Structural Estimation of Partially Observable Markov Decision Processes

Yanling Chang, Alfredo Garcia, *Senior Member, IEEE*, Zhide Wang, and Lu Sun

***Abstract*—Partially observable Markov decision processes (POMDPs) is a well-developed framework for sequential decision-making under uncertainty and partial information. This article considers the (inverse) structural estimation of the primitives of a POMDP based upon data in the form of sequences of observables and implemented actions. We analyze the structural properties of an entropy regularized POMDP and specify conditions under which the model is identifiable without knowledge of the state dynamics. We consider a *soft* policy gradient algorithm to compute a maximum likelihood estimator, and illustrate the approach with an equipment replacement problem.**

***Index Terms*—Dynamic programming, maximum likelihood estimation, observability.**

I. Introduction

While there is extensive literature on the analysis and solution methods for partially observable Markov decision processes (POMDPs), the inverse problem, that is, the estimation of the primitives of a POMDP model based upon observable histories has not been amply examined. To our knowledge, the only method available in the literature for the structural estimation of POMDPs is described in [1]. However, the methodology proposed in [1] assumes the hidden-state dynamics and the observation probabilities are known. In addition, the methodology is heuristic as the search for policies that are consistent with data is restricted to a finite set. Given that an optimal policy for a POMDP is not necessarily unique, a POMDP may not be identifiable. Hence when several reward functions are consistent with data, the estimates obtained with the methodology proposed in [1] are of poor quality [2].

This article develops a methodology for the structural estimation of POMDPs that addresses the described shortcomings. We assume the model parameters are imprecisely known by the agent, and model decision-making in such environment by introducing an information processing cost (see, e.g., [3], [4], [5], [6]) which takes the form of entropy-based regularization. We first characterize the solution of entropy-regularized POMDPs to show the optimal policy is unique. This result is then leveraged to prove that an entropy-regularized POMDP model can be identified if the *priori* belief distribution, the cardinality of the system state space, the discount factor, and the reward for a fixed reference action are given. If the *priori* belief distribution is unknown, a set of model primitives can be identified. Such a set will converge to the true model primitives as the length of the available data sequences increases. We then develop a two-stage maximum likelihood estimation method and a *soft*-policy gradient algorithm to compute maximum likelihood estimators. We provide a finite-time characterization of the algorithm's performance. When rewards are linearly parametrized (with known hidden dynamics as in [1]) the algorithm is guaranteed to identify the parameters maximizing the log-likelihood in finite time.

The method is tested in an optimal engine replacement problem with both synthetic and real data. Using synthetic data, our method can accurately recover the true model primitives both *with* and *without* the knowledge of the *priori* belief distribution. This experiment also illustrates how model misspecification resulting from ignoring partial state observability can lead to models with poor fit. We further apply our method to a widely studied engine replacement real dataset. Compared to the results in [7], our model can dramatically improve the data fit by 17.7% in terms of the log-likelihood. The model also reveals a new feature of engine utilization in the dataset which was hitherto ignored, i.e., buses with engines *believed* to be in worse condition exhibit less utilization (mileage) and higher maintenance costs.

Article Organization: The rest of this article is organized as follows: Section II provides a literature review. Section III introduces an entropy-regularized POMDP and the identification result is in Section IV. Section V provides an estimation method for the POMDP with numerical illustration in Section VI. Finally, Section VII concludes this article.

II. Literature Review

The structural estimation of MDPs has been an active research area in artificial intelligence under the label of inverse reinforcement learning (IRL). A maximum entropy method proposed in [8] has been highly influential. Sample-based algorithms for implementing the maximum entropy method have scaled to scenarios with nonlinear reward functions (see, e.g., [9], [10]). Recently, L^*-based learning algorithms are also developed for MDPs (e.g., [11], [12]).

A literature on structural estimation of MDPs has also been developed in econometrics since [7] (see [13] for a recent survey). In this literature, it is assumed that persistent random reward perturbations are observed by the controller but *not* by the modeler. Hence, from the modeler's standpoint, the controller's observed actions are seen as samples from a randomized policy. Papers addressing computational challenges in estimating the structural MDP parameters include [13], [14], [15], and [16].

III. Entropy Regularized POMDP Model

At each period $t \geq 0$, the state $s_t \in S$ is not directly observable to the controller nor to the external modeler. Both the controller and the external modeler can receive an observation $z_t \in Z$ correlated with the underlying state s_t. We assume finite spaces of action A, state S and observations Z. If the hidden state is s_t and $a_t \in A$ is

Manuscript received 28 June 2022; accepted 15 October 2022. Date of publication 28 October 2022; date of current version 28 July 2023. This work was supported by NSF under Grant 2048395. Recommended by Associate Editor A. A. Malikopoulos. *(Corresponding author: Yanling Chang.)*

Yanling Chang is with the Department of Engineering Technology and Industrial Distribution, and also with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77840 USA (e-mail: yanling.chang@tamu.edu).

Alfredo Garcia, Zhide Wang, and Lu Sun are with the Department of Industrial and Systems Engineering,Texas A&M University, College Station, TX 77840 USA (e-mail: alfredo.garcia@tamu.edu; liang93429@tamu.edu; lusun8825@tamu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3217908.

Digital Object Identifier 10.1109/TAC.2022.3217908

implemented, the reward accrued is $r_{\theta_1}(z_t, s_t, a_t)$, where $\theta_1 \in \mathbb{R}^{p_1}$ for some $p_1 \in \mathbb{N}_+$. The system dynamics are described by probabilities $\mathbb{P}_{\theta_2}(z_{t+1}, s_{t+1}|z_t, s_t, a_t)$, where $\theta_2 \in \mathbb{R}^{p_2}$ for some $p_2 \in \mathbb{N}_+$. Let $\theta = (\theta_1, \theta_2)$.

Let $h_t = \{z_t, \ldots, z_0, a_{t-1}, \ldots, a_0, x_0\} \in H_t$ be a publicly observable history of the dynamic decision process including all past and present observations and all past actions up to time $t > 0$, where $H_t \triangleq Z^{t+1} \times A^t \times X$ and $X \subset \mathbb{R}^{|S|-1}$ is the unit simplex and $x_0 = \{\mathbb{P}(s) : s \in S\} \in X$ is the prior belief distribution (a probability mass vector) over S. We consider randomized policies π adapted to the history of the process, i.e., $\pi(a|h_t) \in [0, 1], a \in A$ and $\sum_{a\in A} \pi(a|h_t) = 1$ for all $h_t \in H_t$.

Let $\mathcal{H}(\pi(\cdot|h_t)) = -\sum_{a\in A} \pi(a|h_t) \log \pi(a|h_t)$ denote the entropy of the distribution $\pi(\cdot|h_t)$ given history $h_t \in H_t$. The controller aims to maximize the expected value of the entropy-regularized discounted reward

$$U_{t,\theta}(h_t) \triangleq \sup_{\pi\in\Pi} \mathbb{E}\left[\sum_{\tau\geq t} \beta^{\tau-t}[r_{\theta_1}(z_\tau, s_\tau, a_\tau) + \alpha\mathcal{H}(\pi(\cdot|h_\tau))]\right] \tag{1}$$

where $\beta \in (0, 1)$ is the discount factor, Π is the set of randomized policies adapted to the history process, and the second term is related to an information processing cost with $\alpha \geq 0$ a scale parameter. Increasing values of $\alpha > 0$ reflect an increasing cost for improving (and exploiting) the information in the estimates available to the agent, whereas $\alpha \to 0^+$ describes the case with negligible costs for improving estimates. Entropy regularized MDP models have also been used in reinforcement learning and IRL [17], [18]. In what follows, with no loss of generality we assume $\alpha = 1$, as the objective function in (1) above can be rescaled by $\frac{1}{\alpha}$.

Let $x_{t,\theta_2} = \mathbb{P}_{\theta_2}(\cdot|h_t) \in X \subset \mathbb{R}^{|S|-1}$ be the belief distribution on the state, i.e., the conditional probability distribution of s_t given history h_t (a column vector). Define the observation probabilities: $\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \triangleq \sum_{s'} \sum_s x_{t,\theta_2}(s)\mathbb{P}_{\theta_2}(z_{t+1}, s'|z_t, s, a_t)$, and the belief update function (in matrix form)

$$\lambda_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \triangleq \frac{x_{t,\theta_2}^\top P_{\theta_2}(z_{t+1}, z_t, a_t)}{\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t)} \tag{2}$$

assuming $\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) \neq 0$, where we denote the (s, s') element of the matrix $[P_{\theta_2}(z_{t+1}, z_t, a_t)]_{s,s'} \triangleq \mathbb{P}_{\theta_2}(z_{t+1}, s'|z_t, s, a_t), s, s' \in S$, $x_{t,\theta_2}^\top$ is the transpose of x_{t,θ_2}, and $[x_{t,\theta_2}^\top P_{\theta_2}(z_{t+1}, z_t, a_t)]_{s_{t+1}} \triangleq \sum_s x_{t,\theta_2}(s)\mathbb{P}_{\theta_2}(z_{t+1}, s_{t+1}|z_t, s, a_t)$. By the Bayes' rule, it is easy to verify that $x_{t+1,\theta_2} = \lambda_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t)$ and $\{x_{t,\theta_2}, t \geq 0\}$ is a controlled Markov process. Finally, we denote $r_{\theta_1}(z_t, x_{t,\theta_2}, a_t) \triangleq \sum_s r_{\theta_1}(z_t, s, a_t)x_{t,\theta_2}(s)$.

Let $h_t \in H_t$ denote a history with current observation $z_t = z$ and belief $x_{t,\theta_2} = x$ and define $V_{t,\theta}(z, x)$ by

$$V_{t,\theta}(z, x) = \max_{\pi(\cdot|z,x)} \left\{\sum_a r_{\theta_1}(z, x, a)\pi(a|z, x) + \mathcal{H}(\pi(\cdot|z, x)) + \beta \sum_{z',a} \sigma_{\theta_2}(z', z, x, a)V_{t+1,\theta}(z', x'(a))\pi(a|z, x)\right\} \tag{3}$$

where $x'(a) = \lambda_{\theta_2}(z', z, x, a)$. Mathematical induction shows that $U_{t,\theta}(h_t) = V_{t,\theta}(z_t, x_t)$ for all $h_t \in H_t$ and hence, (z_t, x_t) is a sufficient statistic for solving (1).

For the infinite-horizon case, let $\mathcal{Q}$ be the Banach space of bounded, measurable functions $Q : Z \times X \times A \to \mathbb{R}$ under the supremum norm $||.||$. Define the *soft* Bellman operator $\mathcal{B}_\theta : \mathcal{Q} \to \mathcal{Q}$ by

$$[\mathcal{B}_\theta Q](z, x, a) = r_{\theta_1}(z, x, a) + \beta \sum_{z'} \sigma_{\theta_2}(z', z, x, a)V(z', x') \tag{4}$$

where $x' = \lambda_{\theta_2}(z', z, x, a)$ and

$$V(z, x) \triangleq \max_{\hat\pi(\cdot|z,x)} \left[\sum_a Q(z, x, a)\hat\pi(a|z, x) + \mathcal{H}(\hat\pi(\cdot|z, x))\right]. \tag{5}$$

Theorem 1: (a) $V(z, x) = \log \sum_a \exp(Q(z, x, a))$ for all $Q \in \mathcal{Q}$. (b) $\mathcal{B}_\theta : \mathcal{Q} \to \mathcal{Q}$ is a contraction mapping with modulus $\beta \in (0, 1)$ and (c) the optimal policy is Markovian and in the form of

$$\pi_\theta(a|z, x) = \frac{\exp Q_\theta(z, x, a)}{\sum_{a'\in A} \exp Q_\theta(z, x, a')} \tag{6}$$

where Q_θ is the unique fixed point of $\mathcal{B}_\theta$.

Remark 1: It is easy to verify Theorem 1 continues to hold for the finite-horizon case. Evidently, the results in this case require that the state-action function $Q_{t,\theta}$ and the policy $\pi_{t,\theta}$ are time-dependent t.

IV. Model Identification

The structure of the POMDP model is uniquely defined by the reward parameter θ_1 and the dynamics parameter θ_2. The conditional observation probabilities $\{\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t)\}$ and conditional choice probabilities $\{\pi_\theta(a_t|z_t, x_{t,\theta_2})\}$ are called the reduced form observation probabilities and choice probabilities under structure $\theta = (\theta_1, \theta_2)$. Under the true structure $\theta^* = (\theta_1^*, \theta_2^*)$, the reduced form probabilities must be equal to those observed: $\forall(z_{t+1}, h_t, a_t)$

$$\underbrace{\hat{\mathbb{P}}(z_{t+1}|h_t, a_t)}_{\text{Data}} = \underbrace{\sigma_{\theta_2^*}(z_{t+1}, z_t, x_{t,\theta_2^*}, a_t)}_{\text{Model}} \tag{7}$$

$$\underbrace{\hat{\mathbb{P}}(a_t|h_t)}_{\text{Data}} = \underbrace{\pi_{\theta^*}(a_t|z_t, x_{t,\theta_2^*})}_{\text{Model}}. \tag{8}$$

where $\hat{\mathbb{P}}(z_{t+1}|h_t, a_t)$ and $\hat{\mathbb{P}}(a_t|h_t)$ are empirical distributions. Magnac and Thesmar [19] defines the observational equivalence and identification as follows.

Definition 1 (Observational equivalence): Let Θ be the set of structures θ, and let $\overset{o}{\iff}$ be observational equivalence. $\forall\theta, \theta' \in \Theta, \theta \overset{o}{\iff} \theta'$ if and only if $\sigma_{\theta_2}(z_{t+1}, z_t, x_{t,\theta_2}, a_t) = \sigma_{\theta'_2}(z_{t+1}, z_t, x_{t,\theta'_2}, a_t)$ and $\pi_\theta(a_t|z_t, x_{t,\theta_2}) = \pi_{\theta'}(a_t|z_t, x_{t,\theta'_2}), \forall z_{t+1}, z_t, a_t$.

Definition 2 (Identification): The model is identified if and only if $\forall\theta, \theta' \in \Theta, \theta \overset{o}{\iff} \theta'$ implies $\theta = \theta'$.

It is well known that the primitives of an MDP cannot be identified in general [19], [20], and POMDPs are not an exception. In addition, in the POMDP, the system dynamics cannot be directly observed. However, the next theorem shows that we could identify the hidden dynamics using two periods of data (including x_0), assuming we know the cardinality of the state space $|S|$, where the data consists of $N \geq 1$ finite histories of pairs $h_{T,i} = \{x_{0,i}, z_{t,i}, a_{t,i}, t = 1, \ldots, T\}$ for $i \in \{1, \ldots, N\}$.

Theorem 2: Assume $|S|$ is known. The hidden system dynamics θ_2 can be uniquely identified from the first two periods of data (including x_0).

Theorem 2 is crucial as once the hidden dynamic θ_2 is identified, we can transform each history trajectory h_t to a belief trajectory $\{x_{t,\theta_2}, t \geq$

$0\}$ by (2), which is only dependent on θ_2. Hence, the POMDP problem is transformed to an MDP with "known" beliefs (unknown if θ_2 is unknown), allowing us to generalize the identification results in [14] and [19] for MDPs to POMDPs.

Theorem 3: The primitives of the (infinite-horizon) POMDP $\theta = (\theta_1, \theta_2)$ cannot be identified in general. However, θ_1 and θ_2 can be uniquely identified from the data, given the initial belief x_0, the cardinality of the state space $|S|$, the discount factor β, and the rewards $r_{\theta_1}(z, s, a^0), z \in Z, s \in S$ for a reference action $a^0 \in A$ are all known.

Corollary 1: The (finite-horizon) POMDP model can be uniquely identified from the data, if x_0, $|S|$ and both the reward structure and the terminal value function $\bar{Q}$ in the reference action $a^0 \in A$ are all known.

It is well known that the knowledge on the discount factor β and the reward values for a reference action are necessary to uniquely identify the primitives of an MDP model (e.g., [7], [14]). The identification result of the POMDP in Theorem 3 requires two additional conditions on the knowledge of the initial belief x_0 and the cardinality of the state space. Theorem 4 next examines the case where the initial belief x_0 is unknown to the modeler. It is an interesting future research question to examine whether or under what conditions the POMDP model is identifiable if $|S|$ is unknown. In practice, the number of possible states can be obtained by domain knowledge for a particular application. A practitioner can also try possible values of $|S|$ to examine which value can best explain the observed behaviors.

Given a finite history $\{z_t, \ldots, z_{t-M}, a_{t-1}, \ldots, a_{t-M}\}$, let λ^M be M applications of λ function for an arbitrary $x_{t-M} \in X$, namely, $\lambda^M_{\theta_2}(z^t_{t-M}, a^{t-1}_{t-M}, x_{t-M}) = \lambda_{\theta_2}(z_t, z_{t-1}, \lambda_{\theta_2}(\ldots \lambda_{\theta_2}(z_{t-M+1}, z_{t-M}, x_{t-M}, a_{t-M})), a_{t-1})$, where $z^t_{t-M} = \{z_t, \ldots, z_{t-M}\}, a^{t-1}_{t-M} = \{a_{t-1}, \ldots, a_{t-M}\}$ for short.

Theorem 4: Assume $|S|$ is known but x_0 is unknown to the modeler. A set of model primitives Θ can be obtained from the data (each $\theta \in \Theta$ is consistent with an unknown $x_0 \in X$), given the discount factor β, and the reward $r_{\theta_1}(Z, S, a^0)$. The set Θ will shrink to the singleton true value θ^* as $M \to \infty$ (hence, $T \to \infty$).

Theorem 4 shows if x_0 is unknown, we can obtain a set of θ_2 (hence θ_1) by repeated applications of λ function and by varying $x_0 \in X$. The set of possible θs will shrink to the singleton true value θ^* as M goes to infinity. In addition, in many applications, it is also possible to obtain some (or a small range of) belief points, i.e., $x_0 \in X' \subset X$. This information of X' can be very helpful in improving the accuracy of the estimates. At last, x_0 can also be treated as a model parameter to estimate in order to understand the initial belief of the decision-making agent (see Section V).

V. Maximum Likelihood Estimation

Define the log-likelihood of the data by

$$\begin{aligned}\log \ell(\theta) &\triangleq \log \prod_{i=1}^{N} \mathbb{P}(h_{T,i}|x_{0,i})x_{0,i} \\ &= \sum_{i=1}^{N}\sum_{t=0}^{T-1}[\log \sigma_{\theta_2}(z_{t+1,i}, z_{t,i}, x_{t,\theta_2,i}, a_{t,i}) \\ &\quad + \log \pi_\theta(a_{t,i}|z_{t,i}, x_{t,\theta_2,i})] + \sum_{i=1}^{N}\log x_{0,i}. \end{aligned} \tag{9}$$

We now introduce a *Soft* policy gradient algorithm for approximately maximizing the log-likelihood (9). A two-stage estimator can be obtained as follows.

First stage: Solve for the value of $\hat{\theta}_2$ that maximizes the value of the first term on the right-hand side in (9) if $x_{0,i}$ is known, or by maximizing the value of the first term and the last term of (9) if $x_{0,i}$ is unknown. Note that given a fixed value of θ_2 and $\forall x_{0,i} \in X$, a sequence of belief trajectories $\{x_{t,\theta_2,i}\}_{t=0}^{T}$ can be recursively computed as $x_{t+1,\theta_2,i} = \lambda_{\theta_2}(z_{t+1,i}, z_{t,i}, x_{t,\theta_2,i}, a_{t,i})$ via (2).

Algorithm 1 *Soft* Policy Gradient Algorithm.

Compute $\hat{\theta}_2$ and $x_{t,i} = x_{t,\hat{\theta}_2,i}\ t = 1, \ldots T, i = 1, \ldots, N$;
Initialize $k = 0$, θ_1^0, $\nabla_{\theta_1}\hat{\ell}(\theta_1^0)$, ϵ and ρ;
while $\left\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\right\| \geq \epsilon$ **do**
  $k \leftarrow k + 1$;
  Compute $\nabla_{\theta_1} Q_{\theta_1^k}(a, z_{t,i}, x_{t,i}), a \in A$, and
  $\nabla_{\theta_1}\hat{\ell}(\theta_1^k) = \sum_{i=1}^{N}\sum_{t=0}^{T-1}\nabla_{\theta_1}\log \pi_{\theta_1^k}(a_{t,i}|z_{t,i}, x_{t,i})$;
  Update parameters $\theta_1^{k+1} = \theta_1^k + \rho\nabla_{\theta_1}\hat{\ell}(\theta_1^k)$;
end

Second stage: Once $\hat{\theta}_2$ (and an estimate of $x_{0,i}$) is obtained, solve for the value of $\hat{\theta}_1$ that maximizes the log of pseudo-likelihood $\hat{\ell}(\theta)$ defined as

$$\hat{\ell}(\theta_1) = \sum_{i=1}^{N}\sum_{t=0}^{T-1}\log \pi_{(\theta_1,\hat{\theta}_2)}(a_{t,i}|z_{t,i}, x_{t,\hat{\theta}_2,i}). \tag{10}$$

We simplify notation by using π_{θ_1} and $x_{t,i}$ to refer to $\pi_{(\theta_1,\hat{\theta}_2)}$ and $x_{t,\hat{\theta}_2,i}$, respectively. For a given value of θ_1, consider the *soft* Bellman equation $Q_{\theta_1}(z, x, a) = r_{\theta_1}(z, x, a) + \beta\sum_{z'}\sigma_{\hat{\theta}_2}(z', z, x, a)V_{\theta_1}(z, x')$, where $x' = \lambda_{\hat{\theta}_2}(z', z, x, a)$. After solving the *soft* Bellman equation for fixed θ_1 we can compute the gradient: $\nabla_{\theta_1}\log \pi_{\theta_1}(a|z, x) = \nabla_{\theta_1}\log\left(\frac{\exp Q_{\theta_1}(z,x,a)}{\sum_{a'\in A}\exp Q_{\theta_1}(z,x,a')}\right) = \nabla_{\theta_1}Q_{\theta_1}(z, x, a) - \sum_{a'}\pi_{\theta_1}(a'|z, x)\nabla_{\theta_1}Q_{\theta_1}(z, x, a')$. The basic steps of a *soft* policy gradient algorithm are listed in Algorithm 1.

Theorem 5: Assume $r_{\theta_1}(z, x, a)$ is twice continuously differentiable in $\theta_1 \in \mathbb{R}^{p_1}$ and $\forall(z, x, a) \in Z \times X \times A$

$$\sup_{\theta_1}\|\nabla_{\theta_1}r_{\theta_1}(z, x, a)\| \leq L_{r,1} < \infty$$

$$\sup_{\theta_1}\left\|\nabla^2_{\theta_1}r_{\theta_1}(z, x, a)\right\| \leq L_{r,2} < \infty.$$

The pseudo-log likelihood has Lipschitz continuous gradients with constant $L \triangleq NT(L_Q + L_V)$. With step size $\rho < \frac{2}{L}$ it holds that

$$\min_{k=1,\ldots,K}\left\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\right\|^2 \leq \frac{1}{\rho\left(1 - \frac{\rho L}{2}\right)}\frac{\hat{\ell}(\theta_1^*) - \hat{\ell}(\theta_1^0)}{K}$$

where θ_1^* maximizes pseudo-log likelihood.

Remark 2: According to Theorem 5, the *least* gradient norm converges *sublinearly* to zero, i.e., it takes $\mathcal{O}(\frac{1}{\epsilon})$ iterations for the *least* gradient norm to be less than or equal to $\epsilon > 0$. Thus, the stopping rule in the algorithm will be met in *finite* time. When rewards are linearly parametrized as $r(z, x, a) = \theta^\top\phi(z, x, a)$ with feature vector $\phi(z, x, a) \geq 0$, the first-order condition is also sufficient as it can be shown that the log-likelihood function (10) is concave.

VI. Illustration: Optimal Equipment Replacement

POMDPs have been widely used in engine maintenance problems (e.g., [21], [22]). In this section, we demonstrate the estimation method

TABLE I
PARAMETER ESTIMATES AND log-LIKELIHOOD OF POMDP-BASED MODEL

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	log-Likelihood
True value	0.039	0.333	0.590	0.181	0.757	0.061	0.949	0.988	0.2	1.2	9.24	*
x_0 known	0.038	0.327	0.596	0.181	0.754	0.064	0.950	0.987	0.2	1.2	9.23	-262,973
x_0 unknown	0.038	0.327	0.596	0.181	0.754	0.064	0.950	0.987	0.2	1.2	9.23	-262,814
$\alpha = 0.001$	0.038	0.331	0.592	0.186	0.750	0.062	0.950	0.987	0.2	1.2	9.23	*

using a bus engine replacement example with both synthetic and real datasets.

The engine deterioration state $s_t \in S = \{0, 1\}$, where "0" is being the "good state" and "1" is being the "bad state." At each month, the available actions are $a_t = 1$ for engine replacement and $a_t = 0$ for regular maintenance (can restore the engine to "like-new" condition with a small probability). The model for hidden state dynamics is

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

and $\mathbb{P}_{\theta_2}(s_{t+1} = 0|s_t, a_t = 1) = 1$. The cumulative mileage ranges from 0 to 437.5 K miles and we discretize it into 175 bins of 2.5 K miles. Thus, the cumulative mileage z_t is in $Z = \{0, 1, 2, \ldots, 174\}$ and z_t is a noisy observation of the true hidden engine deterioration state s_t. The monthly mileage increment $z_{t+1} - z_t$ is limited to $\Delta \in \{0, 1, 2, 3\}$, corresponding to values between $[0, 2.5\,K)$, $[2.5\,K, 5\,K)$, $[5\,K, 7.5\,K)$ and $[7.5\,K, 10\,K)$, respectively. The distribution is parametrized as

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

Similarly, we define $\mathbb{P}_{\theta_3}(\, z_{t+1} = z_t + \Delta |\, z_t, s_t = 1, a_t = 0) = \theta_{3,1,\Delta}, \Delta \in \{0, 1, 2, 3\}$. Furthermore, after a replacement, the mileage restarts from zero: $\mathbb{P}_{\theta_3}(\, z_{t+1} = 0 |\, z_t, s_t, a_t = 1) = 1$.

Per 2500-mile maintenance costs are parametrized by $\theta_{1,0}$ in good state and $\theta_{1,1}$ in bad state (state-dependent)

$$\begin{aligned} r_{\theta_1}(z_t, s_t = 0, a = 0) &= -0.001\theta_{1,0} z_t \\ r_{\theta_1}(z_t, s_t = 1, a = 0) &= -0.001\theta_{1,1} z_t \end{aligned}$$

and the replacement cost is $r_{\theta_1}(z_t, s_t, a = 1) = -RC, RC > 0$. Hence, 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

$$\begin{aligned} r_{\theta_1}(z_t, x_t, a = 0) &= \theta_1^\top \phi(z_t, x_t, a = 0) \\ r_{\theta_1}(z_t, x_t, a = 1) &= \theta_1^\top \phi(z_t, x_t, a = 1) \end{aligned}$$

where $\theta_1^\top = -[0.001\theta_{1,0}\ 0.001\theta_{1,1}\ RC]$ and

$$\begin{aligned} \phi(z_t, x_t, a = 0)^\top &= [z_t x_t\ \ z_t(1 - x_t)\ \ 0] \\ \phi(z_t, x_t, a = 1)^\top &= [0 \quad 0 \quad 1]. \end{aligned}$$

A. Synthetic Dataset

We simulate 3000 buses for 100 decision epochs with ground truth parameters in Table I. The estimation results in Table I shows our algorithm can accurately identify the model parameters. In addition, the prior knowledge on initial belief x_0 does not influence the estimation result in a significant way. Theorem 4 shows if x_0 is unknown, the resulting estimates may deviate from their true values; however, the difference between the estimated results and the true values quickly diminishes as the number of decision epochs increases. Fig. 1 clearly illustrates this fact (where the estimation deviation is caused by varying $x_0 \in X$). Using the same ground truth parameters, we build a classical POMDP and approximate it by the entropy-regularized POMDP with $\alpha = 0.001$. The result in the last row of Table I shows that our method can also estimate the classical POMDP model primitives with high accuracy.

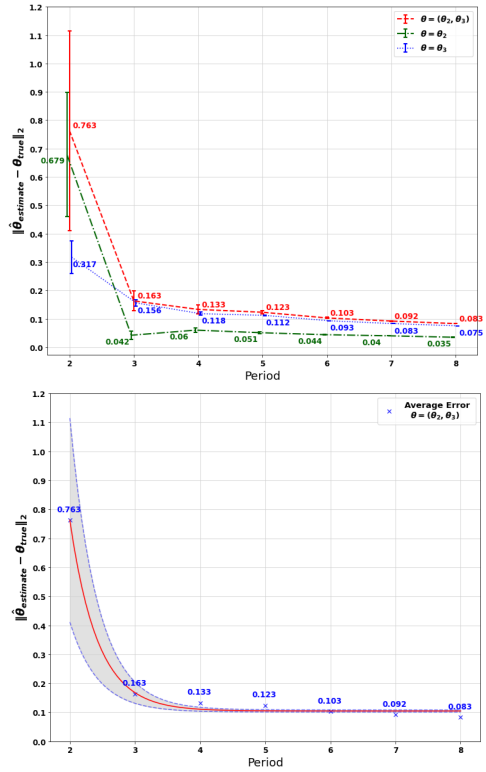

Fig. 1. Estimation results are affected by the prior knowledge of x_0. Without knowing x_0, more periods of data are needed for more accurate estimation. The unknown x_0 induced deviation vanishes quickly, as the number of data periods grows.

Model Misspecification: When the data are generated by a POMDP, using existing MDP-based models will lead to misspecification errors. To see this, we apply the MDP model in [7] to the same synthetic dataset where the cumulative mileage z_t is treated as the state variable. The MDP model is specified as follows: With z_t cumulative mileage after t months, the expected (monthly) maintenance cost is $r_{\theta_1}(z_t, a =$

TABLE II
PARAMETER ESTIMATES AND log-LIKELIHOOD PROVIDED BY THE MDP MODEL FOR THE SYNTHETIC DATA

Parameter	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$	θ_1	RC	log-Likelihood
MDP Model	0.1275	0.6009	0.2569	0.0147	1.1	9.8106	-301,750

TABLE III
PARAMETER ESTIMATES AND log-LIKELIHOOD PROVIDED BY THE POMDP MODEL (STANDARD ERRORS OBTAINED BY BOOTSTRAPPING METHOD ARE IN PARENTHESES)

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	log-Likelihood
Good State	0.039 (.005)	0.335 (.018)	0.588 (.018)	*	*	*	0.949 (.004)	*	0.3 (.3)	*	9.738 (1.052)	-3819
Bad State	*	*	*	0.182 (.008)	0.757 (.008)	0.061 (.006)	*	0.988 (.002)	*	1.3 (.2)		

TABLE IV
PARAMETER ESTIMATES AND log-LIKELIHOOD WITH MDP MODEL

Parameter	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$	θ_1	RC	log-Likelihood
MDP Model [7] p. 1022 (Standard errors in parentheses)	0.119 (0.005)	0.576 (0.008)	0.287 (0.007)	0.016 (0.002)	1.2 (0.3)	10.90 (1.581)	-4495

$0) = -0.001\theta_1 z_t$, and $r_{\theta_1}(z_t, a = 1) = -RC$. The distribution of the monthly mileage increment is parametrized as

$$\mathbb{P}_{\theta_3}(z_{t+1} = z_t + \Delta | z_t, a_t = 0) = \theta_{3,\Delta}, \qquad \Delta \in \{0, 1, 2\}$$
$$\mathbb{P}_{\theta_3}(z_{t+1} = z_t + \Delta | z_t, a_t = 0) = 1 - \theta_{3,0} - \theta_{3,1} - \theta_{3,2}, \qquad \Delta = 3$$

and replacement action resets the mileage state of a bus: $\mathbb{P}_{\theta_3}(z_{t+1} = 0 | z_t, a_t = 1) = 1$.

The MDP estimation results are in Table II. Unsurprisingly, the misspecification error manifests itself by a significant drop in log-likelihood ($\frac{301750-262973}{301750} = 12.9\%$). In general, the modeling options for a given dataset include MDPs (possibly high-order MDPs), POMDP, or other non-Markovian processes. A central question for the modeler is to select an appropriate model which may not necessarily be Markovian. Thus, there is a clarion call for developing estimation methods for possible (non-Markovian) stochastic processes. The developed method for POMDP estimation represents an important building block toward this direction.

B. Real Dataset

We further exhibit our estimation approach by revisiting a real dataset in [7], Group 4 consisting of buses with 1975 GMC engines, and assume the bus maintenance manager makes optimal maintenance/replacement decisions by considering the tradeoff between minimizing maintenance costs and minimizing the costs for potential breakdown due to prolonged course of use. 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. Thus, we fit the data by our POMDP model defined in Section VI-A, and compared our estimation results with the results found in [7] (from an MDP model). In terms of log-likelihood, our POMDP model outperformed the MDP model by $\frac{4495-3819}{3819} = 17.7\%$ (see Tables III and IV). The POMDP model also captures a feature of engine utilization that are conflated in the MDP model: 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 (possibly because buses in worse condition are more likely to experience mechanical problems and be in the shop for repairs). Furthermore, we find that the marginal operation costs ($\theta_{1,1}$) for buses in *bad* state is significantly higher than those ($\theta_{1,0}$) in *good* state (at least about two times, taking the standard deviation into consideration).

VII. CONCLUSION

In this article, we developed a novel estimation method to recover the primitives of a POMDP model based on observable trajectories of the process. We provide a characterization of optimal decisions in an entropy regularized POMDP model by means of *soft* Bellman equation and analyzed identifiability of the model. We then developed a soft policy gradient algorithm to obtain the maximum likelihood estimator and provided a numerical illustration with an application to optimal equipment replacement. Note that computational challenges of our model are obviously not trivial. Thus, a future research direction is to address computational challenges of high dimensional hidden state models.

APPENDIX A

Proof of Theorem 1: The proof closely follows the proof in [18] for the completely observable case. ■

Proof of Theorem 2: For any given two system dynamics of POMDP: $P_{\theta_2}(z,' z, a)$, $P_{\theta_2'}(z,' z, a)$, where $P(z,' z, a) = \{P_{ij}(z,' z, a)\}_{i,j\in S}$, $P_{ij}(z,' z, a) = \mathbb{P}(z_{t+1} = z,' s_{t+1} = j | z_t = z, s_t = i, a_t = a)$, we show that they can be distinguished by the data. Since the dataset contains x_0, z_0, a_0 and $|S|$ is known, we can obtain $\sigma^0_{\theta_2}(z_1, z_0, x_0, a_0) = \sum_s x_0(s)\mathbb{P}_{\theta_2}(z_1|z_0, s, a_0)$, and $\sigma^0_{\theta_2'}(z_1, z_0, x_0, a_0) = \sum_s x_0(s)\mathbb{P}_{\theta_2'}(z_1|z_0, s, a_0)$.

Note that $\sigma^0(z_1, z_0, x_0, a_0) = \hat{\mathbb{P}}(z_1|\zeta_0, a_0)$, where $\hat{\mathbb{P}}(z_1|\zeta_0, a_0)$ is a function of the first period data. Thus, σ can be obtained from the data and $\sigma^0_{\theta_2} = \sigma^0_{\theta_2'}$ if and only if $\mathbb{P}_{\theta_2}(z'|z, s, a) = \mathbb{P}_{\theta_2'}(z'|z, s, a)$. If $\sigma^0_{\theta_2} \neq \sigma^0_{\theta_2'}$, we are done. However, it is possible that there exists s' such that $\mathbb{P}_{\theta_2}(z,' s'|z, s, a) \neq \mathbb{P}_{\theta_2'}(z,' s'|z, s, a)$ but $\mathbb{P}_{\theta_2}(z'|z, s, a) = \mathbb{P}_{\theta_2'}(z'|z, s, a)$. In this case, update belief by (2)

(in matrix form), $x^\top_{1,\theta_2}=\frac{x_0^\top P_{\theta_2}(z_1,z_0,a_0)}{\sigma^0_{\theta_2}(z_1,z_0,x_0,a_0)}$, $x^\top_{1,\theta'_2}=\frac{x_0^\top P_{\theta'_2}(z_1,z_0,a_0)}{\sigma^0_{\theta'_2}(z_1,z_0,x_0,a_0)}$. Then $x_{1,\theta_2}=x_{1,\theta'_2}$ if and only if $P_{\theta_2}(z,'z,a)=P_{\theta'_2}(z,'z,a)$. Now, $\sigma^1_{\theta_2}(z_2,z_1,x_{1,\theta_2},a_1)=\sum_s x_{1,\theta_2}(s)\mathbb{P}_{\theta_2}(z_2|z_1,s,a_1)$ and $\sigma^1_{\theta'_2}(z_2,z_1,x_{1,\theta'_2},a_1)=\sum_s x_{1,\theta'_2}(s)\mathbb{P}_{\theta'_2}(z_2|z_1,s,a_1)$, and again σ^1 is obtainable from the two periods of data as $\sigma^1(z_2,z_1,x_1,a_1)=\hat{\mathbb{P}}(z_2|\zeta_1,a_1)$. Now, $\sigma^1_{\theta_2}=\sigma^1_{\theta'_2}$ if and only if $x_{1,\theta_2}=x_{1,\theta'_2}$, indicating $P_{\theta_2}(z,'z,a)=P_{\theta'_2}(z,'z,a)$ assuming $P(z,'z,a)$ is not rank-1 (each row is the same). ■

Proof of Theorem 3: By Theorem 2, both $\pi(a|z_t,x_t)=\hat{\mathbb{P}}(a|h_t)$ and hidden dynamics $\{P_{\theta_2}(z,'z,a)\}$ can be identified from the data. It is clear to see that per Theorem 1 that $\log(\frac{\pi_\theta(a|z,x)}{\pi_\theta(a^0|z,x)})=Q_\theta(z,x,a)-Q_\theta(z,x,a^0)$ for a fixed reference action $a^0\in A$. We also have $V_\theta(z,x)=\log\sum_{a\in A}\exp\left(Q_\theta(z,x,a)-Q_\theta(z,x,a^0)\right)+Q_\theta(z,x,a^0)$. Note that $Q_\theta(z,x,a^0)=r_{\theta_1}(z,x,a^0)+\beta\mathbb{E}_{z'}[V_\theta(z,'\lambda_{\theta_2}(z,'z,x,a^0))|z,x,a^0]$. Accordingly, $Q_\theta(z,x,a^0)=r_{\theta_1}(z,x,a^0)+\beta\mathbb{E}_{z'}[\log\sum_{a\in A}\exp(Q_\theta(z,'\lambda_{\theta_2}(z,'z,x,a^0),a)-Q_\theta(z,'\lambda_{\theta_2}(z,'z,x,a^0),a^0))|z,x,a^0]+\beta\mathbb{E}_{z'}[Q_\theta(z,'\lambda_{\theta_2}(z,'z,x,a^0),a^0))|z,x,a^0]$. Note

$$Q_\theta(z,x,a^0)=r_{\theta_1}(z,x,a^0)+C+\beta\mathbb{E}_{z'}[Q_\theta(z,'\lambda_{\theta_2}(z,'z,x,a^0),a^0))|z,x,a^0] \quad (11)$$

where $C=\beta\mathbb{E}_{z'}[\log\sum_{a\in A}\frac{\pi(a|z,'\lambda_{\theta_2}(z,'z,x,a^0))}{\pi(a^0|z,'\lambda_{\theta_2}(z,'z,x,a^0))}|z,x,a^0]$, and it can be obtained from the dataset per Theorem 2. It is easy to show that (11) is a contraction mapping; hence, there is a unique solution for $Q_\theta(Z,X,a^0)$, for a given $r_{\theta_1}(Z,X,a^0)$. Thus, $Q_\theta(z,x,a)$ can be identified for all $a\in A$ and consequently $V_\theta(z,x)$ as well. Since $r_{\theta_1}(z,x,a)=Q_\theta(z,x,a)-\beta\mathbb{E}_{z'}[V(z,'\lambda_{\theta_2}(z,'z,x,a))|z,x,a]$ we can get $r_{\theta_1}(z,x,a)$, and consequently, $r_{\theta_1}(z,s,a)$. ■

Proof of Corollary 1: Given terminal value $\bar{Q}_\theta(Z,X,a^0)$ and $r_{\theta_1}(Z,S,a^0)$, we obtain $Q_{t,\theta}(Z,X,a^0)$ via $Q_{t,\theta}(z,x,a^0)=r_{\theta_1}(z,x,a^0)+\beta\mathbb{E}_{z'}[Q_{t+1,\theta}(z,'\lambda_{\theta_2}(z,'z,x,a^0),a^0)|z,x,a^0]+\beta\mathbb{E}_{z'}[\log\sum_{a'\in A}\exp(Q_{t+1,\theta}(z,'\lambda_{\theta_2}(z,'z,x,a^0),a')-Q_{t+1,\theta}(z,'\lambda_{\theta_2}(z,'z,x,a^0),a^0))|z,x,a^0]$. The rest follows exactly as in the proof of Theorem 3. ■

Proof of Theorem 4: Let $\mathcal{D}:X\times X\mapsto\mathbb{R}^+$ be a metric on X, defined as $\mathcal{D}(x,x')\triangleq\max\{d(x,x'),d(x,'x)\}$, where $d(x,x')\triangleq 1-\min\{\frac{x(s)}{x'(s)}:s\in S,x'(s)>0\}\forall x,x'\in X$. Let $\eta(P_{\theta_2}(z,'z,a))=\max\{\mathcal{D}(\lambda_{\theta_2}(z,'z,e_i,a),\lambda_{\theta_2}(z,'z,e_j,a)):i,j\in S\}$, where $e_i\in X$ with 1 on its ith element. The authors in [23] and [24] showed that $\forall x_1,x_2\in X$

$$\mathcal{D}(\lambda_{\theta_2}(z,'z,x_1,a),\lambda_{\theta_2}(z,'z,x_2,a))\le\eta(P_{\theta_2}(z,'z,a))<1$$

where $\eta(P_{\theta_2}(z,'z,a))$ is called a *contraction* coefficient (coefficient of ergodicity) for substochastic matrix $P_{\theta_2}(z,'z,a)$. Then, given a finite history, $\{z_t,\ldots,z_{t-M},a_{t-1},\ldots,a_{t-M}\}$, $\mathcal{D}(\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x_{t-M}),\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x'_{t-M}))\le(\eta(P_{\theta_2}(z,'z,a)))^M\forall x_{t-M},x'_{t-M}\in X$ (see [24, Sec. 2.3]), showing that the effect of x_0 decreases as M increases. The proof of Theorem 2 shows that θ_2 can be uniquely determined by

$$\sigma_{\theta_2}(z_{t+1},z_t,\lambda_{\theta_2}(z_t,z_{t-1},x^*_{t-1},a_{t-1}),a_t)=\hat{\mathbb{P}}(z_{t+1}|h_t,a_t)$$

given the true value x^*_{t-1} is known for h_t. Namely, there exists a unique θ^*_2 such that

$$\sigma_{\theta^*_2}(z_{t+1},z_t,\lambda_{\theta^*_2}(z_t,z_{t-1},x^*_{t-1},a_{t-1}),a_t)=\hat{\mathbb{P}}(z_{t+1}|h_t,a_t)$$

and $\forall\theta_2\neq\theta^*_2$, there is an $\epsilon>0$ such that

$$D(\sigma_{\theta_2}(.,z_t,\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x^*_{t-M}),a_t),\hat{\mathbb{P}}(.|h_t,a_t))\ge\epsilon$$

where x^*_{t-M} is the true belief at $t-M$. Due to the contraction coefficient $\eta(P_{\theta_2}(z,'z,a))<1$, we have

$$D(\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x^*_{t-M}),\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x_{t-M}))\le\eta^M$$

where $\eta=\max_{z,'z,'a}\eta(P_{\theta_2}(z,'z,a))<1$ since $|Z|<\infty,|A|<\infty$. Thus, $\forall\theta_2\neq\theta^*_2$, we have

$$\begin{aligned}&\lim_{M\to\infty}D(\sigma_{\theta_2}(.,z_t,\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x_{t-M}),a_t),\hat{\mathbb{P}}(.|h_t,a_t))\\&=\lim_{M\to\infty}D(\sigma_{\theta_2}(.,z_t,\lambda^M_{\theta_2}(z^t_{t-M},a^{t-1}_{t-M},x^*_{t-M}),a_t),\hat{\mathbb{P}}(.|h_t,a_t))\\&\ge\epsilon,\quad\forall x_{t-M}\in X\end{aligned}$$

indicating θ_2 can be distinguished by the data given M is sufficiently large. Once θ_2 is determined, θ_1 can be determined by Theorem 3, completing the proof. ■

Lemma 1: Under the Assumptions of Theorem 5, $Q_{\theta_1}(z,x,a)$ and $V_{\theta_1}(z,x)$ are also twice continuously differentiable in $\theta_1\in\mathbb{R}^{p_1}$ and $\sup_{\theta_1}\|\nabla^2_{\theta_1}Q_{\theta_1}(z,x,a)\|\le L_Q$, $\sup_{\theta_1}\|\nabla^2_{\theta_1}V_{\theta_1}(z,x)\|\le L_V\forall(z,x,a)\in Z\times X\times A$, where $L_Q\triangleq\frac{1}{1-\beta}L_{r,2}+\frac{2\beta}{(1-\beta)^3}(L_{r,1})^2$, $L_V\triangleq\frac{1}{1-\beta}L_{r,2}+\frac{2}{(1-\beta)^3}(L_{r,1})^2$.

Proof of Lemma 1: For fixed $\theta_1\in\mathbb{R}^{p_1}$, consider the mapping $G^1_{\theta_1}:\mathcal{B}\mapsto\mathcal{B}$ defined as: $[G^1_{\theta_1}g](z,x,a)=\nabla_{\theta_1}r_{\theta_1}(z,x,a)+\beta\sum_{z'}\sigma_{\hat{\theta}_2}(z,'z,x,a)\sum_{a'}\pi_{\theta_1}(a'|z,'x')g(z,'x',a')$, where $x'=\lambda_{\hat{\theta}_2}(z,'z,x,a)$. It follows that $G^1_{\theta_1}$ is a contraction map with unique fixed point $\nabla_{\theta_1}Q_{\theta_1}$ and $\|\nabla_{\theta_1}Q_{\theta_1}(z,x,a)\|\le\frac{1}{1-\beta}L_{r,1}$. Since

$$\nabla_{\theta_1}V_{\theta_1}(z,'x')=\sum_{a'}\pi_{\theta_1}(a'|z,'x')\nabla_{\theta_1}Q_{\theta_1}(z,'x,'a')$$

it follows that $\nabla_{\theta_1}V_{\theta_1}$ also exists and $\|\nabla_{\theta_1}V_{\theta_1}\|\le\frac{1}{1-\beta}L_{r,1}$.

Consider the mapping $G^2_{\theta_1}:\mathcal{B}\mapsto\mathcal{B}$ defined as follows: $[G^2_{\theta_1}g](z,x,a)=\nabla^2_{\theta_1}r_{\theta_1}(z,x,a)+\beta\sum_{z'}\sigma_{\hat{\theta}_2}(z,'z,x,a)\sum_{a'}\nabla_{\theta_1}\pi_{\theta_1}(a'|z,'x')\nabla_{\theta_1}Q_{\theta_1}(z,'x,'a')+\beta\sum_{z'}\sigma_{\hat{\theta}_2}(z,'z,x,a)\sum_{a'}\pi_{\theta_1}(a'|z,'x')g(z,'x,'a')$, where $x'=\lambda_{\hat{\theta}_2}(z,'z,x,a)$. Thus, $G^2_{\theta_1}$ is a contraction map with unique fixed point $\nabla^2_{\theta_1}Q_{\theta_1}$. Since $\nabla^2_{\theta_1}V_{\theta_1}(z,'x')=\sum_{a'}\nabla_{\theta_1}\pi_{\theta_1}(a'|z,'x')\nabla_{\theta_1}Q_{\theta_1}(z,'x,'a')+\sum_{a'}\pi_{\theta_1}(a'|z,'x')\nabla^2_{\theta_1}Q_{\theta_1}(z,'x,'a')$, $\nabla^2_{\theta_1}V_{\theta_1}$ exists, $\|\nabla^2_{\theta_1}Q_{\theta_1}(z,x,a)\|\le\frac{1}{1-\beta}(\|\nabla^2_{\theta_1}r_{\theta_1}(z,x,a)\|+\beta\|\sum_{z'}\sigma_{\hat{\theta}_2}(z,'z,x,a)\sum_{a'}\nabla_{\theta_1}\pi_{\theta_1}(a'|z,'x')\nabla_{\theta_1}Q_{\theta_1}(z,'x,'a')\|)$. Also $\nabla_{\theta_1}\pi_{\theta_1}(a|z,x)=\pi_{\theta_1}(a|z,x)[\nabla_{\theta_1}Q_{\theta_1}(z,x,a)-\sum_{a'}\pi_{\theta_1}(a'|z,x)\nabla_{\theta_1}Q_{\theta_1}(z,x,a')]$.

Hence, $\|\sum_{z'}\sigma_{\hat{\theta}_2}(z',z,x,a)\sum_{a'}\nabla_{\theta_1}\pi_{\theta_1}(a'|z',x')\times\nabla_{\theta_1}Q_{\theta_1}(z',x',a')\|\le\frac{2}{(1-\beta)^2}(L_{r,1})^2$.

Putting all together, we obtain $\|\nabla^2_{\theta_1}Q_{\theta_1}(z,x,a)\|\le\frac{1}{1-\beta}L_{r,2}+\frac{2\beta}{(1-\beta)^3}(L_{r,1})^2=L_Q$. $\|\nabla^2_{\theta_1}V_{\theta_1}\|\le\|\sum_{a'}\nabla_{\theta_1}\pi_{\theta_1}(a'|z',x')\nabla_{\theta_1}Q_{\theta_1}(z',x',a')\|+\|\sum_{a'}\pi_{\theta_1}(a'|z',x')\nabla^2_{\theta_1}Q_{\theta_1}(z',x',a')\|\le\frac{2}{(1-\beta)^3}(L_{r,1})^2+\frac{1}{1-\beta}L_{r,2}=L_V$. ■

Proof of Theorem 5: Recall that $\nabla^2_{\theta_1}\hat{\ell}(\theta_1)=\sum_{i=1}^N\sum_{t=0}^{T-1}\nabla^2_{\theta_1}Q_{\theta_1}(z_{t,i},x_{t,i},a_{t,i})-\nabla^2_{\theta_1}V_{\theta_1}(z_{t,i},x_{t,i})$. By Lemma 1, it follows that $\|\nabla^2_{\theta_1}\hat{\ell}(\theta_1)\|\le L$ with $L\triangleq NT(L_Q+L_V)$ Or equivalently, $\nabla_{\theta_1}\hat{\ell}(\theta_1)$ is Lipschitz continuous in θ_1 with constant L.

By Lipschitz continuous gradients, $\hat{\ell}(\theta_1^{k+1})\ge\hat{\ell}(\theta_1^k)+\nabla\hat{\ell}(\theta_1^k)^\top(\theta_1^{k+1}-\theta_1^k)-\frac{L}{2}\|\theta_1^{k+1}-\theta_1^k\|^2=\hat{\ell}(\theta_1^k)+\rho\left(1-\frac{\rho L}{2}\right)\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\|^2$. Hence, $\rho\left(1-\frac{\rho L}{2}\right)\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\|^2\le\hat{\ell}(\theta_1^{k+1})-\hat{\ell}(\theta_1^k)$. Adding over $k=1,\ldots,K$ we obtain $\rho\left(1-\frac{\rho L}{2}\right)\sum_{k=1}^K\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\|^2\le\hat{\ell}(\theta_1^K)-\hat{\ell}(\theta_1^0)\le\hat{\ell}(\theta_1^*)-\hat{\ell}(\theta_1^0)$, where θ_1^* is a maximizerof log-likelihood. It follows that $\frac{1}{K}\sum_{k=1}^K\|\nabla_{\theta_1}\hat{\ell}(\theta_1^k)\|^2\le\frac{1}{\rho(1-\frac{\rho L}{2})}\frac{\hat{\ell}(\theta_1^*)-\hat{\ell}(\theta_1^0)}{K}$. ■

REFERENCES

[1] J. Choi and K. Kim, "Inverse reinforcement learning in partially observable environments," *J. Mach. Learn. Res.*, vol. 12, pp. 691–730, 2011.

[2] T. Osa et al., "An algorithmic perspective on imitation learning," *Foundations Trends Robot.*, vol. 7, no. 1/2, pp. 1–179, 2018.

[3] N. Tishby and D. Polani, "Information theory of decisions and actions," in *Perception-Action Cycle*, Berlin, Germany: Springer, 2011, pp. 601–636.

[4] P. A. Ortega and D. A. Braun, "Thermodynamics as a theory of decision-making with information-processing costs," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 469, no. 2153, 2013, Art. no. 20120683.

[5] F. Matějka and A. McKay, "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *Amer. Econ. Rev.*, vol. 105, pp. 272–98, Jan. 2015.

[6] L. P. Hansen and J. Miao, "Aversion to ambiguity and model misspecification in dynamic stochastic environments," *Proc. Nat. Acad. Sci.*, vol. 115, no. 37, pp. 9163–9168, 2018.

[7] J. Rust, "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica*, vol. 55, no. 5, pp. 999–1033, 1987.

[8] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. 23rd AAAI Conf. Artif. Intell.*, 2008, pp. 1433–1438.

[9] A. Boularias, J. Kober, and J. Peters, "Relative entropy inverse reinforcement learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, vol. 15, pp. 182–189.

[10] C. Finn, S. Levine, and P. Abbeel, "Guided cost learning: Deep inverse optimal control via policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 49–58.

[11] M. Tappler, B. K. Aichernig, G. Bacci, M. Eichlseder, and K. G. Larsen, "l^*-based learning of Markov decision processes (extended version)," *Formal Aspects Comput.*, vol. 33, pp. 575–615, 2021.

[12] H. Mao, Y. Chen, M. Jaeger, T. D. Nielsen, K. G. Larsen, and B. Nielsen, "Learning deterministic probabilistic automata from a model checking perspective," *Mach. Learn.*, vol. 105, pp. 255–299, 2016.

[13] H. Kasahara and K. Shimotsu, "Estimation of discrete choice dynamic programming models," *J. Appl. Econometrics*, vol. 69, no. 1, pp. 28–58, 2018.

[14] V. J. Hotz and R. A. Miller, "Conditional choice probabilities and the estimation of dynamic models," *Rev. Econ. Stud.*, vol. 60, no. 3, pp. 497–529, 1993.

[15] V. Aguirregabiria and P. Mira, "Dynamic discrete choice structural models: A survey," *J. Econometrics*, vol. 156, pp. 38–67, 2010.

[16] C. L. Su and K. L. Judd, "Constrained optimization approaches to estimation of structural models," *Econometrica*, vol. 80, no. 5, pp. 2213–2230, 2012.

[17] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1352–1361.

[18] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[19] T. Magnac and D. Thesmar, "Identifying dynamic discrete choice processes," *Econometrica*, vol. 70, no. 2, pp. 801–816, 2002.

[20] K. Kim, S. Garg, K. Shiragur, and S. Ermon, "Reward identification in inverse reinforcement learning," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 5496–5505.

[21] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Res.*, vol. 21, no. 5, pp. 1071–1088, 1973.

[22] W. Xu and L. Cao, "Optimal maintenance control of machine tools for energy efficient manufacturing," *Int. J. Adv. Manuf. Technol.*, vol. 104, no. 9, pp. 3303–3311, 2019.

[23] L. K. Platzman, "Optimal infinite-horizon undiscounted control of finite probabilistic systems," *SIAM J. Control Optim.*, vol. 18, pp. 362–380, 1980.

[24] C. C. White and W. T. Scherer, "Finite-memory suboptimal design for partially observed Markov decision processes," *Operations Res.*, vol. 42, no. 3, pp. 439–455, 1994.