arXiv:2008.00500v2 [cs.LG] 27 Nov 2020

Dynamic Discrete Choice Estimation with Partially Observable States and Hidden Dynamics

Yanling Chang, Alfredo Garcia and Zhide Wang
Texas A&M University

December 1, 2020

Abstract

Dynamic discrete choice models are used to estimate the intertemporal preferences of an agent as described by a reward function based upon observable histories of states and implemented actions. However, in many applications, such as reliability and healthcare, the system state is only partially observable or hidden (e.g., the level of deterioration of an engine, the condition of a disease), and the decision maker only has access to information imperfectly correlated with the true value of the hidden state. In this paper, we consider the estimation of a dynamic discrete choice model with state variables and system dynamics hidden to *both* the agent and the modeler, thus generalizing the model in [36] to partially observable cases. We examine the structural properties of the model and prove that this model is still identifiable if the cardinality of the state space, the discount factor, the distribution of random shocks, and the rewards for a given (reference) action are given. We analyze both theoretically and numerically the potential mis-specification errors that may be incurred when the Rust's model is improperly used in partially observable settings. We further apply the model to a subset of dataset in [36] for bus engine mileage and replacement decisions. The results show that our model can improve model fit as measured by the log-likelihood function by 17.73% and the log-likelihood ratio test shows that our model statistically outperforms the Rust's model. Interestingly, our hidden state model also reveals an economically meaningful route assignment behavior in the dataset which was hitherto ignored, i.e. routes with lower mileage are assigned to buses believed to be in worse condition.

1 Introduction

We consider the task of training a model of dynamic decisions by a single human agent based upon the history of implemented actions with hidden (or partially observable) states and system dynamics. Under the assumption of complete state observability, this problem has been widely studied in two strands in the literature where it is referred to as *inverse reinforcement learning* (IRL) or alternatively as *structural estimation*.

The bus engine replacement model developed in Rust's seminal paper [36] has served as a point of reference for this literature over decades. The model assumes that at each time period and for every bus in the fleet, Mr. Zurcher (a superintendent of maintenance of city of Madison's Metropolitan Bus Company) had to choose between replacing the engine or continuing to operate at a cost which included maintenance and loss of ridership in the case of a breakdown. The model assumes optimal engine replacement decisions are a function of accumulated mileage (observable by the researcher) and random cost perturbations (e.g. minor engine repairs) observed by Mr. Zurcher (not observable by the researcher). Hence, Mr. Zurcher's decisions are modeled as resulting from a Markov Decision Process (MDP) subject to independent and identically distributed (i.i.d.) cost perturbations.

Assuming cumulative mileage as sufficient statistic to identify optimal replacement decisions goes counter to much of the reliability literature on models of cumulative damage (see e.g., [39]). Cumulative mileage and other specialized tests only provide informative signals of the true underlying engine state. In fact, identifying a set of performance indicators closely related to the engine degradation state is a critical issue in predictive maintenance literature [42]. These observations suggest that a model of optimal replacement decisions based upon accumulated mileage and privately observed i.i.d cost perturbations is likely to be *misspecified*. Since cumulative mileage is likely not a sufficient statistic to identify optimal engine replacement policies, alternative model specifications should be considered. For example, there is a literature (e.g., [32, 6, 35, 2]) in which MDP-based models with serially correlated unobservables are examined.

In this paper we take a different approach for addressing the possibility that a MDP-based model with i.i.d shocks is misspecified. Rather than modeling Mr. Zurcher's decisions as resulting from a MDP with serially correlated cost perturbations, we model decisions as consistent with a Partially Observable Markov Decision Process (POMDP): Mr. Zurcher's engine replacement (and route assignment decisions) are a function of his evolving Bayesian *belief* on the level of *unobservable* damage or deterioration of each engine. Under this approach, serial correlation in cumulative mileage is an *endogenous* feature of the dynamic decision making process, e.g. in order to minimize maintenance costs, Mr. Zurcher could have conceivably assigned higher mileage routes to buses with engines he *believed* were in a better condition.

Making dynamic decisions when the relevant state variable is only partially observable is a general trait in many application domains. For example, in healthcare settings, the true physiological state of a patient, especially for cancers and chronic diseases [44], is only imperfectly known even with the most advanced testing technologies. In these cases, it is not reasonable to use a model relying on the assumption of the state being observable as it may inevitably incur model mis-specification errors.

We consider a dynamic discrete choice (DDC) model of an agent's decisions as resulting from a partially observable Markov decision process. At each stage, the agent (e.g., Mr. Zurcher) collects observations (e.g., mileage) that are imperfectly correlated with the state and makes decisions based on the *entire* history of his/her observations and actions subject to i.i.d random cost perturbations. Our objective is to develop a new estimation method for determining maximum likelihood estimates of the primitive parameters of the underlying controlled stochastic process, including the agent's reward structure and the system *hidden* dynamics.

We now summarize the main contributions of this paper. First, we show that the updated Bayesian belief on the hidden state is sufficient to characterize optimal dynamic choices. For a given choice of parameter estimates, the evolution of Bayesian beliefs (although latent) can be recursively computed by the modeler for each sample path. Secondly, we prove that a model is still identifiable although the state is not observable by the modeler, if the cardinality of the system state space, the distribution of random i.i.d shocks, the discount factor, and the reward for a reference action are known. Thirdly, when the dimensions of the state and observation sets are the same, we characterize model misspecification error when a Rust's MDP-based model is used to fit data that is generated according to a POMDP process. Finally, we apply the model to the widely studied Rust's engine replacement dataset. We show that our new estimation method can dramatically improve the data fit by 17.73% in terms of the log-likelihood. The log-likelihood ratio test also suggests that our approach performs significantly better than the Rust's model. Furthermore, the results from our model indicate that Mr. Zurcher was trying to optimize route assignments based on engine conditions. This is a feature of Mr. Zurcher's decisions that has been hitherto ignored.

Our work adds a new candidate to the collection of empirical models for dynamic discrete choices. MDP models with serially correlated unobservables and POMDP-based models are competing alternatives that should be empirically tested and compared. For example, a MDP-based model with correlated unobservables may still be rejected even after considering *higher* order models (via for example the Markov property testing procedure developed by [41]). In this case, the proposed POMDP-based approach developed in this paper constitutes an appealing alternative.

The paper is organized as follows. Section 2 provides an overview of the classical dynamic discrete choice model with observable states, which we will compare with and generalize. Section 3 presents our dynamic discrete choice model for partially observable states and hidden dynamics with model formulation in Section 3.1, a preview of estimation results in Section 3.2, and structural results in Section 3.3. The identification results are presented in Section 4. Section 5 applies our POMDP-based hidden state model to the bus engine replacement data in [36], and compares to the results from the Rust's model. In Section 6, we examine the deviation between the quantities that the modeler intends to obtain from what he/she may get when the Rust's MDP model is applied to the partially observable situations. This deviation represents the so called "specification error" in statistical analysis when the selected model poorly represents the underlying data generation process. Section 7 summarizes research results and discusses future research directions.

1.1 Literature Review

Dynamic Discrete Choice Models. The econometrics research on dynamic discrete choice models has been focused on (i) developing efficient algorithms to address computational challenges in estimating the structural parameters such as determining the value function used in the likelihood function estimation [20, 21, 3, 4, 45, 26]; (ii) relaxing restrictive assumptions on the unobservables to the modeler such as considering permanent heterogeneity [6, 5], serially correlated unobservables [7, 35], unobservables correlated across choices [27, 14], etc.; (iii) analyzing approximation errors, inference and validation [19, 23]; (iv) examining nonparametric and semiparametric identification issues and estimation [1, 29]. Dynamic discrete choice estimation methods have been widely used in applications including retirement behaviors [38], occupational choices and career decisions of young men [16], incentives to get teachers to work [15], adult women's mammography decisions [17], trade and labor markets [9], car ownership [12], etc. We remark that the "unobservables" mentioned in the literature specifically refers to the state components that are not observed by the modeler (but they are completely observable to the agent; e.g., see) [13, 25, 22]. On the contrary, the "hidden state" in this paper refers to the situations where the system state is not observable to *both* the agent and the modeler.

There is a strand of literature on DDC structural models to describe situations where agents have uncertainty about some primitives and use Bayesian learning to infer the value of unknowns. For example, [10] reviewed this stream of literature with particular emphasis on dynamic demand models where consumers learn over time about some product characteristics. This literature and our POMDP-based DDC model share some similarity. In fact, the existing Bayesian learning work can be viewed as a special case of our approach where the product characteristics are the internal hidden state *without dynamics*. Our POMDP-based model is general as it allows the internal hidden system state to transition over time according to a MDP.

Inverse Reinforcement Learning. A maximum entropy method proposed in [50] has been highly influential in the computer science literature. This method can be seen as an information theoretic derivation of the Rust's nested loop estimation [36]. Sample-based algorithms for implementing the maximum entropy method have scaled to scenarios with nonlinear reward functions (see e.g., [8, 18]). In [11], the authors extended the maximum entropy estimation method to a partially observable environment, assuming both the transition probabilities and observation probabilities are known with domain knowledge. These methods have also been used for apprenticeship learning where a robot learns from expert-based demonstrations [50]. To our best knowledge, none of these existing works have developed a general methodology to jointly estimate the reward structure and system dynamics based on trajectories of a POMDP.

Partially Observable Markov Decision Processes. POMDPs generalize MDPs by taking into account that in many situations the state is only partially observable to the decision-making agent due to measurement noise and/or limited access to information etc. Since the system state is hidden, the agent may need to consult the complete history (i.e. all past implemented actions and all received observations) in order to identify the optimal decision at a given point in time. Fortunately, for a Bayesian decision maker, the updated Bayes belief which is a probability distribution over all possible system states provides enough information to identify the optimal action. Thus, a POMDP can be seen as a MDP in which the state is the updated Bayes belief distribution ([43, 31, 47, 28] and many others).

Over decades, numerous exact and approximate POMDP algorithms have been developed, such as Sondik's one-pass algorithm ([43]), point-based POMDP algorithms (see a review in [40]), and POMDP-based models have been successfully applied in many real applications (interactive spoken dialog, see [49]). Since a given dataset can be generated by a MDP, high-order MDP, POMDP, or other non-Markovian processes, using MDP-based DDC models may not be always appropriate [41]. Hence, this paper is to leverage the POMDP results and to develop an alternative POMDP-based model for non-Markovian observables.

2 Dynamic Discrete Choice Estimation with Observable States

We now briefly review the problem of constructing a MDP-based model of dynamic decision making by a single human agent based upon the history of implemented actions and completely observed system states in [37].

At time $t \geq 0$, the human agent implements an action a_t from the action space A and receives a reward $r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t)$, where s_t is the system state at time t from the state space S, $r_{\theta_1}(s_t, a_t)$ is the parametrized reward associated to the state-action pair (s_t, a_t), $\theta_1 \in \mathbb{R}^{p_1}$ for some $p_1 \in N_+$, and $\epsilon_t(a_t)$ is a random perturbation. The cardinality of A is finite, $|A| < \infty$.

Upon implementing the action $a_t \in A$, the system state evolves according to a Markov process with parametrized conditional probabilities $P_{\theta_2}(s_{t+1}, \epsilon_{t+1}|s_t, \epsilon_t, a_t)$ where $\theta_2 \in \mathbb{R}^{p_2}$ for some $p_2 \in N_+$. The conditional independence assumption in [37] assumes:

$$P_{\theta_2}(s_{t+1}, \epsilon_{t+1}|s_t, \epsilon_t, a_t) = P(\epsilon_{t+1}|s_{t+1}) P_{\theta_2}(s_{t+1}|s_t, a_t).$$

In addition, the reward perturbation vectors $\{\epsilon_t \in \mathbb{R}^{|A|}, t > 0\}$ are assumed to be independently and identically distributed (i.i.d.) over time with probability measure μ. Thus, the decision process of the agent can be modelled as

$$V_\theta(s_t, \epsilon_t) = \max_{a_t \in A}[r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) + \beta \sum_{s_{t+1} \in S} \int V_\theta(s_{t+1}, \epsilon_{t+1}) P_{\theta_2}(s_{t+1}|s_t, a_t) d\mu(\epsilon_{t+1}|s_{t+1})], \quad (1)$$

where $\theta = (\theta_1, \theta_2)$, and $\beta \in [0, 1)$ is the discount factor. The above equation can be rewritten as

$$V_\theta(s_t, \epsilon_t) = \max_{a_t \in A}[r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) + \beta \sum_{s_{t+1} \in S} \bar{V}_\theta(s_{t+1}) P_{\theta_2}(s_{t+1}|s_t, a_t)] \quad (2)$$

where $\bar{V}_\theta(s_{t+1}) = \int V_\theta(s_{t+1}, \epsilon_{t+1}) d\mu(\epsilon_{t+1}|s_{t+1})$.

A (Markovian) model of the human agent's decisions is a function $\pi_\theta(a|s)$ (belonging to a dimension $|A| - 1$ simplex), which gives the probability that the human agent implements action a when the state is s. This function is also referred to as the conditional choice probability (CCP) function in the literature. When the distribution of the random perturbations $\{\epsilon_t(a), t \geq 0\}$ are i.i.d. standard Gumbel for all $a \in A$, the CCP's are of the form:

$$\pi_\theta(a|s) = \frac{\exp Q_\theta(s, a)}{\sum_{a' \in A} \exp Q_\theta(s, a')} \quad (3)$$

where

$$Q_\theta(s, a) = r_{\theta_1}(s, a) + \beta \sum_{s' \in S} \bar{V}_\theta(s') P_{\theta_2}(s'|s, a) \quad (4)$$

$$\bar{V}_\theta(s) = \gamma + \log \sum_{a \in A} \exp(Q_\theta(s, a)), \quad (5)$$

and $\gamma > 0$ is the Euler's constant. For data in the form of a collection of $N > 0$ independent finite sequences of state-action pairs $\{(s_{t,i}, a_{t,i}), 0 \leq t \leq T\}_{i=1}^{N}$, a likelihood function is defined as:

$$\log \ell(\theta) \triangleq \log \prod_{i=0}^{N} \prod_{t=0}^{T} \pi_\theta(a_{t,i}|s_{t,i}) P_{\theta_2}(s_{t+1,i}|s_{t,i}, a_{t,i})$$
$$= \sum_{i=1}^{N} \left(\sum_{t=0}^{T} \log \pi_\theta(a_{t,i}|s_{t,i}) + \sum_{t=0}^{T-1} \log P_{\theta_2}(s_{t+1,i}|s_{t,i}, a_{t,i}) \right) \tag{6}$$

A model of the agent can be obtained by finding the parameter $\theta^* \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ that maximizes the log-likelihood in (6).

3 Dynamic Discrete Choice for Hidden States and Dynamics

This section generalizes the Rust's framework in [37] for the completely observable case to the partially observable case.

3.1 A POMDP-based Model Formulation

At each decision epoch $t \geq 0$, the value of state s_t is not directly observable to the human agent nor to the external modeler. Both the human agent and the external modeler are able to receive the value of a *public* random variable $z_t \in Z$ correlated with the underlying state s_t, assuming the cardinality of the observation space Z is finite. As in the Rust's model, we assume the human agent observes a private signal (or reward perturbation) $\epsilon_t(a_t)$ when implementing action $a_t \in A$. If the hidden state is s_t, the reward accrued is $r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t)$ where $\theta_1 \in \mathbb{R}^{p_1}$ for some $p_1 \in N_+$. The system dynamics is described by probabilities $P_{\theta_2}(z_{t+1}, \epsilon_{t+1}, s_{t+1}|s_t, a_t)$ where $\theta_2 \in \mathbb{R}^{p_2}$ for some $p_2 \in \mathbb{N}_+$; see Figure 1 for a schematic representation.

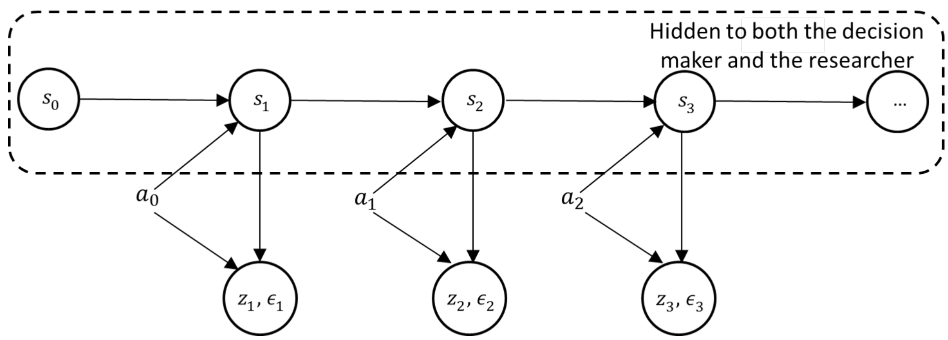


Figure 1: Graphical illustration of the proposed dynamic discrete choice model with partially observable states and hidden dynamics. At each stage, z_t is observed by both the human agent and the modeler, and ϵ_t is privately observed by the human agent. The system state is s_t is hidden to both the agent and the modeler with its own hidden dynamics.

Let $\zeta_t = \{z_t, ..., z_1, a_{t-1}, ..., a_0, x_0\}$ be the publicly received history of the dynamic decision process including all past and present revealed observations and all past actions at time $t > 0$, where $x_0 = \{P(s_0), s_0 \in S\}$ is the prior belief vector over S, assuming $|S| < \infty$. At time t, the objective of the agent is to maximize

$$E\left(\sum_{\tau=0}^{\infty} \beta^\tau r(s_{t+\tau}, a_{t+\tau})|\zeta_t, \epsilon_1, ..., \epsilon_t\right)$$

where the reward structure $r(s_t, a_t)$ is determined by the hidden system state s_t (not z_t) and action a_t. Similar to [37], we assume

Additively Separable (AS) Rewards:

$$\sum_{s_t} P_{\theta_2}(s_t|\zeta_t, \epsilon_t) r_{\theta_1}(s_t, a_t) = \sum_{s_t} P_{\theta_2}(s_t|\zeta_t) r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t); \tag{7}$$

Conditional Independence (CI) Assumption:

$$P_{\theta_2}(z_{t+1}, \epsilon_{t+1}|\zeta_t, \epsilon_t, a_t) = P(\epsilon_{t+1}|z_{t+1})P_{\theta_2}(z_{t+1}|\zeta_t, a_t). \tag{8}$$

Note that different from the CI Assumption in [36], $\{\zeta_t, \epsilon_t\}$ is not Markovian. However, z_{t+1} is a sufficient statistic for ϵ_{t+1}, indicating ϵ_t and ϵ_{t+1} are independent given z_{t+1} and

$$P_{\theta_2}(z_{t+1}, \epsilon_{t+1}, s_{t+1}|s_t, a_t) = P(\epsilon_{t+1}|z_{t+1})P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t). \tag{9}$$

In addition, the conditional probability $P_{\theta_2}(z_{t+1}|\zeta_t, a_t)$ does not depend on ϵ_t.
Under Assumptions AS and CI, the decision making process of the agent can be modelled as

$$\begin{aligned} V_{t,\theta}(\zeta_t, \epsilon_t) = \max_{a_t \in A} \Big\{ &\sum_{s_t} P_{\theta_2}(s_t|\zeta_t) r_{\theta_1}(s_t, a_t) + \epsilon_t(a_t) \\ &+ \beta \sum_{z_{t+1}} \int P_{\theta_2}(z_{t+1}|\zeta_t, a_t) V_{t+1,\theta}(\zeta_{t+1}, \epsilon_{t+1}) d\mu(\epsilon_{t+1}|z_{t+1}) \Big\}, \end{aligned} \tag{10}$$

which can be viewed as a POMDP with perturbations $\{\epsilon_t : t \geq 0\}$ and $\{P_{\theta_2}(s_t|\zeta_t) : t \geq 0\}$ is commonly called the belief over system state s_t, given the history at stage t. We thus explain the random shock as follows. If the modeler knew the values of (ζ_t, ϵ_t), then the modeler could also replicate the solution of the POMDP in (10). Because of this noise, the modeler does not know exactly the true value of belief $\{P_{\theta_2}(s_t|\zeta_t, \epsilon_t)\}$, resulting in a perturbation to the one-step reward.
The objective of the modeler is to identify θ_1 in the reward structure $r_{\theta_1}(s_t, a_t)$, and θ_2 in the dynamics $P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)$ from the publicly received histories $\{\zeta_T^i\}_{i=1}^N$. Consequently, the underlying hidden state transition probabilities can be determined by

$$P_{\theta_2}(s_{t+1}|s_t, a_t) = \sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t), \tag{11}$$

and the observation probabilities $O_{\theta_2}(z_{t+1}|s_{t+1}, s_t, a_t)$ can also be obtained from

$$O_{\theta_2}(z_{t+1}|s_{t+1}, s_t, a_t) = \frac{P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)}{\sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t)}, \tag{12}$$

assuming $\sum_{z_{t+1}} P_{\theta_2}(z_{t+1}, s_{t+1}|s_t, a_t) > 0$.

3.2 Estimation: A Preview of Results

The main challenge of solving Eq. (10) as basis for computing a maximum likelihood estimator is that the cardinality of the history ζ_t grows to infinity as t increases. Let $x_{t,\theta_2}(s)$ denote the Bayesian updated belief for a fixed value of the parameter choice θ_2, i.e. $x_{t,\theta_2}(s) \triangleq P_{\theta_2}(s_t = s|\zeta_t), t > 0$ where we assume $x_0 = \{P(s_0) : s_0 \in S\}$ is given. Note that

$$x_{t,\theta_2} \in X \triangleq \{x \in R^{|S|} : x(s) \geq 0, s \in S, \sum_{s \in S} x(s) = 1\} \subset \mathbb{R}^{|S|},$$

and X is finite dimensional and time invariant. Building upon [43, 47], and many others, we show in Theorem 1 that the updated Bayesian belief distribution is sufficient to identify the optimal dynamic choices. For a fixed value of the parameter choice θ_2, the belief process $\{x_{t,\theta_2} : t > 0\}$ is a controlled Markov process where $x_{t+1,\theta_2} = \lambda_{\theta_2}(z_{t+1}, x_{t,\theta_2}, a_t)$ and

$$\lambda_{\theta_2}(z_{t+1}, x_{t,\theta_2}, a_t) \triangleq \frac{x_{t,\theta_2} P_{\theta_2}(z_{t+1}, a_t)}{\sigma_{\theta_2}(z_{t+1}, x_{t,\theta_2}, a_t)}, \quad (\sigma_{\theta_2}(z_{t+1}, x_{t,\theta_2}, a_t) \neq 0) \tag{13}$$

$$\sigma_{\theta_2}(z_{t+1}, x_{t,\theta_2}, a_t) \triangleq \sum_{s'} \sum_{s} x_{t,\theta_2}(s) P_{\theta_2}(z_{t+1}, s_{t+1} = s'|s_t = s, a_t), \tag{14}$$

and $[x_{t,\theta_2} P_{\theta_2}(z_{t+1}, a_t)]_{s_{t+1}} \triangleq \sum_s x_{t,\theta_2}(s) P_{\theta_2}(z_{t+1}, s_{t+1} | s_t = s, a_t)$.
In what follows we will show that the solution to (10) are conditional choice probabilities of the form:

$$\pi_\theta(a|x) = \frac{\exp Q_\theta(x,a)}{\sum_{a' \in A} \exp Q_\theta(x,a')},$$

where

$$Q_\theta(x,a) = r_\theta(x,a) + \beta \sum_z \sigma_{\theta_2}(z,x,a) V_\theta(x') \quad V_\theta(x') = \gamma + \log \sum_{a'} \exp(Q_\theta(x',a')),$$

and $r_\theta(x,a) \triangleq \sum_s x_{\theta_2}(s) r_{\theta_1}(s,a)$, $x' = \lambda_{\theta_2}(z,x,a)$ and $\gamma > 0$ is Euler's constant.
Given data corresponding to $N \geq 1$ finite histories of pairs $\{(z_{t,i}, a_{t,i})\}_{t=0}^T$ for $i \in \{1, \dots, N\}$, a sequence of trajectories for the belief $\{x_{t,\theta_2,i} : t > 0\}$ can be recursively computed for a fixed value of $\theta = (\theta_1, \theta_2)$ as follows:

$$x_{t+1,\theta_2,i} = \lambda_{\theta_2}(z_{t+1,i}, x_{t,\theta_2,i}, a_{t,i}) = \frac{x_{t,\theta_2,i} P_{\theta_2}(z_{t+1,i}, a_{t,i})}{\sigma_{\theta_2}(z_{t+1,i}, x_{t,\theta_2,i}, a_{t,i})},$$

Thus, the log-likelihood can be written as:

$$\begin{aligned}
\log \ell(\theta) &\triangleq \log \prod_{i=1}^N P(\zeta_{T,i} | x_{0,i}) = \log \prod_{i=1}^N \prod_{t=0}^{T-1} P(z_{t+1,i} | \zeta_{t,i}, a_{t,i}) P(a_{t,i} | \zeta_{t,i}) \\
&= \sum_{i=1}^N \sum_{t=0}^{T-1} \log \sigma_{\theta_2}(z_{t+1,i}, x_{t,\theta_2,i}, a_{t,i}) + \sum_{i=1}^N \sum_{t=0}^{T-1} \log \pi_\theta(a_{t,i} | x_{t,\theta_2,i}). \qquad (15)
\end{aligned}$$

That is, assuming the data is generated by a Bayesian decision maker controlling a partially observable Markov process, the external modeler can construct a POMDP-based DDC model by finding the parameter values that maximize the log-likelihood in (15).

3.3 Structural Results

The results in this section are obtained for a fixed value of θ. Thus, to alleviate notation we drop allusion to the fixed parameter everywhere. The next theorem leverages results in the POMDP literature to show that the belief process $\{x_t, t \geq 0\}$ is a sufficient statistic identifying for optimal dynamic choices.

Theorem 1. *Let $V_t(x_t, \epsilon_t)$ be the solution to:*

$$V_t(x_t, \epsilon_t) = \max_{a_t \in A} \left\{ r(x_t, a_t) + \epsilon_t(a_t) + \beta \sum_{z_{t+1}} \int \sigma(z_{t+1}, x_t, a_t) V_{t+1}(\lambda(z_{t+1}, x_t, a_t), \epsilon_{t+1}) d\mu(\epsilon_{t+1} | z_{t+1}) \right\}. \qquad (16)$$

It follows that $V_t(\zeta_t, \epsilon_t) = V_t(x_t, \epsilon_t)$.

Proof: The proof is by induction. Assume $V_{t+1}(\zeta_{t+1}, \epsilon_{t+1}) = V_{t+1}(x_{t+1}, \epsilon_{t+1})$, then

$$\begin{aligned}
V_t(\zeta_t, \epsilon_t) &= \max_{a_t \in A} \left\{ \sum_{s_t} P(s_t | \zeta_t) r(s_t, a_t) + \epsilon_t(a_t) + \beta \sum_{z_{t+1}} \int P(z_{t+1} | \zeta_t, a_t) V_{t+1}(x_{t+1}, \epsilon_{t+1}) d\mu(\epsilon_{t+1} | z_{t+1}) \right\} \\
&= \max_{a_t \in A} \left\{ r(x_t, a_t) + \epsilon_t(a_t) + \beta \sum_{z_{t+1}} \int \sigma(z_{t+1}, x_t, a_t) V_{t+1}(\lambda(z_{t+1}, x_t, a_t), \epsilon_{t+1}) d\mu(\epsilon_{t+1} | z_{t+1}) \right\} \\
&= V_t(x_t, \epsilon_t)
\end{aligned}$$

where the second equality follows from $\sigma(z_{t+1}, x_t, a_t) = P(z_{t+1} | \zeta_t, a_t)$.

Define

$$Q_t(x_t, a_t) = r(x_t, a_t) + \beta \sum_{z_{t+1}} \int \sigma(z_{t+1}, x_t, a_t) V_{t+1}(\lambda(z_{t+1}, x_t, a_t), \epsilon_{t+1}) d\mu(\epsilon_{t+1}|z_{t+1}). \tag{17}$$

Then,

$$V_t(x_t, \epsilon_t) = \max_{a_t \in A} \{Q_t(x_t, a_t) + \epsilon_t(a_t)\}. \tag{18}$$

We extend the concept of social surplus function in [30] and [37] by

$$G[\{u(x_t, a_t), a_t \in A\}|x_t, z_t] = \int \max_{a_t \in A} [u(x_t, a_t) + \epsilon_t(a_t)] d\mu(\epsilon_t|z_t), \tag{19}$$

for a measurable function $u : X \times A \to R$. Then by the dominated convergence theorem and probability theory, we have the following result (analogous to Theorem 3.1 in [37]).

Theorem 2. *If $\mu(d\epsilon_t|z_t)$ has finite first moments, then the social surplus function (19) exists and it is*

$$G[\{u(x_t, a_t), a_t \in A\}|x_t, z_t] = \sum_{a_t \in A} \pi(a_t|x_t)(u(x_t, a_t) + E[\epsilon_t|Y(u, \epsilon) = a_t, z_t]), \tag{20}$$

where $Y(u, \epsilon) = \arg\max_{a_t \in A}(u(x_t, a_t) + \epsilon_t(a_t))$. Furthermore,

(i) G is a convex function of $\{u(x, a), a \in A\}$;

(ii) G satisfies the additivity property, i.e., for any $\alpha \in R$,

$$G[\{u(x_t, a_t) + \alpha, a_t \in A\}|x_t, z_t] = \alpha + G[\{u(x_t, a_t), a_t \in A\}|x_t, z_t]; \tag{21}$$

(iii) The partial derivative of G with respect to $u(x_t, a_t)$ is the conditional choice probability:

$$\frac{\partial G[\{u(x_t, a_t), a_t \in A\}|x_t, z_t]}{\partial u(x_t, a_t)} = \pi(a_t|x_t). \tag{22}$$

Let $\mathscr{B}$ be the Banach space of bounded, Borel measurable functions $H : X \times A \to R$ under the supremum norm. Define an operator $H : \mathscr{B} \to \mathscr{B}$ by

$$[Hv](x, a) = r(x, a) + \beta \sum_{z'} \sigma(z', x, a) G[\{v(\lambda(z', x, a), a'), a' \in A\}|\lambda(z', x, a), z']. \tag{23}$$

Assume the following regularity conditions (similar to [37] with state replaced by belief x):

(i) (Bounded Upper Semicontinuous) For each $a \in A$, $r(x, a)$ is upper semicontinuous in belief x with bounded expectation and

$$\begin{aligned}
h(x) &:= \sum_{t=1}^{\infty} \beta^t h_t(x) < \infty, \\
h_1(x) &= \max_{a \in A} \sum_{z' \in Z} \sigma(z', x, a) \int \max_{a' \in A} |r(\lambda(z', x, a), a') + \epsilon'(a')| d\mu(\epsilon'|z'), \\
h_t(x) &= \max_{a \in A} \sum_{z' \in Z} \sigma(z', x, a) h_{t-1}(\lambda(z', x, a));
\end{aligned}$$

(ii) (Weakly Continuous) The stochastic kernel $\sigma(\cdot, x, a) = \{\sigma(z, x, a)\}_{z \in |Z|}$ is weakly continuous in $X \times A$;

(iii) (Bounded Expectation) The reward $r \in \mathscr{B}$ and for each $v \in \mathscr{B}$, $Ev \in \mathscr{B}$, where

$$[Ev](x, a) = \sum_{z' \in Z} \sigma(z', x, a) G[\{v(\lambda(z', x, a), a'), a' \in A\}|\lambda(z', x, a), z'].$$

Theorem 3. *Under AS, CI, and the regularity conditions, $H : \mathscr{B} \to \mathscr{B}$ is a contraction mapping with modulus β. Hence, H has a unique fixed point satisfying $Q^* = HQ^*$ and the optimal decision rule δ^* is*

$$\delta^*(x, \epsilon) = \arg\max_{a \in A} \left\{ Q^*(x, a) + \epsilon(a) \right\}. \tag{24}$$

Furthermore, the controlled process $\{z_{t+1}, x_t, a_t\}$ is Markovian with

$$\pi(a|x) = \frac{\partial G[\{Q^*(x, a), a \in A\}|x, z]}{\partial Q^*(x, a)}. \tag{25}$$

Theorem 4. *If the probability measure of ϵ is multivariate extreme-value, i.e.,*

$$\mu(d\epsilon|z) = \prod_{a \in A} \exp\{-\epsilon(a) + \gamma\} \exp[-\exp\{-\epsilon(a) + \gamma\}], \tag{26}$$

where γ is the Euler constant. Then, the agent will select its action a with probability

$$\pi(a|x) = \frac{\exp Q(x, a)}{\sum_{a' \in A} \exp Q(x, a')}, \tag{27}$$

where

$$Q(x, a) = r(x, a) + \beta \sum_{z'} \sigma(z', x, a) \log(\sum_{a'} \exp Q(\lambda(x, z', a), a')).$$

Let

$$V(x, \epsilon) = \max_{a \in A} \{Q(x, a) + \epsilon(a)\}. \tag{28}$$

Proposition 1. *Functions of $Q(x, a)$ and $V(x, \epsilon)$ are convex in x.*

For example, under the extreme-value assumption, $Q(x, a)$ in Theorem 4 is convex on X because $f(x) = \log \sum_i \exp(x_i)$ is a convex function (by Hölder's inequality).

Under the extreme-value assumption, it is clear to see per Theorem 4 that

$$\log\left(\frac{\pi(a|x)}{\pi(a^*|x)}\right) = Q(x, a) - Q(x, a^*). \tag{29}$$

for a fixed reference action $a^* \in A$. Thus, the *Independence from Irrelevant Alternatives* (IIA) property still holds with respect to belief (rather than the hidden state), i.e., the odds of choosing alternative a over the reference action $a^* \in A$ only depends on the attributes of the two choices. The results in Theorems 1-4 allow us to formulate the maximum likelihood estimation problem as defined in (15).

Remark 1. It can be easily verified that Theorems 3-4 and Proposition 1 continue to hold for the case in which the agent is solving a finite horizon problem. Evidently, the results in this case require that δ, CCP, Q, V all depend on t.

Remark 2. Our model is fundamentally different from the existing DDCs with unobserved states (see e.g., [13, 22, 25] and many others). In these papers, an agent's dynamic decisions are assumed to be a function s_t and ϵ_t where both the agent and the modeler can observe s_t but the utility perturbations ϵ_t can only be observed by the agent. In contrast, in our model, the payoff-relevant state s_t is hidden to both the decision-making agent and the modeler and s_t also has its own hidden dynamics. The agent is assumed to use Bayesian update to infer the state s_t based upon his/her entire information history. Table 1 summarizes the main differences between our POMDP-based model and the existing MDP-based DDC models.

	Existing DDC Models (including DDCs with unobserved states)	Our Model
Agent makes decision and generates data via	(high order) MDP	POMDP
Payoff-relevant State	fully observable to the agent but partially observable to the modeler	hidden to both the agent and the modeler
Meaning of "unobservables"	the state component hidden to the modeler but observable to the agent	payoff-relevant state

Table 1: The key differences between our model and the existing MDP-based DDC models.

4 Identification Results

In this section we generalize the identification results in [29] for the Rust's model to our hidden-state model. The structure of the POMDP-based DDC model is defined by parameters :

$$b = \{r(S,A), P(Z,S,A), \mu, \beta\}$$

where $r(S,A) \equiv \{r(s,a) : s \in S, a \in A\}$, $P(Z,S,A) \equiv \{P(z,s'|s,a), z \in Z, s,s' \in S, a \in A\}$. The conditional observation probabilities $\{\sigma_b(z_{t+1}, x_{t,b}, a_t)\}$ and conditional choice probabilities $\{\pi_b(a_t|x_{t,b})\}$ are called reduced form observation probabilities and choice probabilities under structure b, respectively. Under the true structure b_0, we must have

$$\forall (z_{t+1}, \zeta_t, a_t), \qquad \underbrace{\hat{P}(z_{t+1}|\zeta_t, a_t)}_{\text{Data}} = \underbrace{\sigma_{b_0}(z_{t+1}, x_{t,b_0}, a_t)}_{\text{Model}}, \tag{30}$$

$$\underbrace{\hat{P}(a_t|\zeta_t)}_{\text{Data}} = \underbrace{\pi_{b_0}(a_t|x_{t,b_0})}_{\text{Model}}. \tag{31}$$

where $\hat{P}(z_{t+1}|\zeta_t, a_t)$ and $\hat{P}(a_t|\zeta_t)$ are functions of the data.
We now restate the definitions of observational equivalence and identification as in [29].

Definition 1 (Observational equivalence)**.** *Let B be the set of structures b, and let $\overset{o}{\Longleftrightarrow}$ be observational equivalence. $\forall b, b' \in B$, $b \overset{o}{\Longleftrightarrow} b'$ if and only if $\sigma_b(z_{t+1}, x_{t,b}, a_t) = \sigma_{b'}(z_{t+1}, x_{t,b'}, a_t)$ and $\pi_b(a_t|x_{t,b}) = \pi_{b'}(a_t|x_{t,b'}), \forall z_{t+1}, a_t$.*

Definition 2 (Identification)**.** *The model is identified if and only if $\forall b, b' \in B$, $b \overset{o}{\Longleftrightarrow} b'$ implies $b = b'$.*

A distinct feature of our model from the MDP-based models is that the dynamics of the system under study cannot be directly observed as the system state is only partially observable. However, the next theorem shows that we could identify the hidden dynamics using two periods of data, assuming we know the cardinality of the state space of the system.

Theorem 5. *Assume $|S|$ is known. The hidden dynamic $P(Z,S,A)$ (not rank-1) can be uniquely identified from the first two periods of data* [1].

Theorem 5 allows us to generalize the identification results in [20] and [29] for the completely observable case to the partially observable case, by realizing that our model is a belief version of the Rust's model once the hidden dynamics is determined.

Theorem 6. *The hidden state dynamic discrete choice model is not identified. However, $\{r(S,a), a \in A \backslash a^*\}$ and $P(Z,S,A)$ can be uniquely identified from the data, given the cardinality of the state space $|S|$, the discount factor β, the distribution of random shock μ, and the rewards $r(s,a^*), s \in S$ for a reference action $a^* \in A$ are all known.*

For example, under the Gumbel assumption, it is clear to see that

$$\bar{V}(x) = E_\epsilon[V(x,\epsilon)] = \log \sum_{a\in A} \exp\left(Q(x,a) - Q(x,a^*)\right) + Q(x,a^*). \tag{32}$$

Note that

$$Q(x,a^*) = r(x,a^*) + \beta E_z[\bar{V}(\lambda(x,z,a^*))|x,a^*] \tag{33}$$

Combining Eq. (32) and Eq. (33),

$$\begin{aligned} Q(x,a^*) = r(x,a^*) &+ \beta E_z[\log \sum_{a\in A} \exp\left(Q(\lambda(x,z,a^*),a) - Q(\lambda(x,z,a^*),a^*)\right)|x,a^*] \\ &+ \beta E_z[Q(\lambda(x,z,a^*),a^*))|x,a^*], \end{aligned}$$

By Eq. (29) we have

$$Q(x,a^*) = r(x,a^*) + C + \beta E_z[Q(\lambda(x,z,a^*),a^*))|x,a^*], \tag{34}$$

where

$$C = \beta E_z\Big[\log \sum_{a\in A} \frac{\pi(a|\lambda(x,z,a^*))}{\pi(a^*|\lambda(x,z,a^*))}|x,a^*\Big],$$

and it can be obtained from the dataset per Theorem 1 and Theorem 5. It is easy to show that Eq. (34) is a contraction mapping; hence, there is a unique solution for $Q(x,a^*)$, for a fixed given $r(x,a^*)$. Thus, $Q(x,a)$ can be identified by Eq. (29) for all $a\in A$ and consequently $\bar{V}(x)$ as well. Next, since

$$r(x,a) = Q(x,a) - \beta E_z[\bar{V}(\lambda(x,z,a))|x,a]$$

we can get $r(x,a)$, and consequently, $r(s,a)$.
It is an interesting future research question to examine whether or under what conditions the POMDP-based model is identifiable if $|S|$ is unknown. In practice, the number of possible states can be obtained by domain knowledge for a particular application. For example, the possible stages of a cancer or system degradation are likely obtainable. A practitioner can also try possible values of $|S|$ to examine which value can best explain the observed behaviors.

Corollary 1. *For $T<\infty$, the hidden model can be uniquely identified from the data, if μ and both the reward structure and the terminal value function Q_T in the reference action $a^*\in A$ are all known.*

Finally, we remark that the approach to identification based upon a reference action might pose issues to the identification of counterfactuals as in the MDP-based models (see [24]). Detailed discussions on counterfactual identification for the POMDP-based approach is a future research topic.

5 Rust's Model Revisited

To illustrate the application of the proposed methodology, we revisit a subset of Rust's dataset for Mr. Zurcher's engine replacement decisions in [36]. Specifically, Group 4 consisting of buses with 1975 GMC engines. As in [36], we discretize the state space in 175 bins of 2.5K miles. Evidence of positive serial correlation in mileage increments is quite strong as the Durbin-Watson statistic is less than 1.13 for all buses except one with a value of 1.32. The action $a_t=1$ is associated with engine replacement at a cost $RC>0$ whereas $a_t=0$ is the continued operation. The state space S is $S=\{0,1\}$, where "0" is being the "good state" and "1" is being the "bad state". The model for hidden state dynamics is

$$P_{\theta_2}(s_{t+1}|s_t,a_t=0) = \begin{pmatrix} \theta_{2,0} & 1-\theta_{2,0} \\ 1-\theta_{2,1} & \theta_{2,1} \end{pmatrix}$$

and $P_{\theta_2}(s_{t+1}=0|s_t,a_t=1)=1$. Per mile maintenance costs are parametrized by $\theta_{1,0}$ (in good state) and $\theta_{1,1}$ (in bad state). With a belief $x_t\in[0,1]$ of the engine being in good state and z_t cumulative mileage after t months, the expected (monthly) maintenance cost is of the form $(\theta_{1,0}z_t)x_t + (\theta_{1,1}z_t)(1-x_t)$. Monthly mileage increments $\Delta \in \{0,1,2,3\}$ correspond to values between $[0, 2.5K)$, $[2.5K, 5K)$, $[5K, 7.5K)$ and $[7.5K, 10K)$, respectively. The distribution is parametrized as follows

$$
\begin{aligned}
&P_{\theta_3}(z_{t+1} - z_t = \Delta |\, s_t = 0, a_t = 0) = \theta_{3,0,\Delta}, && \Delta \in \{0,1,2\} \\
&P_{\theta_3}(z_{t+1} - z_t = \Delta |\, s_t = 0, a_t = 0) = 1 - \theta_{3,0,0} - \theta_{3,0,1} - \theta_{3,0,2}, && \Delta = 3.
\end{aligned}
$$

Similarly, we define $P_{\theta_3}(z_{t+1} - z_t = \Delta |\, s_t = 1, a_t = 0) = \theta_{3,1,\Delta}, \Delta \in \{0,1,2,3\}$. The estimation results are described in Table 2 where the belief space (i.e. the unit interval) is discretized in a uniform grid of 100 intervals.

Parameter	$\theta_{3,0,0}$	$\theta_{3,0,1}$	$\theta_{3,0,2}$	$\theta_{3,1,0}$	$\theta_{3,1,1}$	$\theta_{3,1,2}$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{1,0}$	$\theta_{1,1}$	RC
Good State	0.040 (.005)	0.335 (.017)	0.587 (.017)	*	*	*	0.949 (.004)	*	0.0009 (.0003)	*	10.073 (.725)
Bad State	*	*	*	0.182 (.008)	0.757 (.008)	0.060 (.006)	* *	0.988 (.002)	*	0.0012 (.0002)	
Log-Likelihood	-3818.853										

Table 2: Parameter estimates and log-likelihood of our hidden state model (standard errors obtained by bootstrapping method are in parentheses)

Compared to the original Rust's model displayed in Table 3, the hidden state model captures an optimal route assignment: the distribution of mileage increments for engines considered in bad state is dominated (in the first-order stochastic sense) by the distribution of mileage increments of engines in good state. Assigning routes with lower mileages to buses in bad state decreases the operational cost of buses in such state. We find that the marginal operation costs (per mile) for buses in *bad* vs. *good* state differ slightly (i.e. the difference is approximately equal to the standard deviation). This is consistent with the optimal route assignment in the sense that if for example $\theta_{1,0} << \theta_{1,1}$ then buses in good condition would be *under*-utilized vis-á-vis those in bad condition for which replacement is not yet justified. This is an economically meaningful feature of Mr. Zurcher's behavior (ignored by Rust's model) which improves model fit as measured by log-likelihood by $\frac{4495-3818}{3818} = 17.73\%$. We remark that MDP-based DDC model with AR(1) unobservables was also studied on the Rust data set in [35]; however, no improvement in the goodness of fit was observed in that study.

Parameter	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$	θ_1	RC
Rust's Model ([36], p. 1022) (Standard errors in parentheses)	0.119 (0.005)	0.576 (0.008)	0.287 (0.007)	0.016 (0.002)	0.0012 (0.0003)	10.90 (1.581)
Log-Likelihood	-4495					

Table 3: Parameter estimates and log-likelihood Rust's model

The significant increment in the log-likelihood function is achieved at the cost of 5 more parameters in dynamics and rewards. Considering that the Rust's model is a special case of our model (where the agent has perfect observability of the system state), we use the log-likelihood ratio test to examine whether our model statistically outperforms the Rust's model in this example. The underlying null hypothesis is H0: there is no significant difference between the Rust's model and our model. The log-likelihood ratio test result in Table 4 shows that the null hypothesis is rejected with a p value very close to zero, indicating that our model statistically outperforms the Rust's model on the Group 4 of the 1975 GMC bus engine dataset.

	Log-likelihood	Degree of Freedom	Statistic	p Value
Hidden state model	-3819	11	1352	$< 10^{-100}$
Rust's model	-4495	6		

Table 4: The log-Likelihood ratio test for both models on the 1975 GMC engine (Group 4) data set

6 Model Misspecification Errors

For a given dataset, possible modeling options include MDP-based DDC models (with serial correlations), POMDP-based models, or other non-Markovian models. Recently, [41] have developed a model selection procedure based on the Markov assumption, because MDP-based models may not always fit the data and estimation results can be misleading if the model is mis-specified. In this section, we examine how far the estimations may deviate from their true values if the existing MDP-based Rust's model is used to fit data generated by an agent making decisions according to a POMDP process. We first analyze the potential specification errors on the system dynamics and the reward structure, assuming $|S| = |Z|$. Then, we use synthetic data to further numerically illustrate this issue in Section 6.1.

As in the previous section, all results below are obtained for a fixed value of θ. Thus, to alleviate notation, we drop allusion to the fixed parameter everywhere. Also, we define $y1 \triangleq \sum_s y(s), \forall y \in R^{|S|}$, so that we can write $\sigma(z,x,a) = xP(z,a)1 = \sum_{s'}\sum_s x(s)P(z,s'|s,a)$, where $P(z,a) \triangleq \{P_{ij}(z,a)\}, P_{ij}(z,a) \triangleq P(z_{t+1} = z, s_{t+1} = j|s_t = i, a_t = a)$.

To examine how the accuracy of z_t reflecting the true state s_t affects the estimation results, the main difference between a MDP-based model and a POMDP-based model, we assume the observation probability $O(z'|s',s,a)$ is independent of s,a and the analytical relation between z_t and s_t is described by

$$O_\eta(z|s) = \begin{cases} 1-\eta, & z = s, \\ \kappa_{s,z}\eta, & z \neq s \end{cases} \tag{35}$$

where $0 \leq \eta \leq 1, \kappa_{s,z} \geq 0, \kappa_{s,s} = 0, \sum_z \kappa_{s,z} = 1, \forall z \in Z, s \in S$ [33]. The parameter η represents the observation accuracy (or observation quality): the observation z_t is the true system state s_t with probability $1-\eta$. We say the observation accuracy of O_η is better than $O_{\eta'}$ if $\eta < \eta'$.
Define

$$\begin{aligned} d(x,x') &= 1 - \min\left\{\frac{x(s)}{x'(s)} : x'(s) > 0\right\}, \\ D_1(x,x') &= \max\{d(x,x'), d(x',x)\}, \\ D(z,a) &= \max\{D_1(\lambda(z,e_i,a), \lambda(z,e_j,a)) : i,j \in S\}, \\ D(z,a,a') &= \max\{D_1(\lambda(z,e_i,a), \lambda(z,e_j,a')) : i,j \in S\}. \end{aligned} \tag{36}$$

We remark that $D(z,a)$ is a contraction coefficient (coefficient of ergodicity) for $P(z,a)$ [48, 34]. For a given sample path ζ_t, let $x_* = x_t = \lambda(z,\tilde{a}_*,\tilde{x}_*)$, $\tilde{x}_* = x_{t-1}$, $\tilde{a}_* = a_{t-1}$ along sample ζ_t for short notation.

Theorem 7. *Assume a sample path ζ_t is given.*

$$\begin{aligned} &P(z' = i|z = k,a) - P(s' = i|s = k,a) \\ &\geq \max_{j\in S}\left[\lambda(k,e_j,\tilde{a}_*) - \lambda(k,\tilde{x}_*,\tilde{a}_*)\right]P(i,a)1 + \sum_{\tilde{s}} x_*(\tilde{s})[P(s' = i|s = \tilde{s},a) - P(s' = i|s = k,a)] \\ &+ \eta\sum_{\tilde{s}}\sum_{s''} x_*(\tilde{s})[\kappa_{s'',z=i}P(s' = s''|s = \tilde{s},a) - P(s' = i|s = \tilde{s},a)]. \end{aligned}$$

Moreover,

$$D_1(P(.|z=k,a), P(.|s=k,a)) \le \max_{\tilde{a}\in A} D(k,\tilde{a},\tilde{a}_*)$$
$$+\max\left\{D_1\left[P(\cdot|s=k,a), \sum_{\tilde{s}} x_*(\tilde{s})P(.|\tilde{s},a)\right], D_1\left[P(\cdot|s=k,a), \sum_{\tilde{s}}\sum_{s''} x_*(\tilde{s})\kappa_{s'',z=.}P(s''|\tilde{s},a)\right]\right\},$$

where e_j is the unit vector in $R^{|S|}$ with the j^{th} element being 1 and all other elements being 0.

Corollary 2. *When $\eta=0$, $P(z'|z,a)=P(s'|s,a), \forall z'=s', s=z, z,z'\in Z, s,s'\in S$.*

For the reward structure, the Rust's model shows $\log\frac{\pi(a|z)}{\pi(a=0|z)} = Q^R(z,a)-Q^R(z,a=0)$, whereas in our hidden state model $\log\frac{\pi(a|x)}{\pi(a=0|x)} = Q(x,a)-Q(x,a=0)$. Thus, we can analyze how the CCP ratios measured by the two models affect the estimated reward. Namely, given

$$\rho_*(z,a) \le |\log\frac{\pi(a|z)}{\pi(a=0|z)} - \log\frac{\pi(a|x_*)}{\pi(a=0|x_*)}| \le \rho^*(z,a), \tag{37}$$

where the minimal and maximal differences (ρ_*,ρ^*) are caused by ignoring history ζ_t leading to the current observation z, we examine how r^R may deviate from the true reward structure r. As only the relative difference between the choice-specific function Q matters, we let $Q^R(Z,a=0)=Q(X,a=0)=0$. Define $\triangle r^R \triangleq \max_{z,z',a}|r^R(z,a)-r^R(z',a)|$, and

$$||\sigma-P^R||_1^* \triangleq \max_{z,a}\frac{1}{2}\sum_{z'}|[\sigma(z',a,x_*)-P(z'|z,a)]| = \max_{z,a,\tilde{x},\tilde{a}}\frac{1}{2}\sum_{z'}|[\sigma(z',a,x_*)-\sigma(z',a,\lambda(z,\tilde{x},\tilde{a}))]|.$$

Theorem 8 below shows that both the deviation in the CCPs and the system dynamic will contribute to the estimation error in the reward structure, where $||.||$ is the infinity norm.

Theorem 8. *Given a sample path ζ_t,*

$$[||r-r^R||+\triangle r^R]+2\beta[||\sigma-P^R||_1^*(||Q^R||+\log|A|)+\frac{1}{2}\log|A|] \ge (1-\beta)\max_{z,a}\rho_*(z,a).$$

Furthermore,

$$||r^R-r|| \le (1+\beta)\max_{z,a}\rho^*(z,a)+2\beta||\sigma-P^R||_1^*||Q^R||+\beta\max_{z,a,x_*}|h(z,a,x_*)|,$$

where $h: X\times A\times Z\to R$ is a continuous function satisfying

(i) $h(z,a,x_*)\to 0$ *if the observation probability $O\to I$, I is the identity matrix,*

(ii) $\max_{z,a,x_*}|h(z,a,x_*)| \le \log|A|+\max_{z,a,a'}|Q^R(z,a)-Q^R(z,a')|$.

Theorem 7, Corollary 2 and Theorem 8 show that the deviation between the estimated values and the true values will increase as the observation quality decreases (i.e., η increases); on the other side, if η is sufficiently small, then the specification error will go to zero (i.e., the existing MDP-based Rust's model is a special case of the POMDP-based model where the observation is perfectly accurate); also see Fig. 2. Hence, when fitting a dataset, a practitioner could use the existing MDP-based Rust's model if the observation quality is sufficiently high; however, the potential model mis-specification can be significant if the observation is only a noisy signal of the true system state; please also see related discussion in [41].

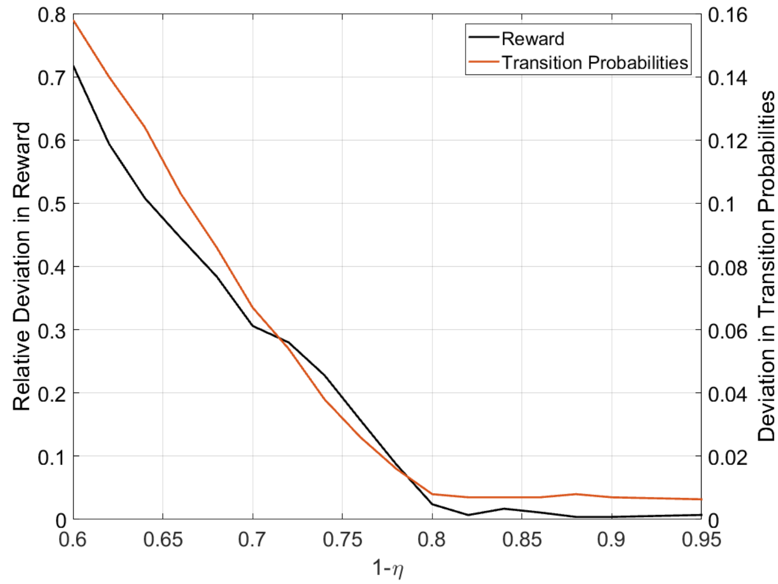


Figure 2: The deviation between the estimated values and the true values as a function of $1-\eta$ (the value of η becomes smaller to the right side, indicating higher observation quality). The graph is plotted using randomly generated synthetic datasets as in Section 6.1

6.1 Numerical Illustrations via Synthetic Data

We randomly generate synthetic data sets where both the system state and dynamics are hidden, in order to test if our approach can correctly identify the reward structure $r_{\theta_1}(s,a)$ and the hidden dynamics $P_{\theta_2}(z',s'|s,a)$, assume the discount factor $\beta(=0.95)$ is known and and $r(a=0)$ is fixed. The belief space is uniformly discretized to 100 intervals over $[0,1]$. Table 5 lists the parameter values (values without parentheses) of an example in the size of $|S|=2, |Z|=2, |A|=3$ with $\eta=0.1$, where the values fitted by our model are included in the parentheses ($r(a=0)=[10,3]$ is fixed and given during the estimation). The result shows that our model can identify the model parameters fairly well, with maximal element-wise deviation of 0.026 in dynamics and 0.08 in reward, using 800 sample trajectories.

We also fit the same data set using the Rust's model in Table 6. The true transition probabilities $P(s'|s,a)$ are obtained by $P(s'|s,a)=\sum_{z'}P(z',s'|s,a)$ and their values are listed in the table without parentheses. The estimation result for the Rust's model (POMDP model) in the parentheses (brackets) shows that the maximal element-wise deviation are 0.275 (0.016) and 2.06 (0.08) in transition probabilities and reward, respectively. Thus, our POMDP model performs better than the Rust's model in the partially observable case. Furthermore, the chi-squared statistic of the log-likelihood ratio test in Table 7 is 1124.46 with degree of freedom 12, rejecting the null hypothesis of no significant difference between the two models with a very small p value close to zero.

In addition, Fig. 3 illustrates the distributions of log-likelihood functions fitted by both models under various randomly generated synthetic data sets. If the data is generated in a completely observable environment, the developed hidden state model will generate the same result as what are estimated by the Rust's model; see Fig. 3(a)(b). However, if the data is generated in a partially observable environment, the log-likelihood function fitted by our model ℓ is significantly greater than the maximal log-likelihood function produced by the Rust's model ℓ^{Rust}; see Fig. 3(c)(d). In fact, $\ell \geq \ell^{Rust}$ almost surely as the Rust's model is a special case of our model where the observation matrix is identity. Furthermore, the likelihood ratio test rejects the Rust's model with high confidence intervals and results generated by the Rust's model can

$a=0$	(z',s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.72 (0.724)	0.08 (0.083)	0.02 (0.013)	0.18 (0.180)
$s=1$	0.00 (0.000)	0.00 (0.000)	0.10 (0.120)	0.90 (0.880)

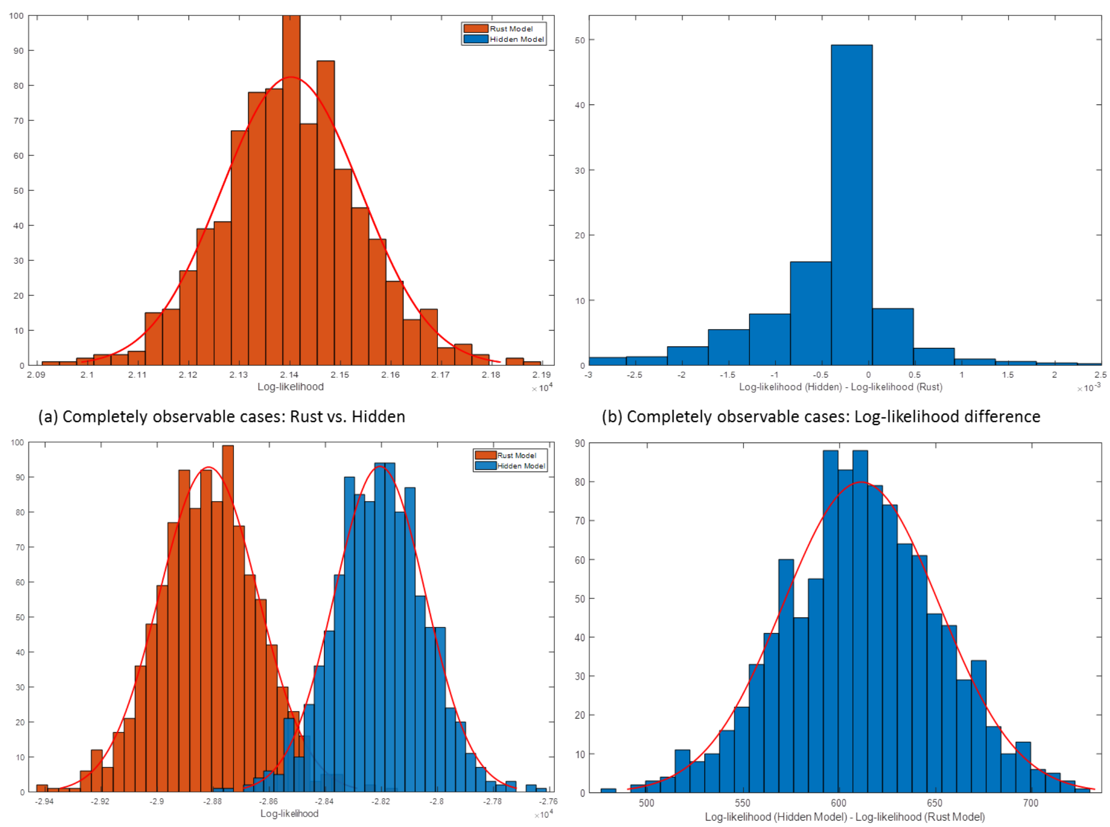
$a=1$	(z',s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.81 (0.784)	0.09 (0.100)	0.01 (0.000)	0.09 (0.116)
$s=1$	0.00 (0.002)	0.00 (0.000)	0.10 (0.118)	0.90 (0.880)

$a=2$	(z',s')			
	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$s=0$	0.90 (0.909)	0.10 (0.090)	0.00 (0.000)	0.00 (0.001)
$s=1$	0.36 (0.350)	0.04 (0.043)	0.06 (0.068)	0.54 (0.539)

$r(s,a)$	a		
	0	1	2
$s=0$	10^*	6 (6.08)	3 (2.95)
$s=1$	3^*	5 (4.99)	7 (6.99)

Table 5: Model parameters of a randomly generated numerical example ($\eta = 0.1$) and the values fitted by our model (in the parentheses). The maximal element-wise difference by the POMDP-based model: 0.026 in dynamics $\{P(z',s'|s,a)\}$, 0.08 in reward $\{r(s,a)\}$.

be misleading and/or erroneous.



(a) Completely observable cases: Rust vs. Hidden

(b) Completely observable cases: Log-likelihood difference

(c) Partially observable cases: Rust vs. Hidden

(d) Partially observable cases: Log-likelihood difference

Figure 3: The log-likelihood distributions of both the Rust's model and our model for a thousand synthetic datasets

$a=0$	$s'=0$	$s'=1$
$s=0$	0.80 (0.704) [0.807]	0.20 (0.296) [0.193]
$s=1$	0.00 (0.275) [0.000]	1.00 (0.725) [1.000]

$a=1$	$s'=0$	$s'=1$
$s=0$	0.90 (0.735) [0.884]	0.10 (0.265) [0.116]
$s=0$	0.00 (0.263) [0.002]	1.00 (0.737) [0.998]

$a=2$	$s'=0$	$s'=1$
$s=0$	1.00 (0.836) [0.999]	0.00 (0.164) [0.001]
$s=1$	0.40 (0.500) [0.393]	0.60 (0.500) [0.607]

$r(s,a)$	a		
	0	1	2
$s=0$	10^*	6 (6.64) [6.08]	3 (4.49) [2.95]
$s=1$	3^*	5 (3.75) [4.99]	7 (4.94) [6.99]

Table 6: Parameter values fitted by the Rust's model in parentheses vs. by POMDP-based model in brackets, $\eta = 0.1$. The maximal element-wise difference by the Rust's model (POMDP model): 0.275 (0.016) in transition probabilities $\{P(s'|s,a)\}$, 2.06 (0.08) in reward $\{r(s,a)\}$.

	Log-likelihood	Degree of Freedom	Statistic	p Value
Hidden state model	-27642.11	22	1124.46	$< 10^{-10}$
Rust's model	-28204.34	10		

Table 7: Log-Likelihood ratio test for our hidden model and the Rust's model.

7 Conclusions

In this paper, we developed a dynamic discrete choice model for the case where the underlying system state and the associated system dynamics are hidden for the decision-making agent. At each decision epoch, the agent infers the system state from possibly noise-corrupted observations on which to base action selection. We formulated the decision making process of the agent on the basis of partially observable Markov decision processes subject to independent and identically distributed random shocks, generalizing the existing dynamic discrete choice models to partially observable settings. We analyzed the structural properties of the proposed hidden state model and proved that the model is still identifiable from sample trajectories if the discount factor, the distribution of the random shock, the reward structure in a reference action, and the possible number of hidden states are known. As the (completely observable) dynamic discrete choice models are widely used in problems where the relevant state variables may only be partially observed, we analyzed the potential model misspecification errors when the Rust's model is used in a partially observable setting. The possible estimation discrepancies were also demonstrated via numerical examples. Finally, we compared our hidden state model to the Rust's model on the bus engine dataset in [36]. The significant improvement in the log-likelihood function and the log-likelihood ratio test strongly suggested that our hidden state model outperforms the Rust's model. Furthermore, our model also revealed economically meaningful features of Mr. Zurcher's behavior ignored by the Rust's model.

As this research represents a first effort on developing dynamic discrete choice models for partially observable systems, future research directions are numerous. For example, our current model assumes that the distribution of the random shock is i.i.d.. Many studies have examined more complicated structures in completely observable dynamic choice models. Thus, extending these analyses to the partially observable case will be value-added. Also, computational challenges of our model are obviously not trivial. Both continuous-state MDP and POMDP suffer from the well-known curse of dimensionality, meaning that to achieve a certain level of accuracy, the number of discretized points have to grow exponentially. Moreover, observations in many real applications can be high dimensional. Thus, another research direction is to address computational challenges of high dimensional hidden state models. Application wise, it will be interesting to apply this model to various relevant applications, indicatively, building predictive models of human control when a relevant psychological trait (e.g., fatigue, attention) is hidden.

Appendix

[1]The proof of Theorem 5 essentially implies that $P_{\theta_2}(z,a)$ can be obtained by $\sigma_{\theta_2}(z, \lambda_{\theta_2}(z_{t+1}, x_t, a_t), a_{t+1}) = \hat{P}(z|\zeta_{t+1}, a_{t+1})$, assuming we know the belief at time t. In many applications, it is reasonable to expect that we could find such belief points. For example, in the engine replacement example, it is acceptable to reason that the state of a newly replace engine is good.

7.1 Structural Results

Proof. Proof of Theorem 2. Eq. (20) is directly from the definition of G and

$$G[\{u(x_t,a_t), a_t \in A\}|x_t, z_t] = \int \sum_{a_t \in A} [u(x_t,a_t) + \epsilon_t(a_t)] 1_{\{Y=a\}} d\mu(\epsilon_t|z_t).$$

$(i) - (iii)$ follow the same idea as in [37]. □

Proof. Proof of Theorem 3. Under the modified regularity conditions, the proof follows the exactly same procedure as in [37] for Theorems 3.2-3.3. The controlled process $\{z_{t+1}, x_t, a_t\}$ is Markovian because the conditional probability of a_t is $P(a_t|x_t)$, the conditional probability of x_{t+1} is provided by $\lambda(z_{t+1}, x_t, a_t)$, and the conditional probability of z_{t+1} is given by $\sigma(z_{t+1}, x_t, a_t)$. □

Proof. Proof of Theorem 4. The result follows by Theorem 3 and [46]. □

Proof. Proof of Proposition 1. The result is obvious by induction, $\sigma \geq 0$, the maximal of convex functions is still convex, and Theorem 2(i). □

7.2 Identification Results

Proof. Proof of Theorem 5. For any given two system dynamics $P_1(z,a)$, $P_2(z,a)$, where $P(z,a) = \{P_{ij}(z,a)\}, P_{ij}(z,a) = P(z_{t+1} = z, s_{t+1} = j | s_t = i, a_t = a)$, we show that they can be distinguished by the data. Since the dataset contains x_0, a_0 and $|S|$ is known, we can obtain $\sigma_0^1(z_1, x_0, a_0) = \sum_{s_1} \sum_s x_0(s) P_1(z_1, s_1|s, a_0) = \sum_s x_0(s) P_1(z_1|s, a_0)$ and $\sigma_0^2(z_1, x_0, a_0) = \sum_{s_1} \sum_s x_0(s) P_2(z_1, s_1|s, a_0) = \sum_s x_0(s) P_2(z_1|s, a_0)$. Note that $\sigma_0(z_1, x_0, a_0) = \hat{P}(z_1|\zeta_0, a_0)$, where $\hat{P}(z_1|\zeta_0, a_0)$ is a function of the first period data. Thus, σ can be obtained from the data and $\sigma_0^1 = \sigma_0^2$ if and only if $P_1(z|s,a) = P_2(z|s,a)$. If $\sigma_0^1 \neq \sigma_0^2$, we are done. However, it is possible that there exists s' such that $P_1(z, s'|s, a) \neq P_2(z, s'|s, a)$ but $P_1(z|s,a) = P_2(z|s,a)$. In this case, update belief by Eq. (13), $x_1^1 = \lambda_1(z_1, x_0, a_0) = \frac{x_0 P_1(z_1, a_0)}{\sigma_0^1(z_1, x_0, a_0)}, x_1^2 = \lambda_2(z_1, x_0, a_0) = \frac{x_0 P_2(z_1, a_0)}{\sigma_0^2(z_1, x_0, a_0)}$. Then $x_1^1 = x_1^2$ if and only if $P_1(z,a) = P_2(z,a)$. Now, $\sigma_1^1(z_2, x_1^1, a_1) = \sum_s x_1^1(s) P_1(z_2|s, a_1)$ and $\sigma_1^2(z_2, x_1^2, a_1) = \sum_s x_1^2(s) P_2(z_2|s, a_1)$, and again σ_1 is obtainable from the two periods of data as $\sigma_1(z_2, x_1, a_1) = \hat{P}(z_2|\zeta_1, a_1)$. Now, $\sigma_1^1 = \sigma_1^2$ if and only if $x_1^1 = x_1^2$, indicating $P_1(z,a) = P_2(z,a)$ assuming $P(z,a)$ is not rank-1. □

Proof. Proof of Theorem 6. By Theorem 1 and Theorem 5, both CCP $\pi(a|x_t) = \hat{P}(a|\zeta_t)$ and hidden dynamics can be identified from the data. Treating belief as the state and since $|\zeta_t| < \infty$, Proposition 1 in [20] and [29] show that there is a one-to-one mapping $q(X): R^{|A|} \to R^{|A|}$, only depending on μ, which maps the choice probability set $\{\pi(a|x)\}$ to the set of the difference in action-specific value function $\{Q(x,a) - Q(x,a^*)\}$, namely,

$$Q(x,a) - Q(x,a^*) = q_a(\{\pi(a'|x)\};\mu), \tag{38}$$

$q_0(.) = 0$, and $q = (q_0, ..., q_{|A|-1})$. Thus, if we know $Q(x,a^*)$, we can recover $Q(x,a), \forall a \in A\backslash a^*$. Note that

$$\begin{aligned}
Q(x,a) &= r(x,a) + \beta E_{z|x,a}\left[E_{\epsilon|z} \max_{a'\in A}\left\{Q(\lambda(z,x,a),a') + \epsilon'(a')\right\}\right] \\
&= r(x,a) + \beta E_{z|x,a}\left[E_{\epsilon|z}\Big[\max_{a'\in A}\left\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),a^*) + \epsilon'(a')\right\}\Big]\right] \\
&+ \beta E_{z|x,a}\left[Q(\lambda(z,x,a),a^*)\right] \\
&= r(x,a) + \beta E_{z|x,a}\Big[G\big[\left\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),a^*), a' \in A\right\} |\lambda(x,z,a), z\big]\Big] \\
&+ \beta E_{z|x,a}\left[Q(\lambda(z,x,a),a^*)\right]
\end{aligned} \tag{39}$$

Because of the mapping q, and that the hidden dynamic $\{P(z,a)\}$ is known, the quantity

$$\beta E_{z|x,a}\Big[G\big[\left\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),a^*), a' \in A\right\} |\lambda(x,z,a), z\big]\Big]$$

is known. Under the assumption of $r(s,a^*) = 0$, we have

$$\begin{aligned}
Q(x,a^*) &= \beta E_{z|x,a^*}\Big[G\big[\left\{Q(\lambda(z,x,a^*),a') - Q(\lambda(z,x,a^*),a^*), a' \in A\right\} |\lambda(x,z,a^*), z\big]\Big] \\
&+ \beta E_{z|x,a^*}\left[Q(\lambda(z,x,a^*),a^*)\right]
\end{aligned} \tag{40}$$

with only unknown $Q(X,a^*)$. It is easy to see there is a unique solution to Eq. (40) due to the contraction mapping theorem. Consequently, all $Q(X,a), a \in A$ can be recovered by Eq. (38). Lastly,

$$\begin{aligned}
r(x,a) &= Q(x,a) \\
&- \beta E_{z|x,a}\Big[G\big[\left\{Q(\lambda(z,x,a),a') - Q(\lambda(z,x,a),a^*), a' \in A\right\} |\lambda(x,z,a), z\big] - Q(\lambda(z,x,a),a^*)\Big].
\end{aligned}$$

As $r(x,a) = \sum_{s\in S} x(s)r(s,a)$, $\{r(s,a) : s \in S, a \in A\}$ can be uniquely determined. □

Proof. Proof of Corollary 1. When $Q_T(X,a^*)$ and $r(s,a^*)$ are known, we can obtain $Q_t(X,a^*)$ via

$$\begin{aligned}
Q_t(x,a^*) &= r(x,a^*) + E_z\Big[G\big[\left\{Q(\lambda(z,x,a^*),a') - Q(\lambda(z,x,a^*),a^*), a' \in A\right\} |\lambda(x,z,a^*), z\big]\Big] \\
&+ E_z\left[Q_{t+1}(\lambda(z,x,a^*),a^*)\right].
\end{aligned} \tag{41}$$

The rest follows exactly as in the proof of Theorem 6. □

7.3 Model Misspecification Errors

The proof of Theorem 7 needs Propositions 2-4 in the following.
In a partially observable setting, we first observe that the dynamic $P(z'|z,a)$ estimated by the Rust's model is in fact a random variable, as its value depends on past histories ending with $z_t = z$, i.e., $\{\zeta_t : z_t = z\}$. That is, given z', z, a, $P(z'|z,a)$ can be any value in

$$\{P(z_{t+1} = z'|z_t = z, a_t = a, \tilde{\zeta}_{t-1}, \tilde{a}_{t-1}) : \forall\tilde{\zeta}_{t-1}, \forall\tilde{a}_{t-1} \in A\} = \{\sigma(z', a, \lambda(z,\tilde{x},\tilde{a})) : \tilde{x} \in X, \tilde{a} \in A\}.$$

For a sample path ζ_t, Proposition 2 first quantifies a *lower bound* on the maximal difference between $P(z'|z,a)$ estimated in the Rust's MDP-based model and its counterpart $\sigma(z',a,x_*)$ in the POMDP-based model, where $x_* = \lambda(z,\tilde{a}_*,\tilde{x}_*)$ is the current belief x_t at ζ_t (with $z_t = z$), and $\tilde{x}_*, \tilde{a}_*$ are the previous period belief x_{t-1} and action a_{t-1} along the history ζ_t. This difference is caused by (improperly) neglecting the information history leading to the current stage (i.e., serial correlations).

Proposition 2.

$$P(z'|z,a) - \sigma(z',a,x_*) \geq \max_{j\in S} \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1 \geq 0, \tag{42}$$

where e_j is the unit vector in $R^{|S|}$ with the j^{th} element being 1 and all other elements being 0.

We remark that the selection of $j^* \in \arg\max \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1$ depends on z'. Hence, Proposition 2 also implies that

$$||P(\cdot|z,a) - \sigma(\cdot,a,x_*)|| \geq \max_{z'\in Z} \left\{ \max_{j\in S} \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1 \right\}, \tag{43}$$

where $||v|| = \max_{z\in Z} |v(z)|$ for any vector $v \in R^{|Z|}$, $P(\cdot|z,a) = \{P(z'|z,a)\}_{z'\in Z}$, and $\sigma(\cdot,a,x_*) = \{\sigma(z',a,x_*)\}_{z'\in Z}$.

Proof. Proof of Proposition 2.

$$\begin{aligned} P(z'|z,a) - \sigma(z',a,x_*) &= \max_{\tilde{x}\in X,\tilde{a}\in A} \sigma(z',\lambda(z,\tilde{x},\tilde{a}),a) - \sigma(z',\lambda(z,\tilde{x}_*,\tilde{a}_*),a) \\ &\geq \max_{j\in S} \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1. \end{aligned}$$

Assume by contradiction that

$$\max_{j\in S} \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1 < 0.$$

Then

$$\lambda(z,e_j,\tilde{a}_*)P(z',a)1 < \lambda(z,\tilde{x}_*,\tilde{a}_*)P(z',a)1, \forall j \in S,$$

which is

$$\frac{e_j P(z,\tilde{a}_*)P(z',a)1}{e_j P(z,\tilde{a}_*)1} < \frac{\tilde{x}_* P(z,\tilde{a}_*)P(z',a)1}{\tilde{x}_* P(z,\tilde{a}_*)1}, \forall j \in S.$$

Note that $\tilde{x}_* = \sum_j \tilde{x}_{*,j} e_j$, thus

$$\frac{\tilde{x}_* P(z,\tilde{a}_*)P(z',a)1}{\tilde{x}_* P(z,\tilde{a}_*)1} = \frac{\sum_j \tilde{x}_{*,j} e_j P(z,\tilde{a}_*)P(z',a)1}{\sum_j \tilde{x}_{*,j} e_j P(z,\tilde{a}_*)1}$$

Recall that if $a,b,c,d > 0$, then $a/b \geq c/d \Leftrightarrow a/b \geq (a+c)/(b+d)$. Thus, we have

$$\frac{\sum_j \tilde{x}_{*,j} e_j P(z,\tilde{a}_*)P(z',a)1}{\sum_j \tilde{x}_{*,j} e_j P(z,\tilde{a}_*)1} \leq \max_{j\in S} \frac{e_j P(z,\tilde{a}_*)P(z',a)1}{e_j P(z,\tilde{a}_*)1},$$

a contradiction. Hence, $\max_{j\in S} \left[\lambda(z,e_j,\tilde{a}_*) - \lambda(z,\tilde{x}_*,\tilde{a}_*)\right] P(z',a)1 \geq 0$. □

Lemma 1.

$$\min_z \{ \frac{\sigma(z,a,x)}{\sigma(z,a,x')} : \sigma(z,a,x') > 0 \} \geq \min_s \{ \frac{x(s)}{x'(s)} : x'(s) > 0 \} \tag{44}$$

$$\max_z \{ \frac{\sigma(z,a,x)}{\sigma(z,a,x')} : \sigma(z,a,x') > 0 \} \leq \max_s \{ \frac{x(s)}{x'(s)} : x'(s) > 0 \} \tag{45}$$

Proof. Proof of Lemma 1.

$$\frac{\sigma(z,x,a)}{\sigma(z,x',a)} = \sum_s \frac{\sum_{s'} P(z,s'|s,a)x(s)}{\sigma(z,x',a)} = \sum_s \frac{\sum_{s'} P(z,s'|s,a)x'(s)}{\sigma(z,x',a)} \frac{x(s)}{x'(s)}$$

Note that $f(s) = \frac{\sum_{s'} P(z,s'|s,a)x'(s)}{\sigma(z,x',a)} \geq 0$ and $\sum_s f(s) = \sum_s \frac{\sum_{s'} P(z,s'|s,a)x'(s)}{\sigma(z,x',a)} = \frac{\sigma(z,x',a)}{\sigma(z,x',a)} = 1$. Thus,

$$\frac{\sigma(z,x,a)}{\sigma(z,x',a)} \geq \min_s \frac{x(s)}{x'(s)}, \forall z$$

leading to inequality (44). The proof for inequality (45) is similar. □

Proposition 3 provides an *upper bound* on the potential difference between $P(z'|z,a)$ and $\sigma(z',x_*,a)$.

Proposition 3. *For any* $\tilde{a} \in A$,

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) \leq D(z,\tilde{a},\tilde{a}_*), \forall \tilde{x} \in X;$$

thus, if $\tilde{a} = \tilde{a}_*$,

$$D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) \leq D(z,\tilde{a}_*), \forall \tilde{x} \in X.$$

Proposition 3 shows that the difference is bounded above by the ergodicity of $P(z,a)$ if $\tilde{a} = \tilde{a}_*$. If $\tilde{a} \neq \tilde{a}_*$, then the discrepancy caused by the two different actions $\tilde{a}$ and $\tilde{a}_*$ will also contribute to the upper bound.

Proof. Proof of Proposition 3. Note that

$$\begin{aligned}&D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,x_*)) = D_1(\sigma(\cdot,a,\lambda(z,\tilde{x},\tilde{a})),\sigma(\cdot,a,\lambda(z,\tilde{x}_*,\tilde{a}_*)))\\ &\leq D_1(\lambda(z,\tilde{x},\tilde{a}),\lambda(z,\tilde{x}_*,\tilde{a}_*)) \leq D(z,\tilde{a},\tilde{a}_*)\end{aligned}$$

where the second to the last inequality is by Lemma 1 in this Appendix, and the last inequality is from the fact that $D_1(x,\rho x' + (1-\rho)x'') \leq \max\{D_1(x,x'), D_1(x,x'')\}, x, x', x'' \in X, 0 \leq \rho \leq 1$,(A.3 in [34]) and $\lambda(z,x,a) = \sum_i \frac{x_i[e_i P(z,a)1]}{[xP(z,a)1]}\lambda(z,e_i,a)$ in A.4 [34]. $\square$

Proposition 4. *Given* $x_t = x, a_t = a, s_t = s$,

$$\begin{aligned}&\sigma(z'=i,x,a) - P(s'=i|s,a)\\ &= \sum_{\tilde{s}} x(\tilde{s})[P(i|\tilde{s},a) - P(i|s,a)] + \eta \sum_{\tilde{s}} x(\tilde{s})[\sum_{s''} \kappa_{s'',z=i} P(s''|\tilde{s},a) - P(i|\tilde{s},a)].\end{aligned}$$

Moreover,

$$D_1[\sigma(\cdot,x,a),P(\cdot|s,a)] = D_1[P(\cdot|s,a),(1-\eta)\sum_{\tilde{s}} x(\tilde{s})P(.|\tilde{s},a) + \eta \sum_{\tilde{s}} x(\tilde{s}) \sum_{s''} \kappa_{s'',z=.} P(s''|\tilde{s},a)],$$

and for $0 < \eta < 1$,

$$\begin{aligned}&D_1[\sigma(\cdot,x,a),P(\cdot|s,a)]\\ &\leq \max\left\{D_1\left[P(\cdot|s,a),\sum_{\tilde{s}} x(\tilde{s})P(.|\tilde{s},a)\right], D_1\left[P(\cdot|s,a),\sum_{\tilde{s}}\sum_{s''} x(\tilde{s})\kappa_{s'',z=.}P(s''|\tilde{s},a)\right]\right\}.\end{aligned}$$

Thus, Proposition 4 indicates that even if $x = e_s$,

$$\sigma(z'=i,x,a) - P(s'=i|s,a) = \eta \sum_{s''} \kappa_{s'',z=i}[P(s''|s,a) - P(i|s,a)] \neq 0.$$

Proof. Proof of Proposition 4. For the first part,

$$\begin{aligned}&\sigma(z'=i,x,a) - P(s'=i|s,a)\\ &= \sum_{\tilde{s}}\sum_{s''} P(z'=i|s'')P(s''|\tilde{s},a)x(\tilde{s}) - P(i|s,a)\\ &= \sum_{\tilde{s}}[P(i|\tilde{s},a) - P(i|s,a)]x(\tilde{s}) + \eta \sum_{\tilde{s}}(\sum_{s''} \kappa_{s'',z=i}P(s''|\tilde{s},a) - P(i|\tilde{s},a))x(\tilde{s}),\end{aligned}$$

by the definition of O. For the second part, note that $\sigma(z'=i,x,a)$ can also be written as

$$\sigma(z'=i,x,a) = (1-\eta)\sum_{\tilde{s}} P(i|\tilde{s},a)x(\tilde{s}) + \eta \sum_{\tilde{s}}\sum_{s''} \kappa_{s'',z=i}P(s''|\tilde{s},a)x(\tilde{s}).$$

Fix a, let $\pi^1(i) = \sum_{\tilde{s}} P(i|\tilde{s},a)x(\tilde{s})$ and $\pi^2(i) = \sum_{\tilde{s}}\sum_{s''} \kappa_{s'',z=i}P(s''|\tilde{s},a)x(\tilde{s})$, then $\pi^1, \pi^2 \in X$. The result follows from (A.3) in [34]. $\square$

Proof. Proof of Theorem 7. Note that $P(z' = i|z = k, a) - P(s' = i|s = k, a) = P(z' = i|z = k, a) - \sigma(z' = i, a, \lambda(z = k, \tilde{x}_*, \tilde{a}_*)) + \sigma(z' = i, a, \lambda(z = k, \tilde{x}_*, \tilde{a}_*)) - P(s' = i|s = k, a)$. The results follow by Proposition 2 and Proposition 4 in the Appendix. In addition, since D_1 is a metric on X, we have

$$\begin{aligned}
&D_1(\sigma(\cdot, a, \lambda(z = k, \tilde{x}, \tilde{a})), P(.|s = k, a)) \\
&\leq D_1[\sigma(\cdot, a, \lambda(z = k, \tilde{x}, \tilde{a})), \sigma(\cdot, x_*, a)] + D_1[\sigma(\cdot, x_*, a), P(\cdot|s = k, a)].
\end{aligned}$$

Hence, Propositions 3 - 4 lead to the result. □

Proof. Proof of Corollary 2. If $\eta = 0$, then $x = e_s$, $D_1[\sigma(\cdot, a, \lambda(z, \tilde{x}, \tilde{a})), \sigma(\cdot, x_*, a)] = 0$, and $D_1[\sigma(\cdot, x_*, a), P(\cdot|s, a)] = 0$. Hence, $D_1[\sigma(\cdot, a, \lambda(z, \tilde{x}, \tilde{a}), P(.|s, a)] = 0$, implying the result. □

Proof. Proof of Theorem 8.

$$\begin{aligned}
\rho_*(z, a) \leq\ & |Q(x_*, a) - Q^R(z, a)| \\
\leq\ & |r(x_*, a) - r^R(z, a)| \\
& + \beta|\sum_{z'} \sigma(z', a, x_*) \log(\sum_{a'} \exp(Q(\lambda(z', x_*, a), a')) - \sum_{z'} P(z'|z, a) \log(\sum_{a'} \exp(Q^R(z', a')))| \\
\leq\ & |x_* r(a) - x_* r^R(a) + x_* r^R(a) - r^R(z, a)| \\
& + \beta|\sum_{z'} \sigma(z', a, x_*)[\log(\sum_{a'} \exp(Q(\lambda(z', x_*, a), a')) - \log(\sum_{a'} \exp(Q^R(z', a')))]| \\
& + \beta|\sum_{z'}[\sigma(z', a, x_*) - P(z'|z, a)] \log(\sum_{a'} \exp(Q^R(z', a')))| \\
\leq\ & ||r - r^R|| + \triangle r^R \\
& + \beta \max_{z', a'}(|Q(\lambda(z', x_*, a), a')) - Q^R(z', a')|) + \beta \log|A| + 2\beta||\sigma - P^R||_1^*(||Q^R|| + \log|A|) \\
\leq\ & [||r - r^R|| + \triangle r^R](1 + \beta) \\
& + \beta^2 \max_{z', a'} \max_{z'', a''}(|Q(\lambda(z'', a', \lambda(z', a, x_*)), a'')) - Q^R(z'', a'')|) + (\beta + \beta^2) \log|A| \\
& + 2(\beta + \beta^2)||\sigma - P^R||_1^*(||Q^R|| + \log|A|) \\
\leq\ & [||r - r^R|| + \triangle r^R]\frac{1}{1-\beta} + \frac{\beta}{1-\beta} \log|A| + 2\frac{\beta}{1-\beta}||\sigma - P^R||_1^*(||Q^R|| + \log|A|),
\end{aligned}$$

where the fourth inequality is by $\max_i\{x_i\} \leq \log \sum_{i=1}^N \exp(x_i) \leq \max_i\{x_i\} + \log|N|$, the fifth inequality is by induction, and the last inequality is because $Q \in \mathscr{B}$ is bounded. The result follows by rearranging the terms. On the other hand,

$$\begin{aligned}
& |r^R(z, a) - r(x_*, a)| \\
& \leq |Q(x_*, a) - Q^R(z, a)| \\
& + \beta|\sum_{z'} \sigma(z', a, x_*) \log(\sum_{a'} \exp(Q(\lambda(z', x_*, a), a')) - \sum_{z'} P(z'|z, a) \log(\sum_{a'} \exp(Q^R(z', a')))|
\end{aligned}$$

Let $a^*(x) \in \arg\max_{a \in A} Q(x, a)$, and $a_R^*(z) \in \arg\max_{a \in A} Q^R(z, a)$. Then, we have

$$\begin{aligned}
& |r^R(z, a) - r(x_*, a)| \\
& \leq \rho^*(z, a) + \beta|\sum_{z'} \sigma(z', a, x_*) Q(\lambda(z', x_*, a), a^*[\lambda(z', x_*, a)]) - \sum_{z'} P(z'|z, a) Q^R(z', a^*[\lambda(z', x_*, a)])| \\
& + \beta|\sum_{z'} \sigma(z', a, x_*) \log(\sum_{a'} \exp(Q(\lambda(z', x_*, a), a') - Q(\lambda(z', x_*, a), a^*[\lambda(z', x_*, a)])) \\
& - \sum_{z'} P(z'|z, a) \log(\sum_{a'} \exp(Q^R(z', a') - Q^R(z', a^*[\lambda(z', x_*, a)]))))|,
\end{aligned}$$

since $\log \sum_{i=1}^{N} \exp(x_i) = a + \log \sum_{i=1}^{N} \exp(x_i - a), \forall a \in R$. Now let

$$\begin{aligned} h(z, a, x_*) = &\sum_{z'} \sigma(z', a, x_*) \log(\sum_{a'} \exp(Q(\lambda(z', x_*, a), a') - Q(\lambda(z', x_*, a), a^*[\lambda(z', x_*, a)])) \\ &- \sum_{z'} P(z'|z, a) \log(\sum_{a'} \exp(Q^R(z', a') - Q^R(z', a^*[\lambda(z', x_*, a)]))) \end{aligned}$$

Then,

$$\begin{aligned} &|r^R(z, a) - r(x_*, a)| \\ &\leq \rho^*(z, a) + \beta|\sum_{z'} \sigma(z', a, x_*) Q(\lambda(z', x_*, a), a^*[\lambda(z', x_*, a)]) - \sum_{z'} \sigma(z', a, x_*) Q^R(z', a^*[\lambda(z', x_*, a)])| \\ &+ \beta|\sum_{z'} \sigma(z', a, x_*) Q^R(z', a^*[\lambda(z', x_*, a)]) - \sum_{z'} P(z'|z, a) Q^R(z', a^*[\lambda(z', x_*, a)])| + \beta|h(z, a, x_*)| \\ &\leq (1 + \beta) \max_{z,a} \rho^*(z, a) + 2\beta||\sigma - P^R||_1^*||Q^R|| + \beta|h(z, a, x_*)|. \end{aligned}$$

Taking the maximum on both sides gives

$$\max_{z,a,x_*} |r^R(z, a) - r(x_*, a)| \leq (1 + \beta) \max_{z,a} \rho^*(z, a) + 2\beta||\sigma - P^R||_1^*||Q^R|| + \beta \max_{z,a,x_*} |h(z, a, x_*)|,$$

indicating the result. Lastly, it is easy to see that $h(z, a, x_*) \to 0$ as $O \to I$, where I is the identity matrix. Furthermore,

$$\begin{aligned} h(z, a, x_*) &\leq \log |A| - \sum_{z'} P(z'|z, a)[Q^R(z', a_R^*(z')) - Q^R(z', a^*[\lambda(z', x_*, a)])] \\ &\leq \log |A| + \max_{z,a,a'} |Q^R(z, a) - Q^R(z, a')|, \\ h(z, a, x_*) &\geq - \log |A| - \sum_{z'} P(z'|z, a)[Q^R(z', a_R^*(z')) - Q^R(z', a^*[\lambda(z', x_*, a)])] \\ &\geq - \log |A| - \max_{z,a,a'} |Q^R(z, a) - Q^R(z, a')|. \end{aligned}$$

Hence, the result follows. $\square$

References

[1] J. H. Abbring and Ø. Daljord. Identifying the discount factor in dynamic discrete choice models. *Quantitative Economics*, 11(2):471–501, 2020.

[2] V. Aguirregabiria, J. Gu, and Y. Luo. Sufficient statistics for unobserved heterogeneity in dynamic structural logit models. *Journal of Econometrics (Forthcoming)*, 2020.

[3] V. Aguirregabiria and P. Mira. Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica*, 70:1519–1543, 2002.

[4] V. Aguirregabiria and P. Mira. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156:38–67, 2010.

[5] P. Arcidiacono and J. B. Jones. Finite mixture distributions, sequential likelihood and the EM algorithm. *Econometrica*, 71(3):933–946, 2003.

[6] P. Arcidiacono and R. Miller. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, 79(6):1823–1867, 2011.

[7] J. R. Blevins. Sequential Monte Carlo methods for estimating dynamic microeconomic models. *Journal of Applied Econometrics*, 31(5):773–804, 2016.

[8] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 182–189. PMLR, 2011.

[9] L. Caliendo, M. Dvorkin, and F. Parro. Trade and labor market dynamics: general equilibrium analysis of the China trade shock. *Econometrica*, 87(3):741–835, 2019.

[10] A. Ching, T. Erdem, and M. Keane. Empirical models of learning dynamics: A survey of recent developments. *Handbook of Marketing Decision Models*, pages 223–257, 2017.

[11] J. Choi and K. Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730, 2011.

[12] C. Cirillo, R. Xu, and F. Bastin. A dynamic formulation for car ownership modeling. *Transportation Science*, 50(1):322–335, 2016.

[13] B. Connault. Hidden rust models. 2016.

[14] D. Diermeier, M. Keane, and A. Merlo. A political economy model of congressional careers. *American Economic Review*, 95(1):347–373, 2005.

[15] E. Duflo, R. Hanna, and S. P. Ryan. Incentives work: getting teachers to come to school. *American Economic Review*, 102(4):1241–1278, 2012.

[16] Z. Eckstein and K. Wolpin. Why youth drop out of high school: The impact of preferences, opportunities and abilities. *Econometrica*, 67(6):1295–1339, 1999.

[17] H. Fang and Y. Wang. Estimating dynamic discrete choice models with hyperbolic discounting, with an application to mammography decisions. *International Economic Review*, 56:565–596, 2015.

[18] C. Finn, S. Levine, and P. Abbeel. Guided cost learning: deep inverse optimal control via policy optimization. In *International conference on Machine learning*, 2016.

[19] I. Hendel and A. Nevo. Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673, 2006.

[20] V. J. Hotz and R. A. Miller. Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies*, 60(3):497–529, 1993.

[21] 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*, 61:265–289, 1994.

[22] Y. Hu and M. Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, 2012.

[23] S. Imai, N. Jain, and A. Ching. Bayesian estimation of dynamic discrete choice models. *Econometrica*, 77(6):1865–1899, 2009.

[24] M. Kalouptsidi, P. T. Scott, and E. Souza-Rodrigues. Identification of counterfactuals in dynamic discrete choice models. *Quantitative Economics (accepted)*, 2019.

[25] H. Kasahara and K. Shimotsu. Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175, 2009.

[26] H. Kasahara and K. Shimotsu. Estimation of discrete choice dynamic programming models. *Journal of Applied Econometrics*, 69(1):28–58, 2018.

[27] M. Keane and K. Wolpin. The career decisions of young men. *Journal of Political Economy*, 105(3):473–522, 1997.

[28] V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, 2016.

[29] T. Magnac and D. Thesmar. Identifying dynamic discrete choice processes. *Econometrica*, 70(2):801–816, 2002.

[30] D. McFadden. Econometric models of probabilistic choice. In *Structural Analysis of Discrete Data*. MIT Press, 1981.

[31] G. E. Monahan. State of the art—a survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[32] A. Norets. Inference in dynamic discrete choice models with serially correlated unobserved state variables. *Econometrica*, 77(5):1665–1682, 2009.

[33] O. L. Ortiz, A. L. Erera, and C. C. White. State observation accuracy and finite-memory policy performance. *Operations Research Letters*, 41:477–481, 2013.

[34] L. K. Platzman. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control and Optimization*, 18:362–380, 1980.

[35] G. Reich. Divide and conquer: Recursive likelihood function integration for hidden Markov models with continuous latent variables. *Operations Research*, 66(6):1457–1470, 2018.

[36] J. Rust. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033, 1987.

[37] J. Rust. Structural estimation of Markov decision processes. *Handbook of econometrics*, 4:3081–3143, 1994.

[38] J. Rust and C. Phelan. How social security and medicare affect retirement behavior in a world of incomplete markets. *Econometrica*, 65(4):781–831, 1997.

[39] M. Sanchez and G. A. Klutke. *Reliability and Life-Cycle Analysis of Deteriorating Systems*. Springer, 2016.

[40] G. Shani, J. Pineau, and R. Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27:1–51, 2013.

[41] C. Shi, R. Wan, R. Song, W. Lu, and L. Leng. Does the Markov decision process fit the data: Testing for markov property in sequential decision making. In *Proceedings of the 37 th International Conference on Machine Learning*. Vienna, Austria, PMLR 119, 2020.

[42] A. Simoes, J. M. Viegas, J. T. Farinha, and I. Fonseca. The state of art of hidden Markov models for predictive maintenance of diesel engines. *Quality and Reliability Engineering International*, 33:2765–2779, 2017.

[43] R. Smallwood and E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

[44] L. N. Steimle and B. T. Denton. Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*. International Series in Operations Research & Management Science, vol 248. Springer, Cham, 2017.

[45] C. L. Su and K. L. Judd. Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5):2213–2230, 2012.

[46] K. Train. *Discrete choice methods with simulation*. Cambridge University Press, 2002.

[47] C. C. White. A survey of solution techniques for the partially observed Markov decision process. *Annals of Operations Research*, 32:215–230, 1991.

[48] C. C. White and W. T. Scherer. Finite-memory suboptimal design for partially observed Markov decision processes. *Operations Research*, 42(3):439–455, 1994.

[49] J. Williams and S. Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21:393–422, 2007.

[50] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.