2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Convergence and Near Optimality of Q-Learning with Finite Memory for Partially Observed Models

Ali Devran Kara and Serdar Yüksel

***Abstract*— Q learning algorithm is a popular reinforcement learning method for finite state/action fully observed Markov decision processes (MDPs). In this paper, we make two contributions: (i) we establish the convergence of a Q learning algorithm for partially observed Markov decision processes (POMDPs) using a finite history of past observations and control actions and show that the limit fixed point equation gives an optimal solution for an approximate belief-MDP. We then provide bounds on the performance of the policy obtained using the limit Q values compared to the performance of the optimal policy for the POMDP, where we also present explicit performance guarantees using recent results on filter stability in controlled POMDPs. (ii) We apply these results to fully observed MDPs with continuous state spaces and establish the near optimality of learned policies via quantization of the state space, where the quantization is viewed as a measurement channel leading to a POMDP model and a history of unit window size is considered. In particular, we show that Q-learning, with its convergence and near optimality properties, is applicable for continuous space MDPs when the state space is quantized.**

I. Introduction

A stochastic control model where the controller can only see a noisy version of the state, is called a Partially Observed Markov Decision Process (POMDP). POMDPs offer a practically rich and relevant, yet mathematically challenging, model. Even in the most basic setup of finite state-action models, the analysis and computation of optimal solutions are complicated.

On approximation methods. The problem of approximate optimality is significantly more challenging for POMDPs compared to the fully observed MDP counterpart. Most of the studies in the literature are algorithmic and computational contributions with few rigorous analytical results. These include [12], [23], [20]. For partially observed setups, [16], building on [15], introduces a rigorous approximation analysis (and explicit methods for quantization of probability measures) and shows that finite model approximations obtained through quantization are asymptotically optimal. [18] presents a notion of approximate information variable and studies near optimality of policies that satisfies the approximate information state property. We refer the reader to the survey papers [9], [21], [3] and the recent book [7] for further structural results as well as algorithmic and computational methods for approximating POMDPs.

On learning for POMDPs. Learning in POMDPs is challenging: if one attempts to learn optimal policies through empirical observations, the analysis and convergence properties become significantly harder to obtain (compared with MDPs) as the observations progress in a non-Markovian fashion even under a memoryless control policy, and the belief state space is uncountable (as it is a space of probability measures). [5] studies a learning algorithm for POMDPs with average cost criteria where a policy improvement method is proposed using random polices and the convergence of this method to local optima is given. [10] and [8] are studies that propose a similar general approach as we present in this paper, however these only provide extensive experimental or numerical results without analytical convergence or rigorous approximation results.

A natural, though optimistic, attempt to learn POMDPs would be to ignore the partial observability and pretend that the noisy observations reflect the true state perfectly. For example, for infinite horizon discounted cost problems, one can construct Q iterations as:

$$Q_{k+1}(y_k,u_k) = (1-\alpha_k(y_k,u_k))Q_k(y_k,u_k) + \alpha_k(y_k,u_k)\left(C_k(y_k,u_k) + \beta \min_v Q_k(Y_{k+1},v)\right) \quad (1)$$

where y_k represents the observations and u_k represents the control actions, $0 < \beta < 1$ is the discount factor, and α_k's are the learning rates. We can further improve this algorithm by using not only the most recent observation but a finite window of past observations and control actions. However, the joint observation and control process is not a controlled Markov process (as only (X_k, U_k) is), and hence the convergence does not directly follow from standard techniques ([4], [19]).

Even if convergence is guaranteed, it is not clear what the limit Q values are, and whether they are meaningful at all. [17] studies the Q learning algorithm for POMDPs by ignoring the partial observability and constructing the algorithm using the most recent observation variable as in (1), and establishes convergence of this algorithm under mild conditions. In our paper, we will consider memory sizes of more than zero for the information variables and a continuous state space, and thus the algorithm in [17] can be seen as a special case of our setup. Different from our work, [17] does not study what the limit of the Q iterations mean, and in particular whether the limit equation corresponds to some approximate MDP model. Furthermore, using longer window sizes reveals the intimate connection between the approximate learning problem and the nonlinear controlled filter stability problem that we will study in detail.

Contributions. (i) In Theorem 2, we show that the Q iterations constructed using finite history variables converge under mild assumptions on the hidden state process, and the limit fixed point equation corresponds to an optimal solution for an approximate belief-MDP model. (ii) We, towards a practically consequential goal, in Theorem 1 establish

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

The authors are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada, Email: {16adk,yuksel}@mast.queensu.ca

bounds for the performance loss of the policy obtained using the approximate belief-MDP when it is used in the original model. This also establishes that under explicit filter stability conditions to be presented, one can guarantee near optimality of the presented algorithm. (iii) In Section V, we study the Q learning problem for fully observed models with continuous state spaces. We show that the problem can be seen as a POMDP with a quantizer channel if we discretize the state space and that Q-learning can be applied to continuous state spaces via our POMDP analysis with its convergence and near optimality results under guaranteed performance bounds.

II. Partially Observed Markov Decision Processes and Belief-MDP Reduction

Let $\mathbb{X} \subset \mathbb{R}^m$ denote a Borel set which is the state space of a partially observed controlled Markov process for some $m \in \mathbb{N}$. Let $\mathbb{Y}$ be a finite set denoting the observation space of the model, and let the state be observed through an observation channel O. The observation channel, O, is defined as a stochastic kernel (regular conditional probability) from $\mathbb{X}$ to $\mathbb{Y}$, such that $O(\cdot\,|x)$ is a probability measure on $\mathbb{Y}$ for every $x \in \mathbb{X}$, and $O(A|\,\cdot\,) : \mathbb{X} \to [0,1]$ is a Borel measurable function for every $A \subset \mathbb{Y}$. $\mathbb{U}$ denotes the action space which is also a finite set.

An *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables $I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \qquad I_0 = \{Y_0\}$, where $U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+$, are the $\mathbb{U}$-valued control actions and $Y_{[0,t]} = \{Y_s,\, 0 \le s \le t\}, \quad U_{[0,t-1]} = \{U_s,\, 0 \le s \le t-1\}$. We define Γ to be the set of all such admissible policies. The update rules of the system are determined by relationships:

$$\Pr\big((X_0, Y_0) \in B\big) = \int_B \mu(dx_0) O(dy_0|x_0), \quad B \in \mathcal{B}(\mathbb{X}\times\mathbb{Y}),$$

where μ is the (prior) distribution of the initial state X_0, and

$$\Pr\Big((X_t, Y_t) \in B \,\Big|\, (X,Y,U)_{[0,t-1]} = (x,y,u)_{[0,t-1]}\Big) = \textstyle\int_B T(dx_t|x_{t-1}, u_{t-1}) O(dy_t|x_t),$$

$B \in \mathcal{B}(\mathbb{X}\times\mathbb{Y}), t \in \mathbb{N}$, where T is the transition kernel of the model which is a stochastic kernel from $\mathbb{X}\times\mathbb{U}$ to $\mathbb{X}$. Note that, although $\mathbb{Y}$ is finite, we here use an integral sign instead of the summation sign for notational convenience.

We let the objective of the agent (decision maker) be the minimization of the infinite horizon discounted cost,

$$J_\beta(\mu, T, \gamma) = E_\mu^{T,\gamma}\left[\sum_{t=0}^{\infty} \beta^t c(X_t, U_t)\right] \tag{2}$$

for some discount factor $\beta \in (0,1)$, over the set of admissible policies $\gamma \in \Gamma$, where $c : \mathbb{X}\times\mathbb{U} \to \mathbb{R}$ is a Borel-measurable stage-wise cost function and $E_\mu^{T,\gamma}$ denotes the expectation with initial state probability measure μ and transition kernel T under policy γ. Note that $\mu \in \mathcal{P}(\mathbb{X})$, where we let $\mathcal{P}(\mathbb{X})$ denote the set of probability measures on $\mathbb{X}$. We define

$$J_\beta^*(\mu, T) = \inf_{\gamma\in\Gamma} J_\beta(\mu, T, \gamma).$$

For the analysis of POMDPs, a common approach is to reformulate the problem as a fully observed MDP, where the decision maker keeps track of the posterior distribution of the state X_t given the available history I_t. In the following section, we formalize this approach.

A. Reduction to fully observed models using belief states

It is by now a standard result that, for optimality analysis, any POMDP can be reduced to a completely observable Markov decision process ([22], [13]), whose states are the posterior state distributions or *beliefs* of the observer or the filter process; that is, the state at time t is

$$z_t := Pr\{X_t \in \cdot\,|Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(\mathbb{X}). \tag{3}$$

We call this equivalent process the filter process. The filter process has state space $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Here, $\mathcal{Z}$ is equipped with the Borel σ-algebra generated by the topology of weak convergence [1]. Then, the transition probability η of the filter process can be constructed as follows. If we define the measurable function

$$F(z,u,y) := Pr\{X_{t+1} \in \cdot\,|Z_t = z, U_t = u, Y_{t+1} = y\}$$

from $\mathcal{P}(\mathbb{X})\times\mathbb{U}\times\mathbb{Y}$ to $\mathcal{P}(\mathbb{X})$ and use the stochastic kernel $P(\cdot\,|z,u) = \Pr\{Y_{t+1} \in \cdot\,|Z_t = z, U_t = u\}$ from $\mathcal{P}(\mathbb{X})\times\mathbb{U}$ to $\mathbb{Y}$, we can write η as

$$\eta(\,\cdot\,|z,u) = \int_{\mathbb{Y}} 1_{\{F(z,u,y)\in\,\cdot\,\}} P(dy|z,u). \tag{4}$$

The one-stage cost function $\tilde{c} : \mathcal{P}(\mathbb{X})\times\mathbb{U} \to [0,\infty)$ of the filter process is given by

$$\tilde{c}(z,u) := \int_{\mathbb{X}} c(x,u) z(dx), \tag{5}$$

which is a Borel measurable function. Hence, the filter process is a completely observable Markov process with the components $(\mathcal{Z}, \mathbb{U}, \tilde{c}, \eta)$.

Even though the belief-MDP reduction approach provides a powerful analytical tool for the analysis of POMDPs, computational challenges are formidable: The belief space $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ is uncountable even if $\mathbb{X}$ were finite, and the computation of the belief state $Pr(X_t \in \cdot|I_t)$ is numerically demanding. Therefore, some approximation of the belief-MDP is usually needed.

In the following section, we provide an alternative MDP reduction and present rigorous approximation results that only make use of a finite history of the information variables and lead to a finite dimensional implementation for near optimality.

III. An Alternative Finite Window Belief-MDP Reduction and its Approximation

A. An alternative finite window belief-MDP reduction

In this section we construct an alternative fully observed MDP reduction using the predictor from N stages earlier and the most recent N information variables (that is, measurements and actions). This new construction allows us to highlight the most recent information variables and *compress* the information coming from the past history via the predictor (to be defined below) as a probability measure valued variable. In what follows, we will sometimes consider the

case with $N = 1$ to make the presentation less complicated. The general case follows from identical steps.

Consider the following state variable at time t:

$$\hat{z}_t = (\pi^-_{t-N}, I^N_t) \tag{6}$$

where, for $N \geq 1$

$$\pi^-_{t-N} = Pr(X_{t-N} \in \cdot | y_{t-N-1}, \dots, y_0, u_{t-N-1}, \dots, u_0),$$
$$I^N_t = \{y_t, \dots, y_{t-N}, u_{t-1}, \dots, u_{t-N}\}$$

and $I^N_t = y_t$ for $N = 0$ with μ being the prior probability measure on X_0. We will refer to π^-_{t-N} as the *predictor* at $t - N$. The state space with this representation is $\hat{\mathcal{Z}} = \mathcal{P}(\mathbb{X}) \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$ where we equip $\hat{\mathcal{Z}}$ with the product topology where we consider the weak convergence topology on the $\mathcal{P}(\mathbb{X})$ coordinate and the usual (discrete) topologies on $\mathbb{Y}^{N+1} \times \mathbb{U}^N$ coordinates.

This new state representation can be mapped to the belief state z_t defined in (3). Consider the map $\psi : \hat{\mathcal{Z}} \to \mathcal{P}(\mathbb{X})$, for some $\hat{z}_t = (\pi^-_{t-N}, I^N_t)$

$$\psi(\hat{z}_t) = \psi(\pi^-_{t-N}, I^N_t) = P^{\pi^-_{t-N}}(X_t \in \cdot | I^N_t)$$
$$= P^{\mu}(X_t \in \cdot | y_t, \dots, y_0, u_{t-1}, \dots, u_0) = z_t$$

such that the map ψ acts as a Bayesian update of π^-_{t-N} using I^N_t. Using this map, we can define the stage-wise cost function and the transition probabilities. Consider the new cost function $\hat{c} : \hat{\mathcal{Z}} \times \mathbb{U} \to \mathbb{R}$, using the cost function $\tilde{c}$ of the belief MDP (defined in (5)) such that

$$\hat{c}(\hat{z}_t, u_t) = \hat{c}(\pi^-_{t-N}, I^N_t, u_t) = \tilde{c}(\psi(\pi^-_{t-N}, I^N_t), u_t)$$
$$= \int_{\mathbb{X}} c(x_t, u_t) P^{\pi^-_{t-N}}(dx_t | y_t, \dots, y_{t-N}, u_{t-1}, \dots, u_{t-N}). \tag{7}$$

Furthermore, we can define the transition probabilities for $N = 1$ as follows: for some $A \in \mathcal{B}(\hat{Z})$ such that

$$A = B \times \{\hat{y}_{t-N+1}, \hat{u}_t, \dots, \hat{u}_{t-N+1}\}, \quad B \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$$

we write

$$\begin{aligned}
&Pr(\hat{z}_{t+1} \in A | \hat{z}_t, \dots, \hat{z}_0, u_t, \dots, u_0) \\
&= Pr(\pi^-_t \in B, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \pi^-_{[t-1,0]}, y_{[t,0]}, u_{[t,0]}) \\
&= \mathbb{1}_{\{y_t, u_t = \hat{y}_t, \hat{u}_t, G(\pi^-_{t-1}, y_{t-1}, u_{t-1}) \in B\}} \\
&\qquad P^{\pi^-_{t-1}}(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}) \\
&= Pr(\pi^-_t \in B, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \pi^-_{t-1}, y_t, y_{t-1}, u_t, u_{t-1}) \\
&= Pr(\hat{z}_{t+1} \in A | \hat{z}_t, u_t) =: \int_A \hat{\eta}(d\hat{z}_{t+1} | \hat{z}_t, u_t).
\end{aligned}$$

where the map G is defined as

$$G(\pi^-_{t-1}, y_{t-1}, u_{t-1}) = P^{\mu}(X_t \in \cdot | y_{t-1}, \dots, y_0, u_{t-1}, \dots, u_0).$$

Hence, we have a proper fully observed MDP, with the cost function $\hat{c}$, transition kernel $\hat{\eta}$ and the state space $\hat{\mathcal{Z}}$.

Note that any policy $\phi : \mathcal{P}(\mathbb{X}) \to \mathbb{U}$ defined for the belief MDP, can be extended to the newly defined finite window belief-MDP using the map ψ, and defining $\hat{\phi} := \phi \circ \psi$. Thus, if an optimal policy can be found for the belief MDP, say ϕ^*, the policy $\hat{\phi}^* = \phi^* \circ \psi$ is an optimal policy for the newly defined MDP.

We now write the discounted cost optimality equation for the newly constructed finite window belief MDP.

$$J^*_\beta(\hat{z}) = \min_{u \in \mathbb{U}} \left(\hat{c}(\hat{z}, u) + \beta \int J^*_\beta(\hat{z}_1) \hat{\eta}(d\hat{z}_1 | \hat{z}, u) \right).$$

B. Approximation of the finite window belief-MDP

We now approximate the MDP constructed in the previous section. Consider the following set for a fixed $\pi^* \in \mathcal{P}(\mathbb{X})$ denoted by $\hat{\mathcal{Z}}^N_{\pi^*}$:

$$\left\{ \pi^*, y_{[0,N]}, u_{[0,N-1]} : y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N \right\}$$

such that the state at time t is $\hat{z}^N_t = (\pi^*, I^N_t)$. Compared to the state $\hat{z}_t = (\pi^-_{t-N}, I^N_t)$ defined in (6), this approximate model uses π^* as the predictor, no matter what the real predictor at time $t - N$ is.

The approximate cost function is defined as

$$\hat{c}(\hat{z}^N_t, u_t) = \hat{c}(\pi^*, I^N_t, u_t) = \tilde{c}(\psi(\pi^*, I^N_t), u_t)$$
$$= \int_{\mathbb{X}} c(x_t, u_t) P^{\pi^*}(dx_t | y_t, \dots, y_{t-N}, u_{t-1}, \dots, u_{t-N}).$$

We define the approximate controlled transition kernel by, with $\hat{z}^N_{t+1} = (\pi^*, I^N_{t+1})$ and $\hat{z}^N_t = (\pi^*, I^N_t)$,

$$\hat{\eta}^N(\hat{z}^N_{t+1} | \hat{z}^N_t, u_t) = \hat{\eta}^N(\pi^*, I^N_{t+1} | \pi^*, I^N_t, u_t)$$
$$:= \hat{\eta}\Big(\mathcal{P}(\mathbb{X}), I^N_{t+1} | \pi^*, I^N_t, u_t \Big). \tag{8}$$

Denoting the optimal value function for the approximate model by J^N_β, we can write the following fixed point equation

$$\begin{aligned}
J^N_\beta(\hat{z}^N) \quad = \quad & \min_{u \in \mathbb{U}} \Bigg(\hat{c}(\hat{z}^N, u) \\
& + \beta \sum_{\hat{z}^N_1 \in \hat{\mathcal{Z}}^N_{\pi^*}} J^N_\beta(\hat{z}^N_1) \hat{\eta}^N(\hat{z}^N_1 | \hat{z}^N, u) \Bigg).
\end{aligned} \tag{9}$$

Since we have a finite model in this approximate setup, there exists an optimal policy ϕ^N that satisfies this fixed point equation. Note that both J^N_β and ϕ^N are defined on the finite set $\hat{\mathcal{Z}}^N_{\pi^*}$. However, we can simply extend them to the set $\hat{\mathcal{Z}}$ by defining for any $\hat{z} = (\pi, y_1, y_0, u_0) \in \hat{\mathcal{Z}}$ (here, with $N = 1$)

$$\tilde{J}^N_\beta(\hat{z}) = \tilde{J}^N_\beta(\pi, y_1, y_0, u_0) := J^N_\beta(\pi^*, y_1, y_0, u_0)$$
$$\tilde{\phi}^N(\hat{z}) = \tilde{\phi}^N(\pi, y_1, y_0, u_0) := \phi^N(\pi^*, y_1, y_0, u_0).$$

We will later prove that Q-value iterations using finite window of information variables converge to the Q-values for the approximate model constructed in this section.

In what follows, we investigate the difference $J_\beta(\hat{z}, \tilde{\phi}^N) - J^*_\beta(\hat{z})$, that is the loss occurring from applying the approximate policy on the original model. Before the result, we introduce the following definition and notation.

Definition 1. For probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$, the *total variation* metric is given by

$$\|\mu - \nu\|_{TV} = \sup_{f : \|f\|_\infty \leq 1} \left| \int f(x) \mu(\mathrm{d}x) - \int f(x) \nu(\mathrm{d}x) \right|.$$

We now define the following filter stability term which establishes an insensitivity bound for different initializations of the filter process

$$L_t := \sup_{\hat{\gamma}\in\hat{\Gamma}} E^{\hat{\gamma}}_{\pi_0^-}[\|P^{\pi_t^-}(X_{t+N}\in\cdot|Y_{[t,t+N]},U_{[t,t+N-1]}) - P^{\hat{\pi}}(X_{t+N}\in\cdot|Y_{[t,t+N]},U_{[t,t+N-1]})\|_{TV}]. \quad (10)$$

The expectation is with respect to the realizations of π_t^- and $Y_{[t,t+N]}, U_{[t,t+N-1]}$ under the true dynamics of the system when the prior distribution of x_0 is given by π_0^-. The proof of the following result can be found in [6, Appendix C].

Theorem 1.

$$\sup_{\hat{z}\in\hat{\mathcal{Z}}}\left|J_\beta(\hat{z},\tilde{\phi}^N) - J^*_\beta(\hat{z})\right| \le \frac{2\|c\|_\infty}{(1-\beta)}\sum_{t=0}^{\infty}\beta^t L_t.$$

where L is defined as in (10).

IV. Q Iterations Using a Finite History of Information Variables and Convergence

Assume that we start keeping track of the information variables

$$I_t^N = \begin{cases} \{y_t, y_{t-1},\dots,y_{t-N},u_{t-1},\dots,u_{t-N}\} & \text{if } N>0 \\ y_t \quad \text{if } N=0. \end{cases}$$

We will construct the Q-value iteration using these information variables. In what follows, we will drop the N dependence in I_t^N, and we take $N=1$ for simplicity of notation. For these new approximate states, we follow the usual Q learning algorithm such that for any $I\in\mathbb{Y}^{N+1}\times\mathbb{U}^N$ and $u\in\mathbb{U}$

$$Q_{k+1}(I,u) = (1-\alpha_k(I,u))Q_k(I,u) + \alpha_k(I,u)\left(C_k(I,u) + \beta\min_v Q_k(I_1^k,v)\right), \quad (11)$$

where $I_1^k = \{Y_{t+1}, y_t,\dots,y_{t-N+1},u_t,\dots,u_{t-N+1}\}$, we put the k dependence to emphasize that the distribution of Y_{t+1} and hence I_1^k are different for every k, the time we hit $\{y_t,y_{t-1},\dots,y_{t-N},u_{t-1},\dots,u_{t-N}\}$ for the k-th time.

To choose the control actions, we use polices that choose the control actions randomly and independent of everything else such that at time t $u_t=u_i$, w.p σ_i for any $u_i\in\mathbb{U}$ with $\sigma_i>0$ for all i.

The algorithm differs from the usual Q-value iteration:

(i) The distribution of I_1^k, which is the consecutive N-window information variable when we hit the (I,u) pair for the k-th time, is generally different for every k and the pair (I,u) is not a controlled Markov process. Furthermore, the controlled transitions are time dependent.

(ii) Here, the cost we observe is $c(x_t,u_t)$ (which is not a direct function of measurements), where $c(x_t,u_t)$ depends on (I_t,u_t) pair randomly and in a time-dependent fashion.

We will observe that if one assumes that the hidden state process $\{x_t\}$ is positive Harris recurrent and in particular admits a unique invariant probability measure, say π^*, under some memoryless randomized exploration policy γ, then the average of the approximate state transitions converges to

$$P^*(I_{t+1}|I_t,u_t) := \hat{\eta}^N((\pi^*,I_{t+1})|(\pi^*,I_t),u_t) \quad (12)$$

with $\hat{\eta}^N$ defined as in (8).

We also have that the sample path averages of the random cost realizations converge to,

$$C^*(I,u) = \hat{c}(\pi^*,I,u) = \int_{\mathbb{X}} c(x,u)P^{\pi^*}(dx|I) \quad (13)$$

where, $P^*(x|I)$ is the Bayesian update of π^*, using I and $\hat{c}(\pi^*,I,u)$ is defined as in (7).

Now consider the following fixed point equation

$$Q^*(I,u) = C^*(I,u) + \beta\sum_{I'}P^*(I'|I,u)\min_v Q^*(I',v) \quad (14)$$

where P^* is defined in (12) and C^* is defined in (13).

For the rest of the paper, we will use the following notation

$$V^*(I) := \min_{v\in\mathbb{U}} Q^*(I,v), \quad V_t(I) := \min_{v\in\mathbb{U}} Q_t(I,v). \quad (15)$$

We note that π^*, P^*, and C^* do not have to be calculated by the controller. We will show that the algorithm converges to (14), when the hidden state process is positive Harris recurrent.

Assumption 1.

1. *$\alpha_t(I,u)=0$ unless $(I_t,u_t)=(I,u)$. Furthermore,*
$$\alpha_t(I,u) = \frac{1}{1+\sum_{k=0}^t 1_{\{I_k=I,u_k=u\}}}$$
We note that this means $\alpha_k(I,u)=\frac{1}{k}$ if $I_k=I,u_k=u$, when k is the instant of the kth visit to (I,u).
2. *Under every memoryless policy, say γ, the hidden state process $\{X_t\}$ is positive Harris recurrent and in particular admits a unique invariant measure π^*_γ.*
3. *During the exploration phase every (I,u) pair is visited infinitely often.*

Theorem 2. *Under Assumption 1,*

i. The algorithm given in (11) converges almost surely to Q^* which satisfies (14).

ii. For any policy γ^N that satisfies $Q^*(I,\gamma^N(I)) = \min_u Q^*(I,u)$, if we assume that the controller has access to at least $N+1$ observations and N control action variables, when it starts acting, we have

$$J_\beta(\mu,T,\gamma^N) - J^*_\beta(\mu,T) \le \frac{2\|c\|_\infty}{(1-\beta)}\sum_{t=0}^{\infty}\beta^t L_t$$

where L_t is defined in (10).

Proof Sketch. We first prove that the process Q_k, determined by the algorithm in (11), converges almost surely to Q^*. We define

$$\Delta_k(I,u) := Q_k(I,u) - Q^*(I,u)$$
$$F_k(I,u) := C_k(I,u) + \beta V_k(I_1^k) - Q^*(I,u)$$
$$\hat{F}_k(I,u) := C^*(I,u) + \beta\sum_{I_1}V_k(I_1)P^*(I_1|I,u) - Q^*(I,u),$$

where V_k is defined in (15). Then, we can write the following iteration

$$\Delta_{k+1}(I,u) = (1-\alpha_k(I,u))\Delta_k(I,u) + \alpha_k(I,u)F_k(I,u).$$

Now, we write $\Delta_k = \delta_k + w_k$ such that

$$\delta_{k+1}(I,u) = (1-\alpha_k(I,u))\delta_k(I,u) + \alpha_k(I,u)\hat{F}_k(I,u)$$
$$w_{k+1}(I,u) = (1-\alpha_k(I,u))w_k(I,u) + \alpha_k(I,u)r_k(I,u)$$

where $r_k := F_k - \hat{F}_k = \beta V_k(I_1^k) - \beta\sum_{I_1} V_k(I_1)P^*(I_1|I,u) + C_k(I,u) - C^*(I,u)$. Next, we define

$$r_k^*(I,u) = \beta V^*(I_1^k) - \beta\sum_{I_1} V^*(I_1)P^*(I_1|I,u) + C_k(I,u) - C^*(I,u)$$

We further separate $w_k = u_k + v_k$ such that

$$u_{k+1}(I,u) = (1-\alpha_k(I,u))u_k(I,u) + \alpha_k(I,u)e_k(I,u)$$
$$v_{k+1}(I,u) = (1-\alpha_k(I,u))v_k(I,u) + \alpha_k(I,u)r_k^*(I,u)$$

where $e_k = r_k - r_k^*$.

In [6, Appendix A], it is shown that $v_k(I,u) \to 0$ almost surely for all (I,u). Now, we go back to the iterations:

$$\delta_{k+1}(I,u) = (1-\alpha_k(I,u))\delta_k(I,u) + \alpha_k(I,u)\hat{F}_k(I,u)$$
$$u_{k+1}(I,u) = (1-\alpha_k(I,u))u_k(I,u) + \alpha_k(I,u)e_k(I,u)$$
$$v_{k+1}(I,u) = (1-\alpha_k(I,u))v_k(I,u) + \alpha_k(I,u)r_k^*(I,u).$$

Note that, we want to show $\Delta_k = \delta_k + u_k + v_k \to 0$ almost surely. The following analysis holds for any path that belongs to the probability one event in which $v_k(I,u) \to 0$. For any such path and for any given $\epsilon > 0$, we can find an $N < \infty$ such that $\|v_k\|_\infty < \epsilon$ for all $k > N$. We now focus on the term $\delta_k + u_k$ for $k > N$:

$$(\delta_{k+1} + u_{k+1})(I,u) = (1-\alpha_k(I,u))(\delta_k + u_k)(I,u) + \alpha_k(I,u)(\hat{F}_k + e_k)(I,u). \quad (16)$$

Observe that for $k > N$, having that $v_k \to 0$ almost surely,

$$(\hat{F}_k + e_k)(I,u) = (F_k - r_k^*)(I,u) \le \beta\max_{I,u}|Q_k(I,u) - Q^*(I,u)| = \beta\|\Delta_k\|_\infty \le \beta\|\delta_k + u_k\|_\infty + \beta\epsilon.$$

By choosing $C < \infty$ such that $\hat{\beta} := \beta(C+1)/C < 1$, for $\|\delta_k + u_k\|_\infty > C\epsilon$, we can write that $\beta\|\delta_k + u_k + \epsilon\|_\infty \le \hat{\beta}\|\delta_k + u_k\|_\infty$. Now with (16),

$$\begin{aligned}(\delta_{k+1} + u_{k+1})(I,u) &= (1-\alpha_k(I,u))(\delta_k + u_k)(I,u) \\ &\quad + \alpha_k(I,u)(\hat{F}_k + e_k)(I,u) \\ &\le (1-\alpha_k(I,u))(\delta_k + u_k)(I,u) + \alpha_k(I,u)\hat{\beta}\|\delta_k + u_k\|_\infty \\ &< \|\delta_k + u_k\|_\infty\end{aligned}$$

Hence $(\delta_{k+1} + u_{k+1})(I,u)$ clearly converges to 0 for $\|\delta_k + u_k\|_\infty > C\epsilon$. One can also show that once the process hits below $C\epsilon$ it always stays there. Thus, taking $\epsilon \to 0$, we can conclude that $\Delta_k = \delta_k + u_k + v_k \to 0$ almost surely.

For item (ii), notice that (14) coincides with the DCOE for the approximate belief MDP defined in (9). Hence, using Theorem 1, we conclude the result for a policy that satisfies $Q^*(I, \gamma^N(I)) = \min_u Q^*(I,u)$. □

A. Convergence to Near Optimality under Filter Stability

Here, we study the L_t term defined in (10).

Definition 2. [2, Equation 1.16] For a kernel operator $K : S_1 \to \mathcal{P}(S_2)$ (that is a regular conditional probability from S_1 to S_2) for standard Borel spaces S_1, S_2, we define the Dobrushin coefficient as:

$$\delta(K) = \inf\sum_{i=1}^{n}\min(K(x,A_i), K(y,A_i)) \quad (17)$$

where the infimum is over all $x, y \in S_1$ and all partitions $\{A_i\}_{i=1}^n$ of S_2.

Example 1. For the following stochastic transition matrix

$$K = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & 0 & \frac{1}{4} \end{pmatrix}$$

The Dobrushin coefficient is the minimum over any two rows where we sum the minimum elements among those rows. For this example, the first and the second rows give $\frac{2}{3}$, the first and the third rows give $\frac{7}{12}$ and the second and the third rows give $\frac{1}{4}$. Then the Dobrushin coefficient is $\frac{1}{4}$.

Let $\tilde{\delta}(T) := \inf_{u\in\mathbb{U}} \delta(T(\cdot|\cdot,u))$.

Theorem 3. *[11, Theorem 3.3] Assume that for $\mu, \nu \in \mathcal{P}(\mathbb{X})$, we have $\mu \ll \nu$ and that $\alpha := (1-\tilde{\delta}(T))(2-\delta(O)) < 1$. Then, for any $\rho < \frac{1}{\alpha}$, and for any realizable $y_{[0,t]}, u_{[0,t-1]}$ under μ and under some policy γ, we have*

$$L_t \le 2\alpha^N \quad (18)$$

for all t where L_t is defined in (10).

The following result is a direct corollary of Theorem 2 and Theorem 3.

Corollary IV.1 (to Theorem 2 and 3)**.** *If we assume: (i) Assumption 1 holds, (ii) $\mathbb{X} \subset \mathbb{R}^m$ for some $m < \infty$, (iii) the transition kernel $T(\cdot|x_0,u_0)$ admits a density function f with respect to a measure ϕ such that $T(dx_1|x_0,u_0) = f(x_1,x_0,u_0)\phi(dx_1)$ and $f(x_1,x_0,u_0) > 0$ for all x_1, x_0, u_0, (iv) $\alpha := (1-\tilde{\delta}(T))(2-\delta(O)) < 1$, then for any policy γ^N that satisfies $Q^*(I,\gamma^N(I)) = \min_u Q^*(I,u)$, we have*

$$|J_\beta(\mu,T,\gamma^N) - J_\beta^*(\mu,T)| \le 2\alpha^N.$$

V. Application: Reinforcement Learning for Continuous Space MDPs via Finite State Approximations with Quantization as a POMDP

In this section, we consider a fully observed system with a continuous state space and construct an approximate Q learning algorithm by discretizing the state space and using a finite subset of the state space for the Q iterations. We assume that $\mathbb{X} \subset \mathbb{R}^d$ is compact, and thus we can choose a finite subset $\hat{\mathbb{X}} = \{x_1, \ldots, x_n\}$ such that

$$\max_{x\in\mathbb{X}}\min_{\hat{x}\in\hat{\mathbb{X}}}|x - \hat{x}| \le \alpha(1/n)^{1/d}$$

for some $\alpha > 0$. We use a nearest neighbor map $\rho : \mathbb{X} \to \hat{\mathbb{X}}$ to choose elements from the finite set $\hat{\mathbb{X}}$ such that at any time instance $t < \infty$, if the state is x_t, we use

$$\hat{x}_t = \rho(x_t) := \arg\min_{\hat{x}\in\hat{\mathbb{X}}}\|\hat{x} - x_t\|$$

for the Q learning algorithm. Note that with this map, we separate $\mathbb{X}$ into n subsets $\{B_1, \dots, B_n\}$ such that for $x_i \in \hat{\mathbb{X}}$ $B_i := \{x \in \mathbb{X} : \rho(x) = x_i\}$. Using the map ρ, we construct the following Q learning algorithm for any $(x, u) \in \mathbb{X} \times \mathbb{U}$

$$Q_{k+1}(\rho(x), u) = (1 - \alpha_k(\rho(x), u))Q_k(\rho(x), u) \tag{19}$$
$$+ \alpha_k(\rho(x), u) \left(C_k(\rho(x), u) + \beta \min_v Q_k(\rho(X_1), v) \right)$$

that is for any true value of the state, we use its representative state from the finite set $\hat{\mathbb{X}}$. To choose the control actions, we again use randomized memoryless policies with positive probability for every action.

We now argue that this approximate iteration can be seen as a special case of the POMDP iteration (11) by considering the discretization as a quantizer channel. If we consider the finite set $\hat{\mathbb{X}}$ as the observation space and define the observation channel as a quantizer such that $O(\hat{x}_i|x) = 1_{\{x \in B_i\}}$, then the algorithm in (19) is the same algorithm as in (11) with $N = 0$.

Thus, we can use the set $\hat{\mathbb{X}}$ to construct the Q learning algorithm. Using the quantizer channel and Theorem 2 (i) for $N = 0$, the algorithm converges to

$$Q^*(\hat{x}, u) = C^*(\hat{x}, u) + \beta \sum_{\hat{x}_1} P^*(\hat{x}_1|\hat{x}, u) \min_v Q^*(\hat{x}_1, v)$$

where, for $\hat{x} \in B$ and $\hat{x}_1 \in B_1$, if we define $\hat{\pi}^*(A) := \frac{\pi^*(A)}{\pi^*(B)}$ for all $A \subset B$ with π^* being the invariant measure, the cost and transitions are defined as

$$C^*(\hat{x}, u) = \int_B c(x, u)\hat{\pi}^*(dx)$$
$$P^*(\hat{x}_1|\hat{x}, u) = \int_B T(B_1|x, u)\hat{\pi}^*(dx).$$

This fixed point equation aligns with the approximate finite model constructed in [14, Chapter 4]. Thus, we arrive at the following result, using Theorem 2 and [14, Theorem 4.38]:

Theorem 4. *Under Assumption 1, if the transition kernel $T(\cdot|x_0, u_0)$ admits a density function f with respect to a measure ϕ such that $T(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$, $f(x_1, x_0, u_0) > 0$ for all x_1, x_0, u_0 and f is Lipschitz continuous in x_0 with constant α_T and if $c(x, u)$ is Lipschitz continuous in x with constant α_c, then the Q learning algorithm in (19) converges and for any policy γ^n that satisfies $Q^*(\hat{x}, \gamma^n(\hat{x})) = \min_u Q^*(\hat{x}, u)$, we have*

$$J_\beta(T, \gamma^n) - J^*_\beta(T) \leq K(\alpha, \alpha_c, \alpha_T, \beta)(1/n)^{1/d}$$

where the constant $K(\alpha, \alpha_c, \alpha_T, \beta)$ is defined in [14, Theorem 4.38].

VI. Concluding Remarks

We studied the convergence of an approximate Q learning algorithm for POMDPs that uses finite window history variables. We first established convergence and then provided the approximate belief-MDP model that the limit fixed equation corresponds to. Furthermore, we provided bounds on the difference between the performance of the policy that is learned via the proposed algorithm and the optimal policy. In particular, we obtained explicit error bounds between the resulting policy's performance and the optimal performance as a function of the memory length and a coefficient related to filter stability. We applied these results to fully observed MDPs with continuous state spaces and established near optimality of learned policies via quantization of the state space, where the quantization is viewed as a measurement channel leading to a POMDP and a history of unit window size was considered. In particular, we showed that Q-learning, with its convergence and near optimality properties, applies for continuous space MDPs when the state space is quantized.

References

[1] P. Billingsley. *Convergence of probability measures*. New York: Wiley, 2nd edition, 1999.

[2] R.L. Dobrushin. Central limit theorem for nonstationary Markov chains. i. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[3] E. A. Hansen. Solving pomdps by searching in policy space. *arXiv preprint arXiv:1301.7380*, 2013.

[4] T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

[5] T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in neural information processing systems*, pages 345–352, 1995.

[6] A. D. Kara and S. Yüksel. Convergence of finite memory q-learning for pomdps and near optimality of learned policies under filter stability. *arXiv preprint arXiv:2103.12158*, 2021.

[7] V. Krishnamurthy. *Partially observed Markov decision processes: from filtering to controlled sensing*. Cambridge University Press, 2016.

[8] L.-Ji. Lin and T. M. Mitchell. Memory approaches to reinforcement learning in non-Markovian domains. *Technical Report CMU-CS-92-138, Carnegie Mellon University, School of Computer Science*, 1992.

[9] W.S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28:47–66, 1991.

[10] A. McCallum. Reinforcement learning with selective perception and hidden state. *Doctoral dissertation, Department of Computer Science, University of Rochester*, 1997.

[11] C. McDonald and S. Yüksel. Exponential filter stability via Dobrushin's coefficient. *Electronic Communications in Probability*, 25, 2020.

[12] J. M. Porta, N. Vlassis, M. T. J. Spaan, and P. Poupart. Point-based value iteration for continuous pomdps. *Journal of Machine Learning Research*, 7(Nov):2329–2367, 2006.

[13] D. Rhenius. Incomplete information in Markovian decision models. *Ann. Statist.*, 2:1327–1334, 1974.

[14] N. Saldi, T. Linder, and S. Yüksel. *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*. Springer, Cham, 2018.

[15] N. Saldi, S. Yüksel, and T. Linder. On the asymptotic optimality of finite approximations to markov decision processes with borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017.

[16] N. Saldi, S. Yüksel, and T. Linder. Finite model approximations for partially observed markov decision processes with discounted cost. *IEEE Transactions on Automatic Control*, 65, 2020.

[17] S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier, 1994.

[18] J. Subramanian and A. Mahajan. Approximate information state for partially observed systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1629–1636, 2019.

[19] J. N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16:185–202, 1994.

[20] N. Vlassis and M. T. J. Spaan. Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, 24:195–220, 2005.

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

[22] A.A. Yushkevich. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. *Theory Prob. Appl.*, 21:153–158, 1976.

[23] R. Zhou and E.A. Hansen. An improved grid-based approximation algorithm for POMDPs. In *Int. J. Conf. Artificial Intelligence*, pages 707–714, Aug. 2001.