finite sample bias. For this reason, we advocate a cross-fitting procedure (see Chernozhukov et al., 2018). In our context, 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.

Note that computation of $\hat{\omega}$ and $\hat{\xi}$ only involve solving linear equations of dimension $\dim(\phi)$ and $\dim(r)$, respectively. This is computationally very cheap. 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 pseudo-MLE estimator

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

It turns out the estimate from (3.9) is sub-optimal under continuous states. We discuss this in greater detail in Section 3.2, where we suggest a locally robust version of (3.9).

3.1. **Discrete states.** Suppose that the underlying states and actions are discrete, and that our algorithm uses basis functions comprised of the set of all discrete elements of x, a. Then the resulting estimate of $h(a, x)$ obtained from our algorithm is exactly the same as that obtained from the standard CCP estimators, if both the choice and transition probabilities were estimated using cell values. To see this, we note the following: First, the standard CCP estimators (see e.g Aguirregabiria and Mira, 2010), estimate $h(a, x)$ by solving the recursive equations

$$\breve{h}(a, x) = z(a, x) + \beta \sum_{x'} \hat{f}_X(x'|a, x) \sum_{a'} \hat{P}(a'|x')\breve{h}(a', x'), \tag{3.10}$$

where $\hat{f}, \hat{P}$ are estimates of f, P obtained as cell estimates. Second, by the results of Tsitsiklis and Van Roy (1997), it can be shown that when the functional approximation saturates all the states, the TD estimate from (3.3), denoted by $\hat{h}(x, a) := \phi(a, x)^{\intercal}\hat{\omega}$ satisfies the equation

$$z(a, x) + \beta \mathbb{E}_n[\hat{h}(a', x')|a, x] = \hat{h}(a, x),$$

where $\mathbb{E}_n[\hat{h}(a', x')|a, x]$ denotes the conditional expectation of $\hat{h}(a', x')$ given a and x under the empirical distribution $\mathbb{P}_n$ (the conditional distribution exists because of the discrete number of states). But for discrete data, $\mathbb{E}_n[\hat{h}(a', x')|a, x]$ is simply

$$\mathbb{E}_n[\hat{h}(a', x')|a, x] = \sum_{x'} \hat{f}_X(x'|a, x) \sum_{a'} \hat{P}(a'|x')\hat{h}(a', x'),$$

and the value of $\hat{h}(a, x)$ and $\breve{h}(a, x)$ coincide exactly. Thus, the two algorithms give identical results (a similar property also holds for $g(a, x)$). Since our estimates $\hat{h}(a, x)$ coincide with those from the standard CCP estimators, the resulting estimate $\hat{\theta}$ is also exactly the same. As a result, the final estimates of θ from both procedures also coincide exactly.

When the states are discrete, Aguirregabiria and Mira (2002) show that the estimation of η is orthogonal to the estimation of θ^*. This holds true for our procedure as well since our estimator is numerically equivalent to the one proposed by Aguirregabiria and Mira (2002). It is important to note, however, that the estimation of the transition probabilities $f_X(x'|a, x)$ is not orthogonal to the estimation of θ^*. This is not too much of an issue with discrete states since any estimate, $\hat{f}_X(x'|a, x)$, of $f_X(x'|a, x)$ converges at parametric rates, so $\sqrt{n}$ consistent estimation of θ is still possible. However, as we will see in Section 3.3, this creates issues once we move to continuous states.

3.2. **Theoretical Properties of TD estimators under continuous states.** We now characterize the formal properties of our TD fixed point estimates of $h(\cdot)$ and $g(\cdot)$. We shall only focus on the case of continuous states, since under discrete states, our prcedure gives exactly the same output as previous methods.

We start by characterizing the estimation error of $h(\cdot)$. Let k_ϕ denote the dimension of ϕ. We shall take $k_\phi \to \infty$ as $n \to \infty$. We impose the following assumptions for the estimation of $h(a,x)$:

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_ϕ.*

(ii) The basis functions are uniformly bounded, i.e $|\phi(a,x)|_\infty \le M$ for some $M < \infty$.

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

(iv) The domain of (a,x) is a compact set, and there exists $L < \infty$ such that $|z(a,x)|_\infty \le L$.

(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, we can set $\alpha = r/d$, 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 Appendix B, where we provide some primitive conditions on $z(a,x)$, $f_X(x'|a,x)$ that ensure existence of r continuous derivatives of $h(a,\cdot)$ for each $a \in \mathcal{A}$. Assumption 1(iv) requires the function $z(a,x)$ to be bounded. A sufficient condition for this is that $z(a,x)$ is continuous (since its domain is bounded).

Finally, Assumption 1(v) specifies the rate at which the dimension of the basis functions are allowed to grow. The rate requirements are also mild, and are the same as those employed for standard series estimation, even though our procedure is not the same as 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 $h(\cdot)$. But the choice of k_ϕ could matter in practice. For this reason we propose selecting k_ϕ 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. However in contrast to standard cross-validation, the performance is measured in terms of the empirical MSE of the TD error $\mathbb{E}_{\text{test}}[\delta^2(a,x;\hat{h})]$ on the test dataset. The value of k_ϕ that is chosen is the one that achieves the lowest mean squared TD error.

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

Theorem 1. *Under Assumptions 1(i) to 1(v), the following hold:*

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

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

(iii) There exists some $C < \infty$ such that with probability approaching one,

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

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

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

We prove Theorem 1 in the Appendix by adapting the results of Tsitsiklis and Van Roy (1997). The first part of Theorem 1 assures that both population and empirical TD fixed points exist. The second part of Theorem 1 implies the approximation bias from $\phi(a,x)^\intercal\omega^*$ is within a $(1-\beta)^{-1}$ factor of that from $P_\phi h(a,x)$. Note that the latter is the best one could do under an L_2 norm, so the theorem assures that we are only a constant away from attaining this. The third part of Theorem 1 characterizes the rate of convergence of $\hat{\omega}$ to ω^*, and the final part of Theorem 1 characterizes the rate of estimation of $h(a,x)$ itself.

In a similar vein, we impose the following assumptions for the estimation of $g(a,x)$. Let k_r denote the dimension of $r(a,x)$.

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.*

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

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

(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 $\eta(a,x) \geq \delta > 0$, where δ is independent of a,x. Additionally, $|\eta(a,x) - \hat{\eta}(a,x)|_\infty = o_p(1)$ and $\|\eta(a,x) - \hat{\eta}(a,x)\|_2^2 = o_p(n^{-1/2})$.

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.

Theorem 2. *Under Assumptions 2(i) to 2(vi), the following hold:*

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

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

(iii) There exists some $C < \infty$ such that with probability approaching one,

$$\left|\hat{\xi} - \xi^*\right| \leq C(1-\beta)^{-1}\sqrt{\frac{k_r}{n}}.$$

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

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

Theorem 1 and Theorem 2 imply that we can estimate $h(a, x)$ and $g(a, x)$ at reasonably fast rates. However we still need to discuss how this relates to consistent estimation of θ^*. We do this below.

3.3. **Continuous states and locally robust estimation.** When the states are continuous, estimation of $h(a, x)$ and $g(a, x)$ is inherently non-parametric. Unlike the case with discrete states, the estimation error from the non-parametric functions does affect the estimation of θ^* to a first order, when using the pseudo-MLE criterion. The reason for this is that $h(a, x)$ and $g(a, x)$ are actually functions of two non-parametric terms: the choice probabilities $\eta(a, x)$, and the transition probabilities $f_X(x'|a, x)$. The TD estimator implicitly takes both into account with a series approximation. Since the estimates for $f_X(x'|a, x)$ and θ^* are not orthogonal under a pseudo-MLE, this extends to the lack of orthogonality between the estimates for $h(a, x)$, $g(a, x)$ and θ^*. Consequently, the pseduo-MLE estimator will converge at slower than parametric rates.

3.3.1. *Construction of the locally robust estimator.* We now describe the construction of a locally robust version of the pseudo-MLE estimator. For the present analysis, let us suppose that $h(x, a)$ and $g(x, a)$ are finite-dimensional, i.e $h(x, a) \equiv \phi(x, a)^\intercal \omega^*$ and $g(x, a) \equiv r(x, a)^\intercal \xi^*$. Denote $(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}) := (a, x, a', x')$, $\mathbf{v} := (\omega, \xi)$, $\mathbf{v}^* := (\omega^*, \xi^*)$ and let

$$Q(a, x; \theta, \mathbf{v}) = \ln \pi_{\theta,\mathbf{v}}(a, x); \quad \pi_{\theta,\mathbf{v}}(a, x) := \frac{\exp\left\{(\phi(a, x)^\intercal \omega)\,\theta + r(a, x)^\intercal \xi\right\}}{\sum_{\breve{a}} \exp\left\{(\phi(\breve{a}, x)^\intercal \omega)\,\theta + r(\breve{a}, x)^\intercal \xi)\right\}}.$$

The true value θ^* is then

$$\theta^* = arg\max_\theta \mathbb{E}\left[Q(a, x; \theta, \mathbf{v}^*)\right].$$

Since the criterion function is convex, we can alternatively identify θ^* using the moment function

$$\mathbb{E}[m(a, x; \theta^*, \mathbf{v}^*)] = 0; \quad m(a, x; \theta, \mathbf{v}) := \partial_\theta Q(a, x; \theta, \mathbf{v}). \tag{3.11}$$

The lack of orthogonality of the estimator based on (3.11) is evident by the fact $\partial_{\mathbf{v}}\mathbb{E}[m(a, x; \theta, \mathbf{v}^*)] \neq 0$. Note that ω^* and ξ^* are in turn estimated using the moment functions

$$\mathbb{E}[\varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega^*)] = 0, \text{ and } \mathbb{E}[\varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi^*)] = 0, \tag{3.12}$$

where, given (3.1) and (3.4),

$$\varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega) := \phi(a, x)z(a, x) + \phi(a, x)\left(\beta\phi(a', x') - \phi(a, x)\right)^\intercal \omega, \text{ and}$$
$$\varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi) := \beta r(a, x)e(a', x'; \hat{\eta}) + r(a, x)\left(\beta r(a', x') - r(a, x)\right)^\intercal \xi.$$

We make use of (3.11) and (3.12) to construct a locally robust moment for θ^*. Following Newey (1994), Ackerberg et al. (2014) and Chernozhukov et al. (2018), this is given by

$$\mathbb{E}[\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta^*, \mathbf{v}^*)] = 0, \tag{3.13}$$

where

$$\zeta(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \mathbf{v}) := m(a, x; \theta, \mathbf{v}) - \mathbb{E}[\partial_\omega m(a, x; \theta, \mathbf{v})]\mathbb{E}[\partial_\omega \varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)]^{-1}\varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)$$
$$- \mathbb{E}[\partial_\xi m(a, x; \theta, \mathbf{v})]\mathbb{E}[\partial_\xi \varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi)]^{-1}\varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi).$$

Note that

$$\mathbb{E}[\partial_\omega \varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)] = \mathbb{E}\left[\phi(a,x)\left(\beta\phi(a',x') - \phi(a,x)\right)^\intercal\right], \text{ and}$$
$$\mathbb{E}[\partial_\xi \varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi)] = \mathbb{E}\left[r(a,x)\left(\beta r(a',x') - r(a,x)\right)^\intercal\right].$$

We can now construct a locally robust estimator for θ^* based on (3.13). Following Chernozhukov et al. (2018), we employ a cross-fitting procedure by randomly splitting the data into two samples $\mathcal{N}_1$ and $\mathcal{N}_2$. We compute $\hat{\omega}$ and $\hat{\xi}$ using one of the samples, say $\mathcal{N}_2$. Denote by $\mathbb{E}_n^{(1)}[\cdot]$ the empirical expectation using only the observations in the first sample. We then obtain $\hat{\theta}$ as the solution to the moment equation

$$\mathbb{E}_n^{(1)}\left[\zeta_n(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \hat{\omega}, \hat{\xi})\right] = 0, \tag{3.14}$$

where

$$\zeta_n(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}; \theta, \mathbf{v}) := m(a,x;\theta,\mathbf{v}) - \mathbb{E}_n^{(1)}[\partial_\omega m(a,x;\theta,\mathbf{v})]\mathbb{E}_n^{(1)}[\partial_\omega \varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)]^{-1}\varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)$$
$$- \mathbb{E}_n^{(1)}[\partial_\xi m(a,x;\theta,\mathbf{v})]\mathbb{E}_n^{(1)}[\partial_\xi \varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi)]^{-1}\varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \xi).$$

The use of cross-fitting or sample splitting is critical. If we had used the entire sample to estimate all of θ^*, ω^* and ξ^*, we would have $\mathbb{E}_n[\varphi_g(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \hat{\omega})] = 0$ for $g \in \{h, e, \eta\}$, which implies $\mathbb{E}_n\left[\zeta_n(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \theta, \hat{\omega}, \hat{\xi})\right] = \mathbb{E}_n\left[m(a,x,\theta,\hat{\omega},\hat{\xi})\right]$. As noted by Chernozhukov et al. (2018), cross-fitting gets rid of the 'own observation bias' that is the source of the degeneracy here.

We will refer to the solution $\hat{\theta}$ of (3.14) as the locally robust pseduo-MLE estimator of θ^*. Note that we would need three way sample splits if we employ cross-fitting procedures for both estimation of θ^* and ξ^*. But the use of cross-fitting for $\hat{\xi}$ is not as critical as that for $\hat{\theta}$, and can be avoided if necessary.

Estimation of $\hat{\theta}$ using (3.14) involves non-convex optimization. Since this could cause difficulties in practice, we recommend a two-step method for computation. We first obtain a preliminary estimate $\hat{\theta}_1$ by solving the empirical analogue of (3.11). This is a convex optimization problem, and is usually very fast. Note that $\hat{\theta}_1$ is consistent for θ under mild regularity conditions, even though its not efficient. We can then use $\hat{\theta}_1$ as the starting point for a Newton-Raphson or some other gradient descent algorithm for finding the root of (3.14).

3.3.2. *Non-parametric analysis.* For the setup of finite-dimensional $h(a,x)$ and $g(a,x)$, it is straightforward to show that the above procedure leads to $\sqrt{n}$ rates of estimation of θ^* (see e.g Newey (1994)). In this paper, we are primarily interested in the case where these quantities are infinite-dimensional. Still, treating the first step as parametric leads to an estimation strategy that is also valid non-parametrically as long as we let the series terms grow to infinity. To show this, we will need to derive the exact form of the adjustment terms in the non-parametric case. To this end, we will make use of the form of the parametric adjustment terms in (3.14) to conjecture the expression for the non-parametric correction term. We shall then verify that this indeed leads to a locally robust estimator.

With the above in mind, consider the adjustment term

$$\hat{\mathcal{A}}_h := \mathbb{E}_n^{(1)}[\partial_\omega m(a,x;\theta,\mathbf{v})]\mathbb{E}_n^{(1)}[\partial_\omega \varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)]^{-1}\varphi_h(\tilde{\mathbf{a}}, \tilde{\mathbf{x}}, \omega)$$

for $h(a,x)$. Denote

$$m(a,x;\theta,h,g) := \partial_\theta Q(a,x;\theta,h,g); \quad Q(a,x;\theta,h,g) := \ln \frac{\exp\{h(a,x)\theta + g(a,x)\}}{\sum_{\breve{a}} \exp\{h(\breve{a},x)\theta + g(\breve{a},x))\}}.$$

Then $\hat{\mathcal{A}}_h$ can be rewritten as

$$\hat{\mathcal{A}}_h = \hat{\lambda}_h(a,x;\theta)\left\{z(a,x) + \beta\phi(a',x')^\intercal\omega - \phi(a,x)^\intercal\omega\right\}, \tag{3.15}$$

where

$$\hat{\lambda}_h(a,x;\theta) := \phi(a,x)^\intercal \mathbb{E}_n^{(1)}\left[(\beta\phi(a',x') - \phi(a,x))\,\phi(a,x)^\intercal\right]^{-1}\mathbb{E}_n^{(1)}\left[\phi(a,x)\partial_h m(a,x;\theta,h,g)\right],$$

and $\partial_h m(\cdot)$ denotes the Fréchet derivative of $m(\cdot)$ with respect to $h(\cdot)$. In Appendix B, we provide a heuristic argument that suggests that the limit, $\mathcal{A}_h$, of $\hat{\mathcal{A}}_h$ as $n, k_\phi \to \infty$ is given by

$$\mathcal{A}_h = \lambda_h(a,x;\theta)\left\{z(a,x) + \beta h(a',x') - h(a,x)\right\},$$

where $\lambda_h(a,x;\theta)$ is the fixed point of the 'backward' dynamic programming operator $\Gamma^\dagger_{h,\theta}[\cdot]$, defined as[5]

$$\Gamma^\dagger_{h,\theta}[f](a,x) := -\partial_h m(a,x;\theta,h,g) + \beta\mathbb{E}\left[f(a^{-\prime},x^{-\prime})|a,x\right].$$

We conjecture $\mathcal{A}_h$ to be the non-parametric correction term for $h(\cdot)$. A similar analysis also applies to the adjustment term for $g(\cdot)$, which we conjecture to be of the form

$$\mathcal{A}_g = \lambda_g(a,x;\theta)\left\{e(a',x';\eta) + \beta g(a',x') - g(a,x)\right\},$$

where $\lambda_g(a,x;\theta)$ is the fixed point of the operator $\Gamma^\dagger_{g,\theta}[\cdot]$, defined as

$$\Gamma^\dagger_{g,\theta}[f](a,x) := -\partial_g m(a,x;\theta,h,g) + \beta\mathbb{E}\left[f(a^{-\prime},x^{-\prime})|a,x\right].$$

Taken together, we conjecture that the locally robust moment is given by

$$\begin{aligned}\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g) := m(a,x;\theta,h,g) &- \lambda_h(a,x;\theta)\left\{z(a,x) + \beta h(a',x') - h(a,x)\right\} \\ &- \lambda_g(a,x;\theta)\left\{e(a',x';\eta) + \beta g(a',x') - g(a,x)\right\}.\end{aligned} \tag{3.16}$$

The above analysis is heuristic. We now verify that the moment in (3.16) is indeed locally robust. A necessary condition for this is that $\partial_h\mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)] = 0$ and $\partial_g\mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)] = 0$, where the derivatives are Gâteaux derivatives with respect to $h(\cdot)$ and $g(\cdot)$ respectively (see Chernozhukov et al. (2018)). To verify these, observe that for any square integrable γ,

$$\begin{aligned}\partial_\tau\mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h+\tau\gamma,g)] = \mathbb{E}[\partial_h m(a,x;\theta,h,g)\gamma(a,x)] &- \mathbb{E}[\beta\lambda_h(a,x;\theta)\gamma(a',x')] \\ &+ \mathbb{E}[\lambda_h(a,x;\theta)\gamma(a,x)].\end{aligned} \tag{3.17}$$

Since $\lambda_h(\cdot)$ is the fixed point of $\Gamma^\dagger_{h,\theta}[\cdot]$, we can expand the third term in (3.17) as

$$\begin{aligned}\mathbb{E}[\lambda_h(a,x;\theta)\gamma(a,x)] &= \mathbb{E}\left[\left\{-\partial_h m(a,x;\theta,h,g) + \beta\lambda_h(a^{-\prime},x^{-\prime};\theta)\right\}\gamma(a,x)\right] \\ &= -\mathbb{E}\left[\partial_h m(a,x;\theta,h,g)\gamma(a,x)\right] + \mathbb{E}[\beta\lambda_h(a,x;\theta)\gamma(a',x')],\end{aligned}$$

[5] In other words, $\lambda_h^*(a_{it},x_{it};\theta) = -\sum_{j=0}^{\infty}\beta^j\partial_h m(a_{i(t-j)},x_{i(t-j)};\theta,h,g)$, i.e it is like a 'backward' value function.

where the second equality uses the fact that $\mathbb{E}[\cdot]$ is a stationary distribution. We thus conclude $\partial_\tau \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h+\tau\gamma,g)]=0$ for all γ, or $\partial_h \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)]=0$, as required. In fact, by similar arguments, we can also show the stronger statement that $\partial_h \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)]=0$ and $\partial_g \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)]=0$ in a Fréchet sense. Additionally, the Fréchet second derivatives $\partial_h^2 \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)]$, $\partial_g^2 \mathbb{E}[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)]$ also exist, and are uniformly bounded for bounded θ.

The locally robust moment (3.16) is infeasible since $\lambda_g(\cdot),\lambda_h(\cdot),h(\cdot)$ and $g(\cdot)$ are unknown. However, in practice we can simply use the estimator from (3.14). Note that the moment function from the latter can be rewritten as

$$\zeta_n(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,\mathbf{v}) = m(a,x;\theta,\mathbf{v}) - \hat{\lambda}_h(a,x;\theta)\left\{z(a,x)+\beta\phi(a',x')^\intercal\omega-\phi(a,x)^\intercal\omega\right\}$$
$$\hat{\lambda}_g(a,x;\theta)\left\{e(a',x';\hat{\eta})+\beta r(a',x')^\intercal\xi - r(a,x)^\intercal\xi\right\}.$$

There is no loss of first order efficiency in replacing $\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)$ with $\zeta_n(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,\mathbf{v})$. This is because, by a similar analysis as for Theorems 1, 2, it can be shown that

$$\sup_\theta \left\|\hat{\lambda}_h(a,x;\theta)-\lambda_h(a,x;\theta)\right\|_2 = O_p\left(\frac{k_\phi}{\sqrt{n}}+k_\phi^{-\alpha_1}\right)=o_p(n^{-1/4}),\ \text{and}$$
$$\sup_\theta \left\|\hat{\lambda}_g(a,x;\theta)-\lambda_g(a,x;\theta)\right\|_2 = O_p\left(\frac{k_r}{\sqrt{n}}+k_r^{-\alpha_2}\right)=o_p(n^{-1/4}),$$

for suitable (k_r,k_ϕ), where α_1,α_2 depend on the smoothness classes of $\partial_h m(a,\cdot;\theta,h,g)$, $\partial_g m(a,\cdot;\theta,h,g)$. Note also that $\phi(a,x)^\intercal\omega$ and $r(a,x)^\intercal\xi$ are L_2 consistent for $h(a,x)$ and $g(a,x)$, respectively, at faster than $n^{-1/4}$ rates. Following the analysis of Chernozhukov et al. (2018), these facts imply that the estimator based on (3.13) has the same limiting distribution as the one based on (3.16). In particular, it achieves parametric rates of convergence. We state the regularity conditions and the theorem below (for the remainder of this section we allow θ^* to be vector valued):

Assumption 3. *(i) $\theta^*\in\Theta$, a compact set, and $\mathbb{E}\left[\zeta(\tilde{\mathbf{a}},\tilde{\mathbf{x}};\theta,h,g)\right]=0 \iff \theta=\theta^*$.*

(ii) $\partial_g m(a,x;\theta,h,g)$ and $\partial_h m(a,x;\theta,h,g)$ are uniformly bounded for all (a,x,θ,h,g).

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

(iv) $n^{-1/2}k_\phi+k_\phi^{-\min\{\alpha,\alpha_1\}}=o_p(n^{-1/4})$ and $n^{-1/2}k_r+k_r^{-\min\{\alpha,\alpha_2\}}=o_p(n^{-1/4})$.

Theorem 3. *Suppose that Assumptions 1 - 3 hold. Then the estimator, $\hat{\theta}$ of θ^*, based on (3.13) is $\sqrt{n}$ consistent, and satisfies*

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

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

The proof of the above theorem follows by verifying the regularity conditions of Chernozkhov et al. (2018, Theorem 16). Since these are more or less straightforward to verify given our previous results, we omit the details.

For inference on $\hat{\theta}$, the covariance matrix V can be estimated as

$$\hat{V}=\left(\hat{G}^\intercal\hat{\Omega}^{-1}\hat{G}\right)^{-1},$$

where

$$\hat{G} = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \frac{\partial \zeta_n(a_{it}, x_{it}, a_{it+1}, x_{it+1}; \hat{\theta}, \hat{\omega}, \hat{\xi})}{\partial \theta^\intercal}, \quad \text{and}$$

$$\hat{\Omega} = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \zeta_n(a_{it}, x_{it}, a_{it+1}, x_{it+1}; \hat{\theta}, \hat{\omega}, \hat{\xi}) \zeta_n(a_{it}, x_{it}, a_{it+1}, x_{it+1}; \hat{\theta}, \hat{\omega}, \hat{\xi})^\intercal.$$

Chernozhukov et al. (2018) provide conditions under which $\hat{V}$ is consistent for V; these are straightforward to verify in our context. Alternatively, one could employ the bootstrap, which remains valid in this context.

4. Extensions

4.1. Stochastic Gradient descent. Computation of $\hat{\omega}$ and $\hat{\xi}$ involves inverting a $(k \times k)$-dimensional matrix. Once k becomes very large, matrix inversion does start to become more demanding. In such cases stochastic gradient descent is a computationally cheap alternative. In particular, we can estimate ω^* in (3.3) using stochastic gradient updates of the form

$$\hat{\omega}^{new} \longleftarrow \hat{\omega}^{old} + \alpha_\omega \left(z_{it} + \beta \phi_{it+1}^\intercal \hat{\omega}^{old} - \phi_{it}^\intercal \hat{\omega}^{old} \right) \phi_{it}, \tag{4.1}$$

where each observation $(z_{it}, \phi_{it}, \phi_{it+1})$ is drawn at random from $\mathbb{P}_n$ i.e, with replacement from the set of all the sample observations. Here α_ω is the learning rate for stochastic gradient descent. In a similar vein we can estimate ξ^* using gradient updates of the form

$$\hat{\xi}^{new} \longleftarrow \hat{\xi}^{old} + \alpha_\xi \left(\beta e_{it+1}(\hat{\eta}) + \beta r_{it+1}^\intercal \hat{\xi}^{old} - r_{it}^\intercal \hat{\xi}^{old} \right) r_{it}, \tag{4.2}$$

where α_ξ is the learning rate for ξ, and $e_{it+1}(\hat{\eta}) := \gamma - \ln \hat{\eta}(a_{it+1}, x_{it+1})$. Estimation of (ω^*, ξ^*) using the gradient updates (4.1) and (4.2) is termed TD learning in the Reinforcement Learning literature. Pseudo-code for our TD learning algorithm is provided in Algorithm 1.

We shall require the following assumption on the learning rates:

Assumption 4. *The learning rates satisfy* $\sum_l \alpha_\omega^{(l)2} \to 0$, $\sum_l \alpha_\xi^{(l)2} \to 0$ *and* $\sum_l \alpha_\omega^{(l)} \to \infty$, $\sum_l \alpha_\xi^{(l)} \to \infty$ *as the number of steps in the algorithm goes to infinity, where* $\alpha_\omega^{(l)}, \alpha_\xi^{(l)}$ *denote the learning rates after* l *steps/updates of the algorithm.*

Assumption 4 is a standard condition on learning rates for stochastic gradient descent algorithms. We can now prove the following theorem on convergence:

Theorem 4. *Suppose that Assumptions 1, 2 and 4 hold. Then, with probability approaching one, the sequence of updates* ω_l *and* ξ_l *converge to* $\hat{\omega}, \hat{\xi}$ *as* $l \to \infty$.

The TD learning algorithm can also be parallelized by running multiple stochastic gradient threads in parallel and using Hogwild!-style asynchronous updates (Niu et al., 2011). Each thread runs parallel instances of the same code with a delayed time start, and independently and asynchronously updates a global parameter that returns ω. This speeds up computation by the order of magnitude of the number of parallel threads.

Algorithm 1 TD learning algorithm for CCP estimation

Initialize all parameters to arbitrary values

Repeat:

Choose $(x_{it}, a_{it}, x_{it+1}, a_{it+1})$ at random, with replacement, from sample data

Calculate the values of $(\phi_{it}, z_{it}, r_{it}, \phi_{it+1}, r_{it+1}, e_{it+1}(\hat{\eta}))$

$\hat{\omega} \longleftarrow \hat{\omega} + \alpha_\omega \left(z_{it} + \beta\phi_{it+1}^\intercal \hat{\omega} - \phi_{it}^\intercal \hat{\omega} \right) \phi_{it}$

$\hat{\xi} \longleftarrow \hat{\xi} + \alpha_\xi \left(\beta e_{it+1}(\hat{\eta}) + \beta r_{it+1}^\intercal \hat{\xi} - r_{it}^\intercal \hat{\xi} \right) r_{it}$

Until: Convergence criteria for $(\hat{\omega}, \hat{\xi})$ are reached

4.2. **Recursive estimation.** So far, we have required knowledge of some initial estimates of the choice probabilities to obtain the values of $e(a', x')$. This is so even as we do eliminate entirely the need for any initial probability values when estimating $h(.)$, as well as the need to estimatef_X. In this section we show how the estimate for η can also be dispensed with, at the expense of a bit more computation. The key insight we exploit is the fact that at the true value θ^* of θ, we will have

$$\eta(a, x) = \frac{\exp\{h(a, x)\theta^* + g(a, x)\}}{\sum_a \exp\{h(a, x)\theta^* + g(a, x)\}}.$$

Thus, if we have a consistent estimator for θ^*, we can use this to obtain an estimate for $\eta(a, x)$. This suggests a recursive procedure for estimating $\eta(\cdot)$ and θ simultaneously.

Note that, even with this recursive procedure, the estimates $\hat{\omega}$ can be obtained directly from (3.3). We do not require any estimate of $\eta(a, x)$ for this. Let $\hat{h}(a, x)$ denote the estimate of $h(a, x)$ that we obtained in the previous sections. We start the recursive procedure by initializing ξ and θ to arbitrary values. Additionally, we also initialize $\eta(a, x)$ by $\hat{\eta}_{(1)}(a, x)$, where the latter is some preliminary estimate of the choice probabilities. Let $\hat{\xi}_{(k)}$ and $\hat{\theta}_{(k)}$ denote the parameter estimates, at the k-th iteration of the procedure. Similarly, let $\hat{\eta}_{(k)}(a, x)$ and $\hat{e}_{(k)}(a, x)$ denote the estimates of $\eta(\cdot)$ and $e(\cdot)$ after k iterations of the procedure. These quantities are then updated as follows: We first update $\hat{\eta}(\cdot)$ as

$$\hat{\eta}_{(k+1)}(a, x) = \frac{\exp\left\{\hat{h}(a, x)\hat{\theta}_{(k)} + r(a, x)^\intercal \hat{\xi}_{(k)}\right\}}{\sum_{\dot{a}} \exp\left\{\hat{h}(\dot{a}, x)\hat{\theta}_{(k)} + r(\dot{a}, x)^\intercal \hat{\xi}_{(k)}\right\}}. \tag{4.3}$$

This enables us to obtain a new estimate of $e(a, x)$,

$$\hat{e}_{(k+1)}(a, x) := \gamma - \ln \hat{\eta}_{(k+1)}(a, x). \tag{4.4}$$

Following this, $\hat{\xi}$ can be updated as

$$\hat{\xi}_{(k+1)} = \mathbb{E}_n \left[r(a, x) \left(r(a, x) - \beta r(a', x') \right)^\intercal \right]^{-1} \mathbb{E}_n \left[\beta r(a, x) \hat{e}_{(k+1)}(a', x') \right]. \tag{4.5}$$

Finally, $\hat{\theta}$ can be updated as

$$\hat{\theta}_{(k+1)} = \arg\max_\theta \sum_{i=1}^{n} \sum_{t=1}^{T-1} \log \frac{\exp\left\{\hat{h}(a_{it}, x_{it})\theta + r(a_{it}, x_{it})^\intercal \hat{\xi}_{(k+1)}\right\}}{\sum_a \exp\left\{\hat{h}(a, x_{it})\theta + r(a, x_{it})^\intercal \hat{\xi}_{(k+1)}\right\}}. \tag{4.6}$$

The above update does not employ the locally robust correction to obtain $\hat{\theta}$. This can be easily rectified using (3.14); we refer to the previous section for the details. We iterate between steps (4.3) - (4.6) until the parameters converge.

Our recursive procedure is very similar to, and influenced by, the NPL algorithm of Aguirregabiria and Mira (2002). Using Monte Carlo simulations, the authors show that the recursive procedure enjoys smaller finite sample bias and variance. This was subsequently proven using higher order expansions by Kasahara and Shimotsu (2008). We similarly expect our recursive procedure to have better finite sample properties.

4.3. **Nonlinear utility functions.** So far we have focused on the case where the utility function is linear in parameters θ^*. This is the most common setup in practice as it simplifies computation considerably. However, in some situations it may be useful to specify the observed utility component to be nonlinear in θ^*. Denote this by $z(a, x; \theta^*)$. We can then estimate θ^* as the maximizer of the pseudo-likelihood criterion

$$Q(\theta) = \sum_{i=1}^{n} \sum_{t=1}^{T-1} \log \frac{\exp\{h(a_{it}, x_{it}; \theta) + g(a_{it}, x_{it})\}}{\sum_a \exp\{h(a, x_{it}; \theta) + g(a, x_{it})\}},$$

where, for each θ, $h(.; \theta)$ and $g(.)$ solve the following recursive expressions:

$$h(a, x; \theta) = z(a, x; \theta) + \beta\mathbb{E}\left[h(a', x'; \theta)|a, x\right], \text{ and}$$
$$g(a, x) = \beta\mathbb{E}\left[e(a', x') + g(a', x')|a, x\right].$$

We can use our TD estimation procedure to obtain a functional approximation $\hat{h}(a, x; \theta)$ for $h(a, x; \theta)$, conditional on each different value of θ. As argued earlier, this step can be computed very fast. Even more appealingly, the term $\mathbb{E}_n\left[\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^\intercal\right]$ employed in the TD estimate (3.3) does not feature $z(a, x; \theta)$, and therefore only has to be inverted once. As long as Assumption 1 holds uniformly over all θ, we can also prove that $\hat{h}(a, x; \theta)$ is uniformly consistent for $h(a, x; \theta)$ at the same rates as before i.e.

$$\sup_{\theta\in\Theta} \left\| h(a, x; \theta) - \hat{h}(a, x; \theta) \right\|_2 = O_p\left(\frac{k_\phi}{\sqrt{n}} + k_\phi^{-\alpha}\right).$$

We can therefore plug in the values of $\hat{h}(.; \theta)$ and $\hat{g}(\cdot)$ to estimate θ^* as

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

Computing $\hat{\theta}$ is now more involved. We recommend employing a derivative-free optimization procedure such as Nelder-Mead.

As before, when the state space is discrete, the above estimator reduces to standard CCP estimation using cell probabilities, as described in Aguirregabiria and Mira (2010). It is also straightforward to make the objective function locally robust to estimation of $\hat{g}(\cdot)$ as well, following the construction in (3.3). However, due to the estimation of $\hat{h}(\cdot; \theta)$, it is not known (to us) if the resulting estimator achieves parametric rates of convergence under continuous states.

5. Incorporating permanent unobserved heterogeneity

In this section, we show how we can model permanent unobserved heterogeneity by pairing the techniques from Section 3 with the sequential Expectation-Maximization (EM) algorithm (Arcidiacono and Jones, 2003). The use of the sequential EM algorithm in CCP estimation under unobserved heterogeneity was first advocated by Arcidiacono and Miller (2011), and we employ a similar approach.

Suppose that in addition to the observed state x, and the choice specific shock e, individuals also base their choice decisions on a random state variable s which is known to the individual, but unobserved to the econometrician. As is common in the literature, we assume a finite set of unobserved states indexed by $\{1, 2, \dots, k, ...K\}$. The number of states is also assumed to be known a priori. Let π_k denote the population probability $P(s = k)$. The value of s for an individual is assumed to be permanent and not change with time. However, we do not place any restrictions on the transition density $f_X(x'|a, x, s)$, which is allowed to change with s.

To simplify the exposition, we will only employ the basic version of the algorithm without locally robustness corrections as in Section 3.3. It is straightforward to incorporate the correction term into the algorithm, but it comes at the expense of higher computational times.

Suppose that the per-period utility is given by $z(a, x, s)\theta$. For each k, define $h_k(a, x)$ and $g_k(a, x)$ as the solutions to

$$h_k(a, x) = z(a, x, k) + \beta\mathbb{E}\left[h_k(a', x')|a, x, s = k\right],$$
$$g_k(a, x) = \beta\mathbb{E}\left[e(a', x') + g_k(a', x')|a, x, s = k\right].$$

To simplify notation, let $h_{itk} := h_k(a_{it}, x_{it})$ and $g_{itk} := g_k(a_{it}, x_{it})$. If these quantities were known, one can estimate (θ, π) by maximizing the integrated pseudo-likelihood

$$Q(\theta, \pi) = \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T-1} \frac{\exp\{h_{itk}\theta + g_{itk}\}}{\sum_a \exp\{h_k(a, x_{it})\theta + g_k(a, x_{it})\}}\right]. \tag{5.1}$$

In reality, of course, $h_k(a, x), g_k(a, x)$ would have to be estimated. To this end, we choose a set of basis functions $\phi(a, x)$ and $r(a, x)$ over the domain of (a, x), and for each k we approximately parameterize

$$h_k(a, x) :\approx \phi(a, x)^\intercal\omega_k^*; \quad g_k(a, x) :\approx r(a, x)^\intercal\xi_k^*.$$

As before, we have chosen to make $h()$ uni-dimensional to simplify the notation. The extension to multiple dimensions is straightforward as one simply treats each dimension separately.

Maximizing (5.1) is not equivalent to Full Information Maximum Likelihood (FIML). As in Arcidiacono and Jones (2003), the identification and asymptotic properties of θ, π are in fact determined by constructing moment conditions that correspond to the first order conditions from maximizing $Q(\theta, \pi)$, augmented with additional moments identifying ω_k^*, ξ_k^* (see below). Together, these moment conditions, which motivate the sequential EM algorithm, can in turn be related to the identification properties of FIML. We refer the reader to Appendix B for the details of this construction.

Similar to Section 3, for each $k = 1, \ldots, K$, ω_k^* is identified as (see Appendix B for details)

$$\omega_k^* = \bar{\mathbb{E}}\left[\mathbb{I}(s=k)\phi(a,x)\left(\phi(a,x) - \beta\phi(a',x')\right)^\intercal\right]^{-1} \bar{\mathbb{E}}\left[\mathbb{I}(s=k)\phi(a,x)z(a,x,k)\right], \tag{5.2}$$

where $\bar{\mathbb{E}}[\cdot]$ differs from $\mathbb{E}[\cdot]$ in also taking the expectation over the distribution of the unobserved state s. In particular, observe that,

$$\bar{\mathbb{E}}\left[\mathbb{I}(s=k)\phi(a,x)z(a,x,k)\right] = \mathbb{E}\left[P(s=k|\mathbf{a},\mathbf{x})\phi(a,x)z(a,x,k)\right],$$

where

$$P(s=k|\mathbf{a},\mathbf{x}) := Pr(s=k|a_1,x_1,\ldots,a_T,x_T).$$

In a similar vein, we also have

$$\bar{\mathbb{E}}\left[\mathbb{I}(s=k)\phi(a,x)\left(\phi(a,x) - \beta\phi(a',x')\right)^\intercal\right] = \mathbb{E}\left[P(s=k|\mathbf{a},\mathbf{x})\phi(a,x)\left(\phi(a,x) - \beta\phi(a',x')\right)^\intercal\right].$$

Denote by $p_{ik} = P(s=k|\mathbf{a}_i,\mathbf{x}_i)$ the probability of being in state k conditional on the realized set of all actions $\mathbf{a}_i$ and observed states $\mathbf{x}_i$ for individual i. Also, let $e_k(a,x) = \gamma - \ln P(a|x, s = k)$, $z_{itk} := z(a_{it}, x_{it}, k)$, $\phi_{it} := \phi(a_{it}, x_{it})$, $e_{itk} := e_k(a_{it}, x_{it})$, and $r_{it} := r(a_{it}, x_{it})$. Then replacing the expectation $\mathbb{E}[.]$ in the previously displayed equations with the sample expectation $\mathbb{E}_n[.]$, we obtain the estimates

$$\hat{\omega}_k = \left[\sum_{i=1}^{n}\sum_{t=1}^{T-1} p_{ik}\phi_{it}\left(\phi_{it} - \beta\phi_{it+1}\right)^\intercal\right]^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T-1} p_{ik}\phi_{it}z_{itk} \tag{5.3}$$

A similar expression also holds for updates to ξ_k:

$$\hat{\xi}_k = \left[\sum_{i=1}^{n}\sum_{t=1}^{T-1} p_{ik}r_{it}\left(r_{it} - \beta r_{it+1}\right)^\intercal\right]^{-1} \sum_{i=1}^{n}\sum_{t=1}^{T-1} \beta p_{ik}r_{it}\dot{e}_{it+1k}, \tag{5.4}$$

where $\dot{e}_{it+1k}$ is the current estimate of e_{it+1k}.

Estimation of $\boldsymbol{\omega}^* := (\omega_1^*, \ldots, \omega_K^*)$, $\boldsymbol{\xi}^* := (\xi_1^*, \ldots, \xi_K^*)$ and θ^* using equations (5.1), (5.3) and (5.4) requires knowledge of the unknown quantities π_k and p_{ik} along with $\dot{e}_{it+1k}$. Furthermore, even if π_k were known, maximizing the integrated likelihood function (5.1) is computationally very expensive. The sequential EM algorithm of Arcidiacono and Jones (2003) solves both issues and provides a computationally cheap alternative to maximizing (5.1). To describe the procedure, let

$$l_{itk}(\theta, \boldsymbol{\omega}, \boldsymbol{\xi}) \equiv \frac{\exp\left\{(\phi_{it}^\intercal\omega_k)\theta + (r_{it}^\intercal\xi_k)\right\}}{\sum_a \exp\left\{(\phi(a,x_{it})^\intercal\omega_k)\theta + r(a,x_{it})^\intercal\xi_k\right\}}.$$

Denote by $\bar{\pi}_k$ and $\hat{p}_{ik}$ the estimates for π_k and p_{ik}. The algorithm consists of two steps: the M-step and the E-step. We first describe the M-step. Here, we update the estimates for ω^*, ξ^* and θ^* based on the current estimates for π_k, p_{ik} and e_{it+1k}. To this end, first note that we can update $\hat{\boldsymbol{\omega}} := (\hat{\omega}_1, \ldots, \hat{\omega}_K)$ and $\hat{\boldsymbol{\xi}} := (\hat{\xi}_1, \ldots, \hat{\xi}_K)$ using (5.3) and (5.4). From these we can in-turn update $\hat{\theta}$ as

$$\hat{\theta} = \arg\max_{\theta}\left[\sum_{i=1}^{n}\sum_{t=1}^{T-1}\sum_{k} p_{ik}\ln l_{itk}\left(\theta, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}}\right)\right]. \tag{5.5}$$

Next, given $\hat{\theta}, \hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\xi}}$, we update $\bar{\pi}_k, \hat{p}_{ik}$ and $\dot{e}_{it+1k}$ for all i, k. This is the E-step of the EM algorithm. This step consists of three parts. In the first part, we use the current $\hat{\theta}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}}$ and $\bar{\pi}_k$

to update $\hat{p}_{ik}$ for each i, k using Bayes' rule:

$$\hat{p}_{ik} \longleftarrow \frac{\hat{\pi}_k \prod_{t=1}^{T-1} l_{itk}(\hat{\theta}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}})}{\sum_{\tilde{k}} \hat{\pi}_{\tilde{k}} \prod_{t=1}^{T-1} l_{it\tilde{k}}(\hat{\theta}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}})}. \tag{5.6}$$

In the second part, we update $\hat{\pi}_k$, for each k, as

$$\hat{\pi}_k \longleftarrow \frac{1}{N} \sum_{i=1}^{N} \hat{p}_{ik}. \tag{5.7}$$

Finally, we also update $\dot{e}_{it+1k}$ for all i, t, k as

$$\dot{e}_{it+1k} \longleftarrow \gamma - \ln l_{it+1k}(\hat{\theta}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}}). \tag{5.8}$$

The E and M steps are iterated until convergence. Any EM based algorithm can only be guaranteed to find a local solution, so it is important in practice to initialize the above procedure with multiple random values of $\theta, \boldsymbol{\omega}, \boldsymbol{\xi}$.

The computational requirements for the EM algorithm are higher due to the iteration between the expectation and maximization steps. However the maximization step is still very fast as we can estimate all the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\xi}$ through a low-dimensional matrix inversion, while computing $\hat{\theta}$ just requires solving a convex optimization problem.

It is also possible to extend our methods to allow for Markovian unobserved heterogeneity, by employing a variant of the classical Baum-Welch algorithm. The computational and statistical details of such a procedure are however more involved and will be described elsewhere.

6. Estimation of dynamic discrete games

So far we have considered applications of our algorithm to single agent models, where we have argued that there are substantial computational and statistical gains from using our procedure. These gains are magnified when extended to estimation of dynamic discrete games.

Our setup is based on Aguirregabiria and Mira (2010). We assume a single Markov-Perfect-Equilibrium setup where multiple players $i = 1, 2, \dots, N$ play against each other in M different markets. We observe the state of play for T time-periods, where both T and the number of players N are assumed 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 $z_i(a_i, a_{-i}, x)^\intercal \theta^*$ for each player i, for some finite dimensional parameter θ^*, where a_i denotes player i's action and a_{-i} denotes the actions of all other players . Evolution of the states in the next period is determined by the transition probability $f_X(x'|a, x)$ where $\mathbf{a} := (a_1, \dots, a_N)$ 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.

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

$$Q(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{t=1}^{T-1} \log \frac{\exp\{h_i(a_{itm}, x_{tm})\theta + g_i(a_{itm}, x_{tm})\}}{\sum_a \exp\{h_i(a, x_{tm})\theta + g_i(a, x_{tm})\}}, \tag{6.1}$$

where $h_i(.)$ and $g_i(.)$ are defined very similarly to $h(.)$ and $g(.)$ in the single agent case, the complication being that they are now player-specific, and the actions of other players need to be partialled out:

$$h_i(a_i, x) = \sum_{a_{-i}} \left(\prod_{j \neq i} P_j(a_j|x) \right) \left[z_i(a_i, a_{-i}, x) + \beta \sum_{x'} f_X(x'|a_i, a_{-i}, x) \sum_{a'} P_i(a'|x') h(a', x') \right]$$

$$g_i(a_i, x) = \beta \sum_{a_{-i}} \left(\prod_{j \neq i} P_j(a_j|x) \right) \left[\sum_{x'} f_X(x'|a_i, a_{-i}, x) \sum_{a'} P_i(a'|x') \left\{ e(a', x') + g(a', x') \right\} \right].$$

Converting the above to expectations gives us

$$\begin{aligned} h_i(a_i, x) &= \mathbb{E}[z_i(a_i, a_{-i}, x)|a_i, x] + \beta \mathbb{E}\left[h(a', x')|a_i, x\right], \\ g_i(a_i, x) &= \mathbb{E}\left[e(a', x') + \beta g(a', x')|a_i, x\right]. \end{aligned} \tag{6.2}$$

In contrast to (2.1) in the single agent case, the expectation now 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 $f_X(.)$ are then used to recursively solve for $h_i(.)$ and $g_i(.)$ using equation (6.2). 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-likelihood (Aguirregabiria and Mira, 2007) or minimum distance estimation (Pesendorfer and Schmidt-Dengler, 2008). Both these approaches have been proposed for discrete states. For continuous states, Bajari et al. (2007) have proposed an alternative method to solve (6.2) by forward Monte Carlo simulation. Though computationally cheaper than discretization (which could give rise to a very high dimension of states), forward simulation is still cumbersome with many continuous states and players.

By contrast, our algorithm is a straightforward extension of the ones we suggested in earlier sections for single agent models. Let $\hat{\eta}_i(a_i, x)$ denote some non-parametric estimate of the choice probabilities for player i. We then (approximately) parameterize $h_i(.)$ and $g_i(.)$ as

$$h_i(a_i, x) :\approx \phi(a_i, x)^\intercal \omega_i^*; \quad g_i(a_i, x) :\approx r(a_i, x)^\intercal \xi_i^*,$$

where, as before, $\phi(a_i, x)$ and $r(a_i, x)$ are comprised of a set of basis functions over the domain of (a_i, x). The dictionary $\phi(\cdot, \cdot), r(\cdot, \cdot)$ could potentially change with i, but for ease of notation we will not make this explicit.

We now proceed to estimate the value weights ω_i^*, ξ_i^* player-by-player exactly as in Section 3:

$$\begin{aligned} \hat{\omega}_i &= \mathbb{E}_n\left[\phi(a_i, x)\left(\phi(a_i, x) - \beta\phi(a_i', x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\phi(a_i, x) z_i(a_i, a_{-i}, x)\right], \\ \hat{\xi}_i &= \mathbb{E}_n\left[r(a_i, x)\left(r(a_i, x) - \beta r(a_i', x')\right)^\intercal\right]^{-1} \mathbb{E}_n\left[\beta r(a_i, x) e(a_i', x'; \hat{\eta}_i)\right], \end{aligned} \tag{6.3}$$

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

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

Remarkably, the estimation strategy in (6.3) does not require partialling out the other players' actions, leading to a tremendous reduction of computation. Indeed, the procedure automatically takes expectations over the actions of the other players using the empirical distribution. To see this in the discrete case, note that the cell average of $z_i(a_i, a_{-i}, x)$ over a_{-i} given (a_i, x) is an unbiased estimator of the expectation $\sum_{a_{-i}} \prod_{j \neq i} P_j(a_j|x) z(a_i, a_{-i}, x)$. This intuition also goes through with continuous states since we use a functional approximation, which provides an automatic regularization for calculating the above expectation 'internally' as long as the dimension of $\phi(.)$ and $r(.)$ is sufficiently small relative to the sample size.

Using the estimated weights, we compute $\hat{h}_i(a, x) = \phi(a, x)^\intercal \hat{\omega}_i$ and $\hat{g}_i(a, x) = \phi(a, x)^\intercal \hat{\xi}_i$. These quantities can then be plugged into the pseudo-MLE (6.1) to obtain an estimate for θ. Alternatively, we can construct a locally robust estimator for θ in analogy with that for single-agent models. To describe this, we recast the pseudo-MLE criterion function in the form $Q(a, x; \theta, \{\omega_i\}, \{\xi_i\}) = \sum_i Q_i(a_i, x; \theta, \omega_i, \xi_i)$, where

$$Q_i(a_i, x; \theta, \omega_i, \xi_i) := \log \frac{\exp\left\{(\phi(a_i, x)^\intercal \omega_i)\theta + r(a_i, x)^\intercal \xi_i\right\}}{\sum_a \exp\left\{(\phi(a, x)^\intercal \omega_i)\theta + r(a, x)^\intercal \xi_i\right\}}.$$

Denote $m_i(a_i, x; \theta, \omega_i, \xi_i) := \partial_\theta Q_i(a_i, x; \theta, \omega_i, \xi_i)$. Note that for each i, ω_i^* and ξ_i^* are estimated using the moment functions

$$\mathbb{E}[\varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega_i^*)] = 0, \text{ and } \mathbb{E}[\varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi_i^*)] = 0, \tag{6.4}$$

where $(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}) := (a_i, x, a_i', x')$, and in view of (6.3),

$$\begin{aligned}
\varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega) &:= \phi(a_i, x) z(a_i, x) + \phi(a_i, x) \left(\beta \phi(a_i', x') - \phi(a_i, x)\right)^\intercal \omega, \text{ and} \\
\varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi) &:= \beta r(a_i, x) e(a_i', x'; \hat{\eta}_i) + r(a_i, x) \left(\beta r(a_i', x') - r(a_i, x)\right)^\intercal \xi.
\end{aligned}$$

Thus the locally robust moment for θ is

$$\mathbb{E}\left[\sum_i \zeta^{(i)}\left(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta^*, \omega_i^*, \xi_i^*\right)\right] = 0, \tag{6.5}$$

where

$$\begin{aligned}
\zeta^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta, \omega_i, \xi_i) := m_i(a_i, x; \theta, \omega_i, \xi_i) &- \mathbb{E}[\partial_{\omega_i} m_i(a_i, x; \theta, \omega_i, \xi_i)] \mathbb{E}[\partial_{\omega_i} \varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega_i)]^{-1} \varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega_i) \\
&- \mathbb{E}[\partial_{\xi_i} m_i(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta, \omega_i, \xi_i)] \mathbb{E}[\partial_{\xi_i} \varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi_i)]^{-1} \varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi_i).
\end{aligned}$$

For computation, we employ cross-fitting as in the single-agent setting and randomly split the markets into two samples $\mathcal{N}_1$ and $\mathcal{N}_2$. We compute $\{\hat{\omega}_i\}, \{\hat{\xi}_i\}$ using one of the samples, say $\mathcal{N}_2$. Denote by $\mathbb{E}_n^{(1)}[\cdot]$ the empirical expectation defined earlier in this section, but constructed only from observations in $\mathcal{N}_1$. The locally robust estimate $\hat{\theta}$ is the solution to the moment equation

$$\mathbb{E}_n^{(1)}\left[\sum_i \zeta_n^{(i)}\left(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta, \hat{\omega}_i, \hat{\xi}_i\right)\right] = 0, \tag{6.6}$$

where

$$\begin{aligned}
\zeta_n^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta, \omega_i, \xi_i) := m_i(a_i, x; \theta, \omega_i, \xi_i) &- \mathbb{E}_n^{(1)}[\partial_{\omega_i} m_i(a_i, x; \theta, \omega_i, \xi_i)] \mathbb{E}_n^{(1)}[\partial_{\omega_i} \varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega_i)]^{-1} \varphi_h^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \omega_i) \\
&- \mathbb{E}_n^{(1)}[\partial_{\xi_i} m_i(a_i, x; \theta, \omega_i, \xi_i)] \mathbb{E}_n^{(1)}[\partial_{\xi_i} \varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi_i)]^{-1} \varphi_g^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}, \xi_i).
\end{aligned}$$

Note that $\hat{\theta}$ can be equivalently computed player-by-player, via $\mathbb{E}_n^{(1)}\left[\zeta_n^{(i)}(\tilde{\mathbf{a}}_i, \tilde{\mathbf{x}}; \theta, \hat{\omega}_i, \hat{\xi}_i)\right] = 0$, if there were no common parameters θ across players, i.e if we could be partition $\theta \equiv (\theta_1, \ldots, \theta_N)$.

The locally robust estimator (6.6) has the same form as (3.14), except for there being separate correction terms for the estimates $\hat{\omega}_i, \hat{\xi}_i$ of each player i. Its theoretical properties are equivalent to, and can be derived in the same manner, as that for single agent models. We therefore omit these for brevity.

By the same reasoning as in Section 3.1, it possible to show that with discrete states, $h_i(.)$ and $g_i(.)$ are numerically identical to the estimates obtained by plugging in cell estimates $\hat{P}_j(\cdot|x)$ and $\hat{f}_X(.)$ in (6.2). This implies the psuedo-likelihood with plug-in estimates for $h(.)$ and $g(.)$ is not efficient even with discrete states, as discussed by Aguirregabiria and Mira (2007). However the values of $h(.)$ and $g(.)$ can be plugged into other, more efficient objectives, such as our locally robust estimator or the minimum distance estimator of Pesendorfer and Schmidt-Dengler (2008). With continuous states, one would need to employ locally robust corrections even for the minimum distance estimator to recover parametric rates of convergence for θ. To this end, we can use the fact that the minimum distance estimator can be characterized by a moment criterion. Combining this with the moments implied by (6.3) for ω and ξ, it is easy to see how the construction of Section 3.3 can be extended to the minimum distance estimator.

One could also use a recursive version of our algorithm as in Section 4.2. This is equivalent to full information MLE under some additional conditions (see Kasahara and Shimotsu, 2012). Finally, it is also straightforward to incorporate the other extensions from Section 4 to the setup of dynamic games.

7. Simulations

We run Monte Carlo Simulations to test our estimation method. We start with the simplest version of our algorithm for DDC models described in Section 3, before moving on to the recursive version of our algorithm from Section 4.2, and introducing permanent unobserved heterogeneity as described in Section 5. Our simulations for the DDC models are based on a modified version of the Rust (1987) engine replacement problem. We start by describing the setup in Section 7.1, and provide the simulation results in Section 7.2.

In a second set of Monte Carlo simulations, we test our estimation method for dynamic discrete games. Our simulations for these models are based on the dynamic firm entry game used in Aguirregabiria and Mira (2007). We describe the setup of this game in Section 7.3, before providing our simulation results for dynamic discrete games in Section 7.4.

7.1. **Bus Engine Replacement Problem.** Consider the following version of the Rust (1987) bus engine replacement problem which is adapted from Arcidiacono and Miller (2011). Each period $t = 1, ..., T; T < \infty$, Harold Zurcher decides whether to replace the engine of a bus ($a_t = 0$), or keep it ($a_t = 1$). Denote his action by $j \in \{0, 1\}$. Each bus is characterized by a permanent type $s \in \{1, 2\}$, and the mileage accumulated since the last engine replacement $x_t \in \{1, 2, ...\}$. Harold Zurcher observes both s and x_t. As in Section 3, we start by also treating both s and x_t as observed to the econometrician.

Mileage increases by one unit if the engine is kept in period t and is set to zero if the engine is replaced. The current period payoff for keeping the engine is given by $\theta_0+\theta_1 x_t+\theta_2 s+e_{1t}$, where $\theta^* \equiv \{\theta_0, \theta_1, \theta_2\}$ are the structural parameters of interest, and e_{jt} is a choice-specific transitory shock that follows a Type 1 Extreme Value distribution. As in Arcidiacono and Miller (2011), we normalize the current period payoff of replacing the engine to e_{0t}.

When deciding whether to keep or replace the engine, Harold Zurcher solves a DDC problem and sequentially maximizes the following discounted sum of payoffs:

$$E\left[\sum_{t=1}^{T} \beta^t \left\{a_t(\theta_0 + \theta_1 x_t + \theta_2 s) + e_{jt}\right\}\right],$$

where β is a discount factor that we set to 0.9.

Define the ex-ante value functions in period t as the discounted sum of current and future payoffs before the shock e_{jt} is realized and before decision a_t is made, conditional on choosing optimally in every period including t. Denote these ex-ante value functions by $V(x_t, s)$. Further define the conditional value functions $v_j(x, s)$ as the current period payoff of choice j net of e_{jt}:

$$v_j(x,s) = \begin{cases} \beta V(0,s) & j = 0 \\ \theta_0 + \theta_1 x_t + \theta_2 s + \beta V(x+1, s) & j = 1. \end{cases}$$

Denote by $p_0(x, s)$ the conditional probability of replacing the engine given x and s. Given the distributional assumptions about the shocks, this will be given by

$$p_0(x,s) = \frac{1}{1 + \exp[v_1(x,s) - v_0(x,s)]}.$$

To carry out the simulations, we recursively derive the value functions $v_j(x, s)$ for each possible combination of x, s and t. We then use these to compute the conditional replacement probabilities for the same set of combinations of variables. We generate data for 1000 buses and 2000 time periods. The mileage of each bus is first set to zero in $t = 0$. We then simulate the choices a_t using the conditional replacement probabilities $p_0(x, s)$. Finally, we restrict the generated data to 30 time periods between $t = 1000$ and $t = 1030$. This is to ensure that our data is close to being drawn from a stationary model. Our final dataset consists of types s, mileages x_t and choices a_t for 1000 buses with 30 time period observations each.

7.2. **Simulation Results - DDC Model .** This section reports the simulation results for the single-agent DDC model described above. We start by providing results for our basic algorithm, before showing simulations using the recursive version of our algorithm where we avoid the initial estimation of choice probabilities. Finally, we provide simulations for a setting with permanent unobserved heterogeneity.

For our basic algorithm, we use the locally robust version of the estimator described in 3.3. To highlight the gain in using the locally robust version of our estimator, we also generate results for the version of our estimator which is suboptimal under continuous state variables (see Section 3.1). We run 1000 simulations with 1000 buses and 30 time periods each. Each round of the simulations proceeds by first generating a dataset as described in Section 7.1.

To generate the locally robust estimator, we randomly split this dataset into two samples, $\mathcal{N}_1$ and $\mathcal{N}_2$. We then parameterize $h(a,x)$ and $g(a,x)$ using a polynomial in s, x_t and a_t with $k_\phi = k_r = 16$ terms. [6]The choice probabilities η are estimated using a logit model that is a function of the state variables s and x_t, where the same polynomial is used as before. Using only observations from the first sample $\mathcal{N}_1$ and the estimated choice probabilities $\hat{\eta}$, we then estimate the ω parameters using equation 3.3, and the ξ parameters using equation 3.5. Using the observations from the second sample $\mathcal{N}_2$, we finally obtain estimates for the θ^* parameters as the solution to the moment equations 3.14. Following the outlined cross-fitting procedure, we repeat the estimation using the observations from the second sample $\mathcal{N}_2$ to obtain estimates for ω and ξ, and the observations from the first sample $\mathcal{N}_1$ to obtain estimates for the θ^* parameters. Our final $\hat{\theta}$ is a weighted average of the θ^* estimates from the two samples. In contrast to the locally robust estimator, the non-locally robust version of our estimator uses the full sample to estimate all parameters and obtains $\hat{\theta}$ using equation 3.9.

Panel A. in Table 1 shows the results. Column (1) reports the true parameters of the model. Columns (2) – (4) report the results for the version of our estimator which is suboptimal under continuous state variables. The results for our locally robust estimator are reported in columns (5) – (7). Column (5) shows that our estimator produces parameter estimates which are closely centered around the true values. The absolute bias after 1000 simulations is less than half of a percent for all three parameters. These results are comparable to those found by Arcidiacono and Miller (2011) in a similar version of the bus engine replacement problem. However, in contrast to their CCP method, our estimator does not exploit a finite dependence property. When comparing the results from our locally robust estimator to the results from the suboptimal estimator in column (2), it can be seen that the absolute bias is smaller for all three parameter estimates. However the variance of the locally robust estimator is higher which is due to the sample splitting employed in the locally robust procedure. Overall, while the locally robust estimator is supposed to work better than the non-robust version in theory, we find that there is no important difference between the two versions of the algorithm in practice.

To generate simulation results for the recursive version of our algorithm, we follow the steps outlined in Section 4.2. As before, we run 1000 simulations with 1000 buses and 30 time periods each. We also provide results for both the non-locally robust and the locally robust estimator, where the latter are generated by splitting the sample as outlined above. Panel B. in Table 1 shows the results. Columns (2) and (5) show that our estimator produces estimates that are closely centered around the true values of the three structural parameters. When comparing the locally robust estimator to the non-robust estimator, the absolute bias is smaller for θ_0, but higher for θ_1 and θ_2. As in the simulations for the basic algorithm, the variance is higher when using the locally robust estimator.

In a final set of simulations, we introduce permanent unobserved heterogeneity into our setting by assuming that the permanent bus type $s \in \{1, 2\}$ is unknown to the researcher. To generate

[6]These include all terms of the third order polynomial plus pairwise interactions of x_t^3 with s and a_t, and three-wise interactions of both x_t^2 and x_t^3 with $s$and a_t.

Table 1. Simulation Results - DDC Model I

		not locally robust			locally robust		
	DGP (1)	TDL (2)	bias (3)	MSE (4)	TDL (5)	bias (6)	MSE (7)
A. Basic Algorithm							
θ_0 (intercept)	2.0	1.9804	-0.0196	0.0077	1.9928	-0.0072	0.0101
		(0.0859)			(0.1005)		
θ_1 (mileage)	-0.15	-0.1492	0.0008	1.2e-05	-0.1496	0.0004	2.5e-05
		(0.0033)			(0.0050)		
θ_2 (bus type)	1.0	1.0029	0.0029	0.0033	0.9997	-0.0003	0.0060
		(0.0574)			(0.0773)		
B. Recursive Algorithm							
θ_0 (intercept)	2.0	1.9806	-0.0194	0.0077	1.9967	-0.0033	0.0148
		(0.0858)			(0.1218)		
θ_1 (mileage)	-0.15	-0.1493	0.0007	1.2e-05	-0.1509	-0.0009	3.6e-05
		(0.0033)			(0.0059)		
θ_2 (bus type)	1.0	1.0037	0.0037	0.0033	1.0111	0.0111	0.0047
		(0.0574)			(0.0676)		

Notes: The table reports results for 1000 simulations. Column (1) shows the true parameter values in the model. Columns (2) and (5) report the mean and standard deviations for the estimated parameters. Columns (2)-(4) are based on the estimation method without correction function, columns (5)-(7) report results for the locally robust estimator. For both methods, the absolute bias and the mean squared error are reported in columns (3)/(4) and (6)/(7), respectively.

results for these simulations, we follow the steps outlined in Section 5 where we pair our techniques with a sequential EM algorithm (Arcidiacono and Jones, 2003). The results are shown in Table 2. As before, our algorithm produces parameter estimates that are closely centered around the true values. Compared to the results without permanent unobserved heterogeneity, the standard deviation of our estimates is slightly higher due to the uncertainty around the bus type s.

7.3. **Firm Entry Game .** Consider the following firm market entry game described in Aguirregabiria and Mira (2007). There are $i = 1, ..., 5$ firms (players), deciding whether 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 $\Pi_{itm} = \theta_{RS} ln(S_{tm}) + \theta_{RN} ln(1 + \sum_{j \neq i} a_{jtm}) - \theta_{FC,i} - \theta_{EC}(1 - a_{i(t-1)m}) + \varepsilon_{itm}$, where $ln(S_{tm}) \in \{1, 2, 3, 4, 5\}$ is a measure of consumer market size of market m in period t with $ln(S_{tm})$ following a first order Markov process, and ε_{itm} is a transitory shock that follows a logistic distribution. The profit of not entering is normalized to zero, and the discount factor β is set to 0.95 in this application. The parameters $\theta^* \equiv \{\theta_{RS}, \theta_{RN}, \theta_{FC,i}, \theta_{EC}\}$ are the structual parameters of interest. The state variables in this setting are given by the

Table 2. Simulation Results - DDC Model II

	DGP (1)	TDL (2)	bias (3)	MSE (4)
Unobserved Heterogeneity				
θ_0 (intercept)	2.0	1.9794	-0.0206	0.0158
		(0.1239)		
θ_1 (mileage)	-0.15	-0.1491	0.0009	1.5e-05
		(0.0038)		
θ_2 (bus type)	1.0	0.9992	0.0008	0.0098
		(0.0991)		

Notes: The table reports results for 1000 simulations. Column (1) shows the true parameter values in the model. Column (2) reports the mean and standard deviations for the estimated parameters. The absolute bias and mean squared error are reported in columns (3) and (4). The results are based on the estimation method without correction function.

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 follow Aguirregabiria and Mira (2007) and choose specific values for the structural parameters θ^* ($\theta_{RS} = 1, \theta_{RN} = 1, \theta_{FC,1} = 1.9, \theta_{FC,2} = 1.8, \theta_{FC,3} = 1.7, \theta_{FC,4} = 1.6, \theta_{FC,5} = 1.5, \theta_{EC} = 1$) and the transition probabilities $ln(S_{tm})$,[7] and solve for the Markov-Perfect-Equilbrium of the game. This is done by finding the firms' conditional value functions $\nu_j(S_{tm}, a_{(t-1)m})$ for each of the $2^5 \times 5 = 160$ 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 equilbrium probabilties, we compute the equilibrium distribtuion of state variables.

We generate data for $1,000$ markets with $T = 2$ time periods. To do so, we start by drawing a combination of state variables from the equilibrium distribution. Based on the state variables, we then draw choices a_{itm} for t=1 using the equilibrium choice probabilities. To generate data for the second period, we first draw $S_{(t+1)m}$ using the transition probabilities for market size, and then derive new choices $a_{(t+1)m}$ based on the period-$(t+1)$ state variables and the choice probabilities.

7.4. **Simulation Results - Dynamic Discrete Games .** We run 1000 simulations with $1,000$ markets and $T = 2$ time periods. Each round of the simulations begins by generating new data as described in Section 7.3. We then parameterize $h_i(a_i, x)$ and $g_i(a_i, x)$ using a polynomial in $S_{tm}, a_{(t-1)m}$ and a_{itm}. The choice probabilities are estimated using individual logit models for each firm with S_{tm} and $a_{(t-1)m}$ as explanatory variables. We then estimate the parameters ω_i and ξ_i individually for each player using equation 6.3. Finally, we obtain estimates for the θ^* parameters as the solutions to the pseudo-likelihood function 6.1.

The results are shown in Table 3.

[7]The matrix of transition probabilities for S_{tm} is given by $\begin{pmatrix} 0.8 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.6 & 0.2 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.6 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.2 & 0.6 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.2 & 0.8 \end{pmatrix}$.

Table 3. Simulation Results - Dynamic Discrete Game

	DGP (1)	TDL (2)	bias (3)	MSE (4)
θ_{RS} (market size)	1.0			
θ_{RN} (number of entrants)	1.0			
$\theta_{FC,1}$ (fixed cost firm 1)	1.9			
$\theta_{FC,2}$ (fixed cost firm 2)	1.8			
$\theta_{FC,3}$ (fixed cost firm 3)	1.7			
$\theta_{FC,4}$ (fixed cost firm 4)	1.6			
$\theta_{FC,5}$ (fixed cost firm 5)	1.5			
θ_{EC} (entry cost)	1.0			

Notes: The table reports results for 1000 simulations. Column (1) shows the true parameter values in the model. Column (2) reports the mean and standard deviations for the estimated parameters. The absolute bias and mean squared error are reported in columns (3) and (4). The results are based on the estimation method without correction function.

8. Conclusions

We propose a new estimator for DDC models which overcomes previous computational limitations by combining traditional CCP estimation approaches with a TD method from the Reinforcement Learning literature. In making use of simple matrix inversion techniques, our estimator is computationally very cheap and therefore fast. Unlike previous estimation methods, it is able to handle large state spaces in settings where a finite dependence property does not hold. This is of particular importance in settings with continuous state variables where discretization often gives rise to a very high-dimensional state space, or for the estimation of dynamic discrete games. At the same time, our estimator is as efficient as other approaches in simple versions of the DDC problem. We prove the statistical properties of our estimator and show that it is consistent and converges at parametric rates. Preliminary Monte Carlo simulations using a version of the famous Rust (1987) engine replacement problem confirm these properties in practice.

References

D. Ackerberg, X. Chen, J. Hahn, and Z. Liao, “Asymptotic efficiency of semiparametric two-step gmm,” *Review of Economic Studies*, vol. 81, no. 3, pp. 919–943, 2014.

V. Aguirregabiria and P. Mira, “Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models,” *Econometrica*, vol. 70, no. 4, pp. 1519–1543, 2002.

——, "Sequential estimation of dynamic discrete games," *Econometrica*, vol. 75, no. 1, pp. 1–53, 2007.

——, "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, vol. 156, no. 1, pp. 38–67, 2010.

P. Arcidiacono and J. B. Jones, "Finite mixture distributions, sequential likelihood and the em algorithm," *Econometrica*, vol. 71, no. 3, pp. 933–946, 2003.

P. Arcidiacono and R. A. Miller, "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, vol. 79, no. 6, pp. 1823–1867, 2011.

P. Arcidiacono, P. Bayer, F. A. Bugni, and J. James, "Approximating high-dimensional dynamic models: Sieve value function iteration," in *Structural Econometric Models.* Emerald Group Publishing Limited, 2013, pp. 45–95.

P. Bajari, C. L. Bankard, and J. Levin, "Estimating dynamic models of imperfect competition," *Econometrica*, vol. 75, no. 5, pp. 1331–1370, 2007.

A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations.* Springer Science & Business Media, 2012, vol. 22.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins, "Locally robust semiparametric estimation," *Working paper*, 2018.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

G. Hall, G. J. Hitsch, G. Pauletto, and J. Rust, "A comparison of discrete and parametric approximation methods for continuous-state dynamic programming problems," 2000.

V. J. Hotz and R. A. Miller, "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, vol. 60, no. 3, pp. 497–529, 1993.

V. J. Hotz, R. A. Miller, S. Sanders, and J. Smith, "A simulation estimator for dynamic models of discrete choice," *The Review of Economic Studies*, vol. 61, no. 2, pp. 265–289, 1994.

C. Johnson, "Positive definite matrices," *The American Mathematical Monthly*, vol. 77, no. 3, pp. 259–264, 1970.

H. Kasahara and K. Shimotsu, "Pseudo-likelihood estimation and bootstrap inference for structural discrete marvok decision models," *Journal of Econometrics*, vol. 146, no. 1, pp. 92–106, 2008.

——, "Sequential estimation of structural models with a fixed point constraint," *Econometrica*, vol. 80, no. 5, pp. 2303–2319, 2012.

M. P. Keane and K. I. Wolpin, "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence," *the Review of economics and statistics*, pp. 648–672, 1994.

W. K. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica*, vol. 62, no. 6, pp. 1349–1382, 1994.

——, "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, vol. 79, no. 1, pp. 147–168, 1997.

F. Niu, B. Recht, C. Re, and S. J. Wright, “Hogwild!: A lock-free approach to parallelizing stochastic gradient descent (nips 2011),” *Advances in Neural Information Processing Systems 24*, 2011.

M. Pesendorfer and P. Schmidt-Dengler, “Asymptotic least squares estimators for dynamic games,” *Review of Economic Studies*, vol. 75, no. 3, pp. 901–928, 2008.

J. Rust, “Optimal replacement of gmc bus engines: An empirical model of harold zurcher,” *Econometrica*, vol. 55, no. 5, pp. 999–1033, 1987.

V. Semenova, “Machine learning for dynamic models of imperfect information and semiparametric moment inequalities,” *arXiv preprint arXiv:1808.02569*, 2018.

R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

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

J. N. Tsitsiklis and B. Van Roy, “An analysis of temporal-difference learning with function approximation,” *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.

Appendix A. Proofs of main results

In what follows we shall drop the functional argument (a, x) when the context is clear and denote $f' \equiv f(a', x')$ for different functions f.

We start with some useful lemmas:

Lemma 1. *There exists a unique fixed point to the operator $P_\phi \Gamma_z$. If Assumption 1(i) holds, this fixed point is given by $\phi^\intercal \omega^*$, where ω^* is such that $\mathbb{E}\left[\phi\left(z + \beta\phi'^\intercal\omega^* - \phi^\intercal\omega^*\right)\right] = 0$.*

Proof. First off, we note that Γ_z, and therefore $P_\phi \Gamma_z$, are both contraction maps with the contraction factor β. This implies that that $P_\phi \Gamma_z$ has a unique fixed point. Clearly, this fixed point must lie in the space $\mathcal{L}_\phi$. Let us denote this as $\phi^\intercal \omega^*$.

Now for any function $f \in \mathcal{L}_\phi$,

$$\begin{aligned} P_\phi \Gamma_z[f] - f &= \phi^\intercal \mathbb{E}[\phi\phi^\intercal]^{-1}\mathbb{E}\left[\phi\left(z + \beta f'\right)\right] - \phi^\intercal \mathbb{E}[\phi\phi^\intercal]^{-1}\mathbb{E}[\phi f] \\ &= \phi^\intercal \mathbb{E}[\phi\phi^\intercal]^{-1}\mathbb{E}\left[\phi\left(z + \beta f' - f\right)\right]. \end{aligned}$$

Since $\phi^\intercal \omega^*$ is the fixed point, we must have

$$\phi^\intercal \mathbb{E}[\phi\phi^\intercal]^{-1}\mathbb{E}\left[\phi\left(z + \beta\phi'^\intercal\omega^* - \phi^\intercal\omega^*\right)\right] = 0.$$

But ϕ is linearly independent and $E[\phi\phi^\intercal]^{-1}$ is non-singular, by Assumption 1(i). Hence it must be the case

$$\mathbb{E}\left[\phi\left(z + \beta\phi'^\intercal\omega^* - \phi^\intercal\omega^*\right)\right] = 0.$$

This completes the proof the lemma. $\square$

For the next Lemma, we shall use the following definition of a negative-definite matrix: a square, possibly asymmetric, matrix A is said to be negative definite with the coefficient $\bar{\lambda}(A)$ if

$$\sup_{|w|=1} w^\intercal A w \le \bar{\lambda}(A) < 0.$$

For a symmetric negative-definite matrix, we have that $\bar{\lambda}(A) = \text{maxeig}(A)$, where $\text{maxeig}(\cdot)$ represents the maximal eigenvalue. We can similarly define a positive definite matrix with the coefficient $\underline{\lambda}(A)$. If the latter is also symmetric, then $\underline{\lambda}(A) = \text{mineig}(A)$.

We note that under our definition, if A is negative definite, it is also invertible. This holds even if the matrix is asymmetric, see e.g Johnson (1970).

Lemma 2. *Under Assumption 1(i), the matrix $A := \mathbb{E}\left[\phi\left(\beta\phi' - \phi\right)^\intercal\right]$ is negative definite with $\bar{\lambda}(A) \le -(1-\beta)\underline{\lambda}(\mathbb{E}[\phi\phi^\intercal])$, and is therefore invertible.*

Proof. The idea for this proof is taken from Tsitsiklis and van Roy (1997). Recall the definition of $\phi^\intercal \omega^*$ as the fixed point to $P_\phi \Gamma_z[\cdot]$ from Lemma 1. We shall now show that

$$(\omega - \omega^*)^\intercal A(\omega - \omega^*) \le -(1-\beta)\underline{\lambda}(\mathbb{E}[\phi\phi^\intercal])\left|\omega - \omega^*\right|^2 \ \forall\ \omega \in \mathbb{R}^k.$$

Observe that

$$
\begin{aligned}
A(\omega-\omega^*) &= \mathbb{E}\left[\phi\left(z+\beta\phi'^{\intercal}\omega-\phi^{\intercal}\omega\right)\right]-\mathbb{E}\left[\phi\left(z+\beta\phi'^{\intercal}\omega^*-\phi^{\intercal}\omega^*\right)\right] \\
&= \mathbb{E}\left[\phi\left(z+\beta\phi'^{\intercal}\omega-\phi^{\intercal}\omega\right)\right],
\end{aligned}
$$

since the second expression in the first equation is 0. Now,

$$
\begin{aligned}
\mathbb{E}\left[\phi\left(z+\beta\phi'^{\intercal}\omega-\phi^{\intercal}\omega\right)\right] &= \mathbb{E}\left[\phi(a,x)\left(z(a,x)+\beta\mathbb{E}\left[\phi(a',x')^{\intercal}\omega|a,x\right]-\phi(a,x)^{\intercal}\omega\right)\right] \\
&= \mathbb{E}\left[\phi\left(\Gamma_z[\phi^{\intercal}\omega]-\phi^{\intercal}\omega\right)\right] \\
&= \mathbb{E}\left[\phi\left(P_\phi\Gamma_z[\phi^{\intercal}\omega]-\phi^{\intercal}\omega\right)\right],
\end{aligned}
$$

where the last equality holds since $\mathbb{E}\left[\phi(I-P_\phi)[f]\right]=0$ for all f. We thus have

$$
\begin{aligned}
(\omega-\omega^*)^{\intercal}A(\omega-\omega^*) &= \mathbb{E}\left[\left(\omega^{\intercal}\phi-\omega^{*\intercal}\phi\right)\left(P_\phi\Gamma_z[\phi^{\intercal}\omega]-\phi^{\intercal}\omega\right)\right] \\
&= \mathbb{E}\left[\left(\omega^{\intercal}\phi-\omega^{*\intercal}\phi\right)\left(P_\phi\Gamma_z[\phi^{\intercal}\omega]-\phi^{\intercal}\omega^*\right)\right]-\left\|\phi^{\intercal}\omega-\phi^{\intercal}\omega^*\right\|_2^2.
\end{aligned}
$$

Since $P_\phi\Gamma_z[\cdot]$ is a contraction mapping with contraction factor β, it follows

$$
\left\|P_\phi\Gamma_z[\phi^{\intercal}\omega]-\phi^{\intercal}\omega^*\right\|_2^2=\left\|P_\phi\Gamma_z[\phi^{\intercal}\omega]-P_\phi\Gamma_z\left[\phi^{\intercal}\omega^*\right]\right\|_2^2\leq\beta\left\|\phi^{\intercal}\omega-\phi^{\intercal}\omega^*\right\|_2^2.
$$

In view of the above,

$$
\begin{aligned}
(\omega-\omega^*)^{\intercal}A(\omega-\omega^*) &\leq -(1-\beta)\left\|\phi^{\intercal}\omega-\phi^{\intercal}\omega^*\right\|_2^2 \\
&= -(1-\beta)(\omega-\omega^*)^{\intercal}\mathbb{E}[\phi\phi^{\intercal}](\omega-\omega^*) \\
&\leq -(1-\beta)\underline{\lambda}(\mathbb{E}[\phi\phi^{\intercal}])\left|\omega-\omega^*\right|^2.
\end{aligned}
$$

This completes the proof of the lemma. □

Lemma 3. *Suppose that Assumption 1(i) holds. Then,*

$$
\left\|h-\phi^{\intercal}\omega^*\right\|_2\leq(1-\beta)^{-1}\left\|h-P_\phi[h]\right\|_2.
$$

Proof. Recall that $h(\cdot,\cdot)$ is the unique fixed point of Γ_z, and similarly, $\phi^{\intercal}\omega^*$ is the unique fixed point of $P_\phi\Gamma_z$. The operator Γ_z is a contraction mapping with contraction factor β. Furthermore, the projection operator P_ϕ is linear, and $\left\|P_\phi[f]\right\|_2\leq\|f\|_2$ for any function f. Thus

$$
\begin{aligned}
\left\|h-\phi^{\intercal}\omega^*\right\|_2 &\leq \left\|h-P_\phi[h]\right\|_2+\left\|P_\phi[h]-P_\phi\Gamma_z[\phi^{\intercal}\omega^*]\right\|_2 \\
&\leq \left\|h-P_\phi[h]\right\|_2+\left\|h-\Gamma_z[\phi^{\intercal}\omega^*]\right\|_2 \\
&= \left\|h-P_\phi[h]\right\|_2+\left\|\Gamma_z[h]-\Gamma_z[\phi^{\intercal}\omega^*]\right\|_2 \\
&\leq \left\|h-P_\phi[h]\right\|_2+\beta\left\|h-\phi^{\intercal}\omega^*\right\|_2.
\end{aligned}
$$

Rearranging the above expression proves the desired claim. □

For the proofs of Theorems 1-2, we shall 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 $P\equiv P_1\times\cdots\times P_T$. We will denote $E[\cdot]$ as the expectation over

P. Furthermore, we shall 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 $\mathbb{E}[\cdot]$ appropriately: Let $\mathbb{P}$ denote the relative frequency of occurrence of (a, x, a', x') in the data as $n \to \infty$. Let $\mathbb{E}[\cdot]$ denote the corresponding expectation over $\mathbb{P}$. Note that P is different from $\mathbb{P}$ since the latter provides the distribution of (a, x, a, x') after dropping the time index. However, the two are related since for any function f, we can write $\mathbb{E}[f(a, x, a', x')] = (T-1)^{-1} \sum_{t=1}^{T-1} E[f(a_{it}, x_{it}, a_{it+1}, x_{it+1})]$ (we could alternatively use this as the definition of $\mathbb{E}[\cdot]$ itself). These updated definitions of $\mathbb{P}$ and $\mathbb{E}[\cdot]$ are applicable wherever these notations are used 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 consider t as also a part of 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.** That ω^* exists follows from Lemma 1. To prove that $\hat{\omega}$ exists, it suffices to show that $\hat{A} := \mathbb{E}_n [\phi (\beta\phi' - \phi)^\intercal]$ is invertible with probability approaching 1. Recall that by our notation above, $\hat{A} = (n(T-1))^{-1} \sum_i \sum_{t=1}^{T-1} \phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$, while $A = (T-1)^{-1} \sum_{t=1}^{T-1} E[\phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$. We can thus write $\left|\hat{A} - A\right| \le (T-1)^{-1} \sum_{t=1}^{T-1} \left|\hat{A}_t - A_t\right|$, where $\hat{A}_t := n^{-1} \sum_i \phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal$ and $A_t := E[\phi_{it}(\beta\phi_{it+1} - \phi_{it})^\intercal]$. Now, by Assumption 1(ii), $|\phi(a, x)|_\infty \le M$ independent of k_ϕ. We then have

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

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 $\bar{\lambda}(\hat{A}) \le \bar{\lambda}(A) + \left|\hat{A} - A\right| \le \bar{\lambda}(A) + o_p(1)$. Since $\bar{\lambda}(A) < 0$, this proves that $\bar{\lambda}(\hat{A}) < 0$ with probability approaching 1, and subsequently, that $\hat{A}$ is invertible. This completes the proof of the first claim.

The second claim follows directly from Lemma 3 and Assumption 1(iii).

For the third claim, let us 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

$$
(\hat{\omega} - \omega^*)^\intercal(-\hat{A})(\hat{\omega} - \omega^*) = (\hat{\omega} - \omega^*)^\intercal(b - \hat{b}) + (\hat{\omega} - \omega^*)^\intercal(\hat{A} - A)\omega^*. \tag{A.1}
$$

Now, earlier in the proof we have showed that $\left|\hat{A} - A\right| = O_p(k_\phi/\sqrt{n})$. Hence it follows $\underline{\lambda}(-\hat{A}) \ge \underline{\lambda}(-A) + o_p(1)$. We thus have

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

with probability approaching 1, for any constant $c \in (0,1)$. In view of (A.1) and (A.2),

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

with probability approaching 1.

It thus remains to bound $\left|\hat{b} - b\right|$ and $\left|\hat{A}\omega^* - A\omega^*\right|$. By similar arguments 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

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

This proves

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

In a similar vein,

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

as long as

$$\mathbb{E}\left[|\phi(\beta\phi - \phi)^\intercal \omega^*|^2\right] = O(k_\phi).$$

But the latter is true under Assumptions 1(ii)-(iv) since

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

and

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

where the second inequality uses the facts $\|\phi^\intercal\omega^* - h\|_2 = O(k_\phi^{-\alpha})$ (as shown in the second claim of this theorem), and $|h(\cdot,\cdot)|_\infty \leq (1-\beta)^{-1}|z(\cdot,\cdot)|_\infty < (1-\beta)^{-1}L$ (which can be easily verified using (2.1) and Assumption 1(iv)). Combining the above, we thus conclude there exists $C < \infty$ such that

$$|\hat{\omega} - \omega^*| \leq C\sqrt{\frac{k_\phi}{n}},$$

with probability approaching one. This completes the proof of the third claim.

Finally, to prove the last claim, observe that

$$\begin{aligned} \|\phi^\intercal\hat{\omega} - h\|_2^2 &\leq 2\|\phi^\intercal\hat{\omega} - \phi^\intercal\omega^*\|_2^2 + 2\|\phi^\intercal\omega^* - h\|_2^2 \\ &= 2(\hat{\omega} - \omega^*)^\intercal\mathbb{E}[\phi\phi^\intercal](\hat{\omega} - \omega^*)^{1/2} + 2\|\phi^\intercal\omega^* - h\|_2^2 \\ &\leq \bar{\lambda}(\mathbb{E}[\phi\phi^\intercal])O_p\left(\frac{k_\phi}{n}\right) + O_p(k_\phi^{-\alpha}), \end{aligned}$$

where the second inequality follows from the second and third claims of this Theorem. But

$$\bar{\lambda}(\mathbb{E}[\phi\phi^\intercal]) \leq \|\phi\|_2^2 \leq M^2 k_\phi,$$

by Assumption 1(iv). Combining the above proves the last claim.

A.2. **Proof of Theorem 2.** We note that the proofs of the first two claims follows from analogous arguments as used in the proof of Theorem 1. We thus only need consider the third claim of the theorem. The fourth claim is a straightforward consequence of this.

Recall that we use a cross-fitting procedure for estimating ξ^*. Let n_1, n_2 denote the sample sizes in the two folds. Also let $\hat{\eta}_1, \hat{\xi}_1$ and $\hat{\eta}_2, \hat{\xi}_2$ denote the estimates of η and ξ^* from the two folds. We shall show that $|\hat{\xi}_1 - \xi| = O_p(\sqrt{k_r/n})$. By a symmetric argument, we will also have $|\hat{\xi}_2 - \xi| = O_p(\sqrt{k_r/n})$, from which we can conclude $|\hat{\xi} - \xi| = O_p(\sqrt{k_r/n})$. To this end, let $A_r := \mathbb{E}[rr^\intercal]$, $b_r := \mathbb{E}[r(a,x)e(a',x')]$, $\hat{A}_r^{(1)} := \mathbb{E}_n^{(1)}[rr^\intercal]$ 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 observations from the first block. We shall also employ the notation $\psi(a,x,a',x';\eta) := r(a,x)e(a',x';\eta)$ and $\psi_{it}(\eta) := r(a_{it},x_{it})e(a_{it+1},x_{it+1};\eta)$.

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, we find that the only difference is in the treatment of $|\hat{b}_r^{(1)} - b_r|$. As in that proof, define $\hat{b}_{rt}^{(1)} := n^{-1}\sum_i \psi_{it}(\hat{\eta}_2)$ and $b_{rt} := E[\psi_{it}(\eta)]$. 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

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

Hence the claim follows once we show

$$\hat{b}_{rt}^{(1)} - b_{rt} = \frac{1}{n_1}\sum_{i=1}^{n_1}\{\psi_{it}(\eta) - E[\psi_{it}(\eta)]\} + o_p\left(\sqrt{k_r/n}\right). \tag{A.3}$$

We now prove (A.3). Let $\mathcal{N}_2$ denote the set of all observations in the second fold. We have

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

First consider the term R_{1nt}. Define

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

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

$$E\left[|R_{1nt}|^2\Big|\mathcal{N}_2\right] = \frac{1}{n_1}E\left[|\delta_{it}|^2\Big|\mathcal{N}_2\right] = \frac{1}{n_1}E\left[|\psi_{it}(\hat{\eta}_2) - \psi_{it}(\eta)|^2\Big|\mathcal{N}_2\right]. \tag{A.4}$$

Now for any (a,x,a',x'), we can note from the definition of $\psi(\cdot)$ that with probability approaching 1,

$$\begin{aligned}
|\psi(a,x,a',x';\hat{\eta}_2) - \psi(a,x,a',x';\eta)| &\le |r(a,x)|\{|\ln\hat{\eta}_2 - \ln\eta| + |\hat{\eta}_2 - \eta|\} \\
&\le M\sqrt{k_r}\{|\ln\hat{\eta}_2 - \ln\eta| + |\hat{\eta}_2 - \eta|\} \\
&\le M\sqrt{k_r}(2\delta^{-1} + 1)|\hat{\eta}_2 - \eta|, \qquad \text{(A.5)}
\end{aligned}$$

where the second inequality follows from Assumption 2(iii), and the third inequality follows from Assumption 2(v).[8] Thus in view of (A.4) and (A.5), there exists $C < \infty$ such that

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

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

$$|R_{1nt}| = o_p(\sqrt{k_r/n}). \tag{A.6}$$

Next consider the term R_{2nt}. We note that $E[\psi_{it}(\eta)]$ is twice Fréchet differentiable. In the main text we have shown that $\partial_\eta E[\psi_{it}(\eta)] = 0$ (c.f equation (3.7)). Furthermore, following some straightforward algebra it is possible to show $|\partial_\eta^2 E[\psi_{it}(\eta)]| \leq C_1\sqrt{k_r}$, for some $C_1 < \infty$, as long as η is bounded away from 0 (as assured by Assumption 2(v)). Hence

$$E\left[|R_{2nt}|^2\Big|\mathcal{N}_2\right] \leq C_1\sqrt{k_r}E\left[|\hat{\eta}_2(a_{it+1}, x_{it+1}) - \eta(a_{it+1}, x_{it+1})|^2\Big|\mathcal{N}_2\right]$$
$$\leq C_1T\sqrt{k_r}\|\hat{\eta}_2 - \eta\|_2^2 = o_p(k_r/n). \tag{A.7}$$

Together, (A.6) and (A.7) imply (A.3), which concludes the proof of the theorem.

A.3. **Proof of Theorem 4.** Let ω_l denote the lth update of ω. From Algorithm 1, we observe that the gradient updates are of the form

$$\omega_{l+1} = \omega_l + \alpha_\omega^{(l)}\left(z_{it}\phi_{it} - \phi_{it}(\phi_{it} - \beta\phi_{it+1})^\intercal\omega_l\right).$$

By standard results on stochastic approximation algorithms (see, e.g, Benveniste et al. (2012), Theorem 17), the above sequence of updates converges to a fixed point $\hat{\omega}$ satisfying

$$\mathbb{E}_n[z\phi - \phi(\phi - \beta\phi')^\intercal\hat{\omega}] = 0$$

as long as (1) $\mathbb{E}_n[z\phi]$ is finite, (2) $A_n := \mathbb{E}_n[\phi(\phi - \beta\phi')^\intercal]$ is negative definite, and (3) the learning rate $\alpha_\omega^{(k)}$ satisfies the requirements specified Assumption 3. The first condition is obviously satisfied under Assumption 1(iii). The second condition, that A_n is negative definite, has already been shown in the context of the proof of Theorem 1. Hence, with probability approaching 1, all the three conditions are satisfied and the sequence ω_k converges to $\hat{\omega}$. A similar analysis also applies to gradient descent updates of ξ.

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

Appendix B. Supplemental results

B.1. **Primitive conditions for Assumptions 1,2.** In the main text, we introduced Assumptions 1,2 for deriving the properties of Temporal-difference estimators for continuous states. Among other conditions, we required $h(a,x)$ and $g(a,x)$ to be well approximated by a series expansion at a $k^{-\alpha}$ rate. Newey (1997) shows that for splines and power series, we can set $\alpha = r/d$, where r is the number of continuous derivatives of $h(a,\cdot), g(a,\cdot)$, and d is the dimension of x. In determining the derivatives, we may assume without loss of generality that x is continuous; as otherwise we can always condition on the discrete elements of x by including the relevant indicator and interaction terms in the approximation spaces $\mathcal{L}_\phi, \mathcal{L}_r$ (note however that support of a is always assumed to be discrete in the setting of this paper). The following then provides primitive conditions on $z(a,x), e(a,x), f_X(x'|a,x)$ to ensure existence of r continuous derivatives of $h(a,x), g(a,x)$ for each $a \in \mathcal{A}$:

Assumption S. *For each $a \in \mathcal{A}$, the functions $z(a,x), e(a,x), f_X(x'|a,x)$ are uniformly bounded and r times continuously differentiable. Furthermore, $\sup_{a,x} \int |\partial_x^k f_X(x'|a,x)|dx' < \infty$ for $k = 1,\dots,r$.*

We shall now demonstrate that Assumption S implies $h(a,x)$ is uniformly bounded and continuously differentiable when $r \geq 1$. The extension to higher order derivatives, and to the function $g(a,x)$, follows by similar arguments.

We start by showing that $h(a,x)$ is uniformly bounded. Recall that $h(a,x)$ is the fixed point of the dynamic programming operator $\Gamma_z[\cdot]$. Define M_0 to be any positive real number such that $|z(a,x)|_\infty < \beta M_0$ (such a number exists by Assumption S). Now for any f such that $|f|_\infty < M_0$, we have

$$|\Gamma_z[f]|_\infty \leq |z(a,x)|_\infty + (1-\beta)M_0 < M_0.$$

In other words, $\Gamma_z[\cdot]$ maps the space $\mathcal{S}_0 \equiv \{f : |f|_\infty \leq M_0\}$ onto itself. Hence by the properties of contraction mappings, the fixed point of $\Gamma_z[\cdot]$ must lie in $\mathcal{S}_0$, i.e $|h(a,x)|_\infty \leq M_0$.

We now show that $h(a,x)$ is continuously differentiable in x for all a. Let L_1, L_2 be positive real numbers such that $|\partial_x z(a,x)|_\infty \leq L_1$ and $\sup_{a,x} \int |\partial_x f_X(x'|a,x)|\, dx' \leq L_2$ for all $a \in \mathcal{A}$ (the existence of these quantities is assured by Assumptions S). Now, for any $f \in \mathcal{S}_0$,

$$\begin{aligned}|\partial_x \Gamma_z[f](a,x)|_\infty &\leq |\partial_x z(a,x)|_\infty + (1-\beta)\sup_{a,x}\left|\int f(a',x')P(a'|x')\partial_x f_X(x'|a,x)da'dx'\right| \\ &\leq L_1 + (1-\beta)M_0 L_2 := M_1.\end{aligned}$$

Defining $\mathcal{S}_1 \equiv \{f : |\partial_x f|_\infty \leq M_1\}$, we have thus shown $\Gamma_z[\mathcal{S}_0] \subseteq \mathcal{S}_1$. Since we also have $\Gamma_z[\mathcal{S}_0] \subseteq \mathcal{S}_0$, it follows $\Gamma_z[\mathcal{S}_0 \cap \mathcal{S}_1] \subseteq \mathcal{S}_0 \cap \mathcal{S}_1$. Hence $h(a,x) \in \mathcal{S}_0 \cap \mathcal{S}_1$.

B.2. **Heuristic derivation of non-parametric adjustment terms.** Section 3.3.2 in the main text provides the expression for the locally robust moment (3.13). We obtained this expression by using the form of the parametric adjustment terms in (3.14) to conjecture the expression for the non-parametric correction term. This section provides the heuristic analysis underlying this conjecture.

We start by considering the parametric adjustment term for $h(a,x)$:

$$\hat{\mathcal{A}}_h := \mathbb{E}_n^{(1)}[\partial_\omega m(a,x;\theta,\mathbf{v})]\mathbb{E}_n^{(1)}[\partial_\omega \varphi_h(\tilde{\mathbf{a}},\tilde{\mathbf{x}},\omega)]^{-1}\varphi_h(\tilde{\mathbf{a}},\tilde{\mathbf{x}},\omega).$$

Denote

$$m(a,x;\theta,h,g) := \partial_\theta Q(a,x;\theta,h,g); \quad Q(a,x;\theta,h,g) := \ln \frac{\exp\{h(a,x)\theta + g(a,x)\}}{\sum_{\breve{a}} \exp\{h(\breve{a},x)\theta + g(\breve{a},x))\}}.$$

As noted in the main text, $\hat{\mathcal{A}}_h$ can be rewritten as

$$\hat{\mathcal{A}}_h = \hat{\lambda}_h(a,x;\theta)\left\{z(a,x) + \beta\phi(a',x')^\intercal\omega - \phi(a,x)^\intercal\omega\right\},$$

where

$$\hat{\lambda}_h(a,x;\theta) := \phi(a,x)^\intercal \mathbb{E}_n^{(1)}\left[\left(\beta\phi(a',x') - \phi(a,x)\right)\phi(a,x)^\intercal\right]^{-1}\mathbb{E}_n^{(1)}\left[\phi(a,x)\partial_h m(a,x;\theta,h,g)\right],$$

and $\partial_h m(\cdot)$ denotes the Fréchet derivative of $m(\cdot)$ with respect to $h(\cdot)$. We now provide a heuristic derivation for the limit, $\mathcal{A}_h$, of $\hat{\mathcal{A}}_h$ as $n, k_\phi \to \infty$.

To this end, let us keep the dimension k_ϕ fixed for now and define

$$\hat{\vartheta} := \mathbb{E}_n^{(1)}\left[\left(\beta\phi(a',x') - \phi(a,x)\right)\phi(a,x)^\intercal\right]^{-1}\mathbb{E}_n^{(1)}\left[\phi(a,x)\partial_h m(a,x;\theta,h,g)\right].$$

Note that $\hat{\lambda}_h(a,x;\theta) = \phi(a,x)^\intercal\hat{\vartheta}$. Now, in the limit as $n \to \infty$, we can expect $\hat{\vartheta} - \vartheta \to 0$, where

$$\vartheta := \mathbb{E}\left[\left(\beta\phi(a',x') - \phi(a,x)\right)\phi(a,x)^\intercal\right]^{-1}\mathbb{E}\left[\phi(a,x)\partial_h m(a,x;\theta,h,g)\right].$$

Since $\mathbb{E}[\cdot]$ is a stationary distribution, $\mathbb{E}\left[\beta\phi(a',x')\phi(a,x)^\intercal\right] = \mathbb{E}\left[\beta\phi(a,x)\phi(a^{-\prime},x^{-\prime})^\intercal\right]$, where $(a^{-\prime},x^{-\prime})$ denotes the one-step backward quantities corresponding to (a,x). In view of this, a bit of rearrangement of the previous display equation gives us

$$\mathbb{E}\left[\phi(a,x)\left\{-\partial_h m(a,x;\theta,h,g) + \beta\phi(a^{-\prime},x^{-\prime})^\intercal\vartheta - \phi(a,x)^\intercal\vartheta\right\}\right] = 0. \tag{B.1}$$

Define $\lambda_h^*(a,x;\theta) := \phi(a,x)^\intercal\vartheta$, noting also that this is the limit of $\hat{\lambda}_h(a,x)$ as $n \to \infty$. Given (B.1), we then have

$$\mathbb{E}\left[\phi(a,x)\left\{-\partial_h m(a,x;\theta,h,g) + \beta\lambda_h^*(a^{-\prime},x^{-\prime}) - \lambda_h^*(a,x)\right\}\right] = 0.$$

The above equation shares a high degree of similarity with (3.1). Indeed, backtracking the analysis leading to (3.1), we see that $\lambda_h^*(a,x;\theta)$ can be interpreted as the fixed point of the projected 'backward' dynamic programming operator $P_\phi\Gamma_{h,\theta}^\dagger[\cdot]$, where

$$\Gamma_{h,\theta}^\dagger[f](a,x) := -\partial_h m(a,x;\theta,h,g) + \beta\mathbb{E}\left[f(a^{-\prime},x^{-\prime})|a,x\right].$$

While we have supposed the dimension of $\phi(\cdot)$ to be fixed so far, as $k_\phi \to \infty$, we can expect $\lambda_h^*(a,x;\theta) \to \lambda_h(a,x;\theta)$, where the latter is the fixed point of $\Gamma_{h,\theta}^\dagger[\cdot]$ itself. From the above discussion, we thus conjecture that the limit of $\hat{\mathcal{A}}_h$ is given by

$$\mathcal{A}_h = \lambda_h(a,x;\theta)\left\{z(a,x) + \beta h(a',x') - h(a,x)\right\},$$

where we have also replaced $\phi(a,x)^\intercal\omega$ in (3.15) with its limit $h(a,x)$. This is our conjecture for the adjustment term corresponding to $h(\cdot)$.

A similar analysis also applies to the adjustment term for $g(\cdot)$, which we conjectured to be of the form

$$\mathcal{A}_g = \lambda_g(a, x; \theta)\left\{e(a', x'; \eta) + \beta g(a', x') - g(a, x)\right\},$$

where $\lambda_g(a, x; \theta)$ is the fixed point of the operator $\Gamma^{\dagger}_{g,\theta}[\cdot]$, defined as

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

B.3. **Unobserved heterogeneity and sequential EM.** Arcidiacono and Jones (2003) describe a class of two step estimation procedures for which the sequential EM algorithm may be applied. Here, we verify that Temporal-difference estimation with permanent unobserved heterogeneity does indeed into fall into such a class of procedures, and the algorithm described in Section 5 is therefore an instance of sequential EM.

First, let us describe the identification of $\boldsymbol{\omega}^*, \boldsymbol{\xi}^*$. Let $\Gamma^{(k)}_z[\cdot]$ denote the dynamic programming operator

$$\Gamma^{(k)}_z[f](a, x) := z(a, x, k) + \beta \mathbb{E}[f(a', x')|a, x, s = k].$$

Clearly, $h_k(a, x)$ is the unique fixed point of $\Gamma^{(k)}_z[\cdot]$. We define our approximation $\phi(a, x)^{\intercal}\omega^*_k$ to $h_k(a, x)$ as the fixed point of the projected dynamic programming operator $P^{(k)}_\phi \Gamma^{(k)}_z[\cdot]$, where $P^{(k)}_\phi$ is the projection operator into the conditional linear span of $\mathcal{L}_\phi$ given $s = k$, i.e

$$P^{(k)}_\phi[f](a, x) := \phi(a, x)^{\intercal}\mathbb{E}[\phi(a, x)\phi(a, x)^{\intercal}|s = k]^{-1}\mathbb{E}[\phi(a, x)f(a, x)|s = k].$$

By a similar argument as in Lemma 1 in the Appendix A, the fixed point of $P^{(k)}_\phi \Gamma^{(k)}_z[\cdot]$ satisfies

$$\mathbb{E}\left[\phi(a, x)\left\{z(a, x, k) + \beta\phi(a', x')^{\intercal}\omega^*_k - \phi(a, x)^{\intercal}\omega^*_k\right\}\middle| s = k\right] = 0,$$

or equivalently, assuming $P(s = k) \neq 0$, that

$$\mathbb{E}\left[\mathbb{I}(s = k)\phi(a, x)\left\{z(a, x, k) + \beta\phi(a', x')^{\intercal}\omega^*_k - \phi(a, x)^{\intercal}\omega^*_k\right\}\right] = 0.$$

The above enables us to identify ω^*_k as

$$\omega^*_k = \mathbb{E}\left[\mathbb{I}(s = k)\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^{\intercal}\right]^{-1}\mathbb{E}\left[\mathbb{I}(s = k)\phi(a, x)z(a, x, k)\right]. \tag{B.2}$$

By similar arguments as in Lemma 2 in Appendix A, $\mathbb{E}\left[\mathbb{I}(s = k)\phi(a, x)\left(\phi(a, x) - \beta\phi(a', x')\right)^{\intercal}\right]$ is indeed non-singular as long as $\beta < 1$ and $\mathbb{E}[\phi(a, x)\phi(a, x)^{\intercal}|\, s = k]$ is non-singular. Equation (B.2) was described as (5.2) in the main text. The identification of ξ^*_k follows by similar arguments.

Let us now suppose that $h_k(x, a)$ and $g_k(x, a)$ are truly finite-dimensional, i.e $h_k(x, a) \equiv \phi(x, a)^{\intercal}\omega^*_k$ and $g_k(x, a) \equiv r(x, a)^{\intercal}\xi^*_k$. Denote

$$l_k(a, x; \theta, \boldsymbol{\omega}, \boldsymbol{\xi}) \equiv \frac{\exp\left\{(\phi(a, x)^{\intercal}\omega_k)\theta + r(a, x)^{\intercal}\xi_k)\right\}}{\sum_a \exp\left\{(\phi(a, x)^{\intercal}\omega_k)\theta + r(a, x)^{\intercal}\xi_k\right\}},$$

$$P(s = k|\mathbf{a}_i, \mathbf{x}_i, \theta, \boldsymbol{\omega}, \boldsymbol{\xi}) = \frac{\pi^*_k \prod_{t=1}^{T-1} l_k(a_{it}, x_{it}; \theta, \boldsymbol{\omega}, \boldsymbol{\xi})}{\sum_{\tilde{k}} \pi^*_{\tilde{k}} \prod_{t=1}^{T-1} l_{\tilde{k}}(a_{it}, x_{it}; \theta, \boldsymbol{\omega}, \boldsymbol{\xi})},$$

and

$$Q_k(\mathbf{a}_i, \mathbf{x}_i; \theta, \boldsymbol{\omega}, \boldsymbol{\xi}) = \sum_{t=1}^{T-1} \ln l_k(a_{it}, x_{it}; \theta, \boldsymbol{\omega}, \boldsymbol{\xi}).$$

At the population level, we may suppose that (θ^*, π^*) is the unique solutions to the FIML problem

$$(\theta^*, \pi^*) = \arg\max_{\theta,\pi} \bar{\mathbb{E}}\left[\ln\left\{\pi_{s_i} \cdot \prod_{t=1}^{T-1} l_k(a_{it}, x_{it}, s_i; \theta, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) \cdot \prod_{t=1}^{T-1} f_X(x_{it+1}|a_{it}, x_{it}, s) \cdot f_{x_1}(x_{i1}|s_i)\right\}\right], \tag{B.3}$$

where $\bar{\mathbb{E}}[\cdot]$ denotes the expectation over the population distribution of $(\mathbf{a}_i, \mathbf{x}_i, s_i)$ and $f_{x_1}(x_{i1}|s)$ is the probability density of the intial state x_{i1} given s. Taking the First Order Condition (FOC) of (B.3) with respect to θ gives

$$\begin{aligned} 0 &= \bar{\mathbb{E}}\left[\partial_\theta \ln \prod_{t=1}^{T-1} l_k(a_{it}, x_{it}, s_i; \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\right] \\ &= \mathbb{E}\left[\sum_k P(s=k|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\partial_\theta Q_k(\mathbf{a}, \mathbf{x}; \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\right]. \end{aligned} \tag{B.4}$$

In a similar vein, taking the FOC with respect to π_k gives

$$\mathbb{E}\left[P(s=k|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) - \pi_k^*\right] = 0 \ \forall k. \tag{B.5}$$

These moment conditions have to be augmented by the identification conditions for ω_k^*, ξ_k^* which, in view of (B.2), are given by

$$\mathbb{E}\left[P(s=k|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\phi(a,x)\left\{z_k(a,x) + \beta\phi(a',x')^\intercal\omega_k^* - \phi(a,x)^\intercal\omega_k^*\right\}\right] = 0 \ \forall k,$$
$$\mathbb{E}\left[P(s=K|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)r(a,x)\left\{\beta e_K(a',x') + \beta r(a',x')^\intercal\xi_k^* - r(a,x)^\intercal\xi_k^*\right\}\right] = 0 \ \forall k.$$

Combining the above with the definitional requirement $e_k(a,x) := \gamma - \ln l_k(a,x;\theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)$, we thus see that $(\theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*, \pi^*)$ can be identified as the solution to the population moment condition:

$$\mathbb{E}_{(\mathbf{a},\mathbf{x})}\begin{bmatrix} \sum_k P(s=k|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\partial_\theta Q(\mathbf{a}, \mathbf{x}; \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) \\ P(s=1|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\phi(a,x)\left\{z_1(a,x) + \beta\phi(a',x')^\intercal\omega_1^* - \phi(a,x)^\intercal\omega_1^*\right\} \\ \vdots \\ P(s=K|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)\phi(a,x)\left\{z_K(a,x) + \beta\phi(a',x')^\intercal\omega_K^* - \phi(a,x)^\intercal\omega_K^*\right\} \\ P(s=1|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)r(a,x)\left\{\beta e_1(a',x') + \beta r(a',x')^\intercal\xi_1^* - r(a,x)^\intercal\xi_1^*\right\} \\ \vdots \\ P(s=K|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*)r(a,x)\left\{\beta e_K(a',x') + \beta r(a',x')^\intercal\xi_K^* - r(a,x)^\intercal\xi_K^*\right\} \\ P(s=1|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) - \pi_1^* \\ \vdots \\ P(s=K|\mathbf{a}, \mathbf{x}, \theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) - \pi_K^* \\ e_1(\bar{a},\bar{x}) - \gamma + \ln l_1(\bar{a},\bar{x};\theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) \ \forall\ (\bar{a},\bar{x}) \\ \vdots \\ e_K(\bar{a},\bar{x}) - \gamma + \ln l_K(\bar{a},\bar{x};\theta^*, \boldsymbol{\omega}^*, \boldsymbol{\xi}^*) \ \forall\ (\bar{a},\bar{x}) \end{bmatrix} = 0. \tag{B.6}$$

The moment conditions (B.4), (B.5) for (θ, π) are equivalent to the first order conditions obtained from maximizing the pseudo-likelihood (5.1) in the main text. Also, even when $h_k(x,a)$ and

$g_k(x,a)$ are not finite dimensional, the above holds as long as we let the dimensions k_ϕ, k_r of $\phi(a,x), r(a,x)$ grow to infinity.

The moment equation (B.6) fits into the class of models considered by Arcidiacono and Jones (2003); see, e.g their equation (11). Hence, we can follow Arcidiacono and Jones (2003) in applying the sequential EM algorithm to this model. Let $(\theta, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}}, \hat{\pi})$ denote the solution to the sample analogue of (B.6) after replacing $\mathbb{E}[\cdot]$ with its empirical counterpart $\mathbb{E}_n[\cdot]$. It is easy to verify that $(\theta, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\xi}}, \hat{\pi})$ also constitute a fixed point of the sequential EM algorithm. Unfortunately, as with all sequential EM algorithms, this analysis does not tell us whether the algorithm converges, or if it has the increasing likelihood property. Studying the convergence properties of sequential EM is an important avenue for future research.