Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Convergence of Finite Memory Q Learning for POMDPs and Near Optimality of Learned Policies Under Filter Stability

Ali Devran Kara, Serdar Yüksel

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 48, No. 4, November 2023, pp. 2066–2093
ISSN 0364-765X (print), ISSN 1526-5471 (online)
https://pubsonline.informs.org/journal/moor

Convergence of Finite Memory Q Learning for POMDPs and Near Optimality of Learned Policies Under Filter Stability

Ali Devran Kara,[a,]* Serdar Yüksel[b]

[a] Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48103; [b] Department of Mathematics and Statistics, Queen's University, Kingston, Ontario K7L 3N6, Canada

*Corresponding author

Contact: alikara@umich.edu, https://orcid.org/0000-0001-8119-6620 (ADK); yuksel@mast.queensu.ca (SY)

Received: February 16, 2021
Revised: January 9, 2022
Accepted: October 11, 2022
Published Online in Articles in Advance: November 21, 2022

MSC2020 Subject Classification: Primary: 60J20, 60J05, 93E11

https://doi.org/10.1287/moor.2022.1331

Abstract. In this paper, for partially observed Markov decision problems (POMDPs), we provide the convergence of a Q learning algorithm for control policies using a finite history of past observations and control actions, and consequentially, we establish near optimality of such limit Q functions under explicit filter stability conditions. We present explicit error bounds relating the approximation error to the length of the finite history window. We establish the convergence of such Q learning iterations under mild ergodicity assumptions on the state process during the exploration phase. We further show that the limit fixed point equation gives an optimal solution for an approximate belief Markov decision problem (MDP). We then provide bounds on the performance of the policy obtained using the limit Q values compared with the performance of the optimal policy for the POMDP, in which we also present explicit conditions using recent results on filter stability in controlled POMDPs. Whereas there exist many experimental results, (i) the rigorous asymptotic convergence (to an approximate MDP value function) for such finite memory Q learning algorithms and (ii) the near optimality with an explicit rate of convergence (in the memory size) under filter stability are results that are new to the literature to our knowledge.

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

Keywords: reinforcement learning • partially observed MDP • nonlinear filtering

1. Introduction

Partially observed Markov decision problems (POMDPs) offer a practically rich and relevant and mathematically challenging model. Even in the most basic setup of finite state–action models, the analysis and computation of optimal solutions is complicated. The existence of optimal policies is, in general, established via converting or reducing the original partially observed stochastic control problem to a fully observed Markov decision problem (MDP) with probability measure valued (belief) states, leading to a belief MDP. However, computing an optimal policy for this fully observed model and so for the original POMDP using classical methods (such as dynamic programming, policy iteration, linear programming) is not simple even if the original system has finite state and action spaces because the state space of the fully observed (reduced) model is always uncountable. Furthermore, when the dynamics are not known, learning theoretic methods are not as comprehensively and conclusively studied as the fully observed counterpart for MDPs, mainly because of the technical subtleties as we discuss further.

1.1. On Approximation Methods

The problem of approximate optimality is significantly more challenging compared with the fully observed counterpart. Most of the studies in the literature are algorithmic and computational contributions. These include Porta et al. [27] and Zhou and Hansen [44], which develop computational algorithms, utilizing structural convexity/concavity properties of the value function under the discounted cost criterion. Vlassis and Spaan [36] provide an insightful algorithm that may be regarded as a quantization of the belief space; however, no rigorous convergence results are provided. Smith and Simmons [32] and Pineau et al. [26] also present quantization-based algorithms for the belief state, in which the state, measurement, and the action sets are finite.

For partially observed setups Saldi et al. [30], building on Saldi et al. [29], introduce a rigorous approximation analysis (and explicit methods for quantization of probability measures) after establishing weak continuity conditions on the transition kernel defining the belief MDP via the nonlinear filter (Feinberg et al. [5], Kara et al. [15]) and show that finite model approximations obtained through quantization are asymptotically optimal and the control policies obtained from the finite model can be applied to the actual system with asymptotically vanishing error as the number of quantization bins increases. Another rigorous set of studies is from Zhou et al. [42, 43] in which the authors provide an explicit quantization method for the set of probability measures containing the belief states, in which the state space is parametrically representable under strong density regularity conditions. The quantization is done through the approximations as measured by the Kullback–Leibler divergence (relative entropy) between probability density functions. Subramanian and Mahajan [33] present a notion of approximate information variable and study near optimality of policies that satisfy the approximate information state property.

We refer the reader to the survey papers by Lovejoy [19], White [38], and Hansen [8] and the recent book by Krishnamurthy [16] for further structural results as well as algorithmic and computational methods for approximating POMDPs. Notably, for POMDPs, Krishnamurthy [16] presents structural results on optimal policies under monotonicity conditions of the value function in the belief variable.

1.2. On Learning for POMDPs

Learning in POMDPs is challenging for the reasons discussed: if one attempts to learn optimal policies through empirical observations, then the analysis and convergence properties become significantly harder to obtain as the observations progress in a non-Markovian fashion and the belief state is uncountable. Jaakkola et al. [12] study a learning algorithm for POMDPs with average cost criteria in which a policy improvement method is proposed using random polices and the convergence of this method to local optima is given. McCallum [20] and Lin and Mitchell [18] propose the same approach as we use in this paper, and they use a finite memory of history to construct learning algorithms. They provide extensive experimental results; however, both lack a rigorous convergence or approximation result.

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

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

where y_k represents the observations and u_k represents the control actions. We can further improve this algorithm by using not only the most recent observation, but a finite window of past observations and control actions because we can infer information on the true state from the past data. Two main problems with this approach are that (i) first, the (Y_k, U_k) process is not a controlled Markov process (as only (X_k, U_k) is), and the cost realizations $C_k(y_k, u_k)$ depend on the observation process in a random and time-dependent fashion, and hence, the convergence of this approach does not follow directly from usual techniques (Jaakkola et al. [11], Tsitsiklis [34]), and (ii) second, even if the convergence is guaranteed, it is not immediate what the limit Q values are or whether they are meaningful at all. In particular, it is not known what MDP model gives rise to the limit Q values.

Singh et al. [31] study (1), that is, the Q learning algorithm for POMDPs, by ignoring the partial observability and constructing the algorithm using the most recent observation variable (for which the state, action, and measurement spaces were all assumed finite) and establish convergence of this algorithm under mild conditions (notably that the hidden state process is uniquely ergodic under the exploration policy, which is random and puts positive measure to all action variables). In our paper, we consider memory sizes of more than zero for the information variables and a continuous state space, and thus, the algorithm in Singh et al. [31] can be seen as a special case of our setup. Different from our work, however, Singh et al. [31] does not study what the limit of the iterations mean and, in particular, whether the limit equation corresponds to some MDP model. In this paper, we rigorously construct the approximate belief MDP that the limit equation satisfies, which gives an operational and practical conclusion regarding the analysis of the algorithm. Furthermore, we use different window sizes, which turns out to be crucial for the performance of the learned policy: using longer window sizes reveals the intimate connection between the approximate learning problem and the nonlinear controlled filter stability problem that we study in detail. This ultimately leads to near optimality of the N-window variation of (1) with an explicit approximation and robustness error bound as a function of N and a computable/boundable coefficient related to filter stability.

Another motivation for our study is the following: often one deals with problems in which not only the specification of an MDP is unknown, but whether the problem is an MDP in the first place may not be known. The simplest extension perhaps is that of a POMDP in which one is tempted to view the measurements as the state or finite window of measurement and actions as the state. A question that has not been resolved fully is whether a Q learning algorithm for such a setup indeed converges, and the next question is, if it does, to what does it converge. Our answer to the first question is positive under mild conditions, and the second question is, under filter stability conditions, that the convergence is to near optimality with an explicit error bound between the performance loss and the memory window size.

1.3. On Finite Memory Approximations and Relations with Controlled Filter Stability

In our paper, we see, perhaps not surprisingly, that filter stability is an essential ingredient for the learning algorithm to arrive at optimal or near optimal solutions. In other words, how fast the process forgets its initial prior distribution when updated with the information variables is a key aspect for the performance of the approximate Q values determined using most recent information variables. Unlike fully observed systems, the system (belief MDP) states cannot be visited infinitely often for POMDPs because there are uncountably many belief states and the measurements collected should somehow present approximate information on the belief states through conditions related to filter stability. We make this intuition precise in our paper. We also note that, in optimal control theory, it is a standard result that (time-invariant) output feedback control performs poorly compared with state-feedback and, in the absence of observability, this holds for all memory lengths.

We end the literature review section by mentioning particularly related studies on finite memory control for POMDPs. White and Scherer [39] is a particularly related work that studies approximation techniques for POMDPs using finite memory with finite state, action, and measurements. The POMDP is reduced to a belief MDP, and the worst and best case predictors prior to the N most recent information variables are considered to build an approximate belief MDP. The original value function is bounded using these approximate belief MDPs that use only finite memory, in which the finiteness of the state space is critically utilized. Furthermore, a loss bound is provided for a suboptimally constructed policy that only uses finite history, in which the bound depends on a specific ergodicity coefficient (which requires restrictive sample path contraction properties). In this paper, we consider more general signal spaces and more relaxed filter stability requirements and, in particular, establish explicit rates of convergence results. We also rigorously construct the finite belief MDP considering the approximate Q learning algorithm, whereas White and Scherer [39] only focus on the approximation aspect of POMDPs.

In Yu and Bertsekas [40], the authors study near optimality of finite window policies for average cost problems in which the state, action, and observation spaces are finite; under the condition that the liminf and limsup of the average cost are equal and independent of the initial state, the paper establishes the near optimality of (nonstationary) finite memory policies. Here, a concavity argument building on Feinberg [4] (which becomes consequential by the equality assumption) and the finiteness of the state space are crucial. The paper shows that, for any given $\epsilon > 0$, there exists an ϵ-optimal finite window policy. However, the authors do not provide a performance bound related to the length of the window, and in fact, the proof method builds on convex analysis.

In a recent paper (Kara and Yüksel [14]), we establish near optimality of finite window policies using a different approach by considering the belief MDP directly and quantizing the belief space with a nearest neighbor map (under a metric on probability measures that induces the weak convergence topology) that uses finite window information variables. In particular, the results in that paper did not establish the convergence of a Q learning algorithm and, strictly speaking, required the knowledge of the belief state to choose the nearest element from the finite set. As we see later, the approximate Q learning algorithm does not necessarily choose the nearest element from the finite set induced by the window information variables. Thus, in this paper, we explicitly only use the memory variables directly for the approximation. We also note that the approximation method presented in Kara and Yüksel [14] only works for a restricted values of the discount factor, which depends on the system components and the filter stability terms, whereas the method used in this paper does not put any restrictions on the discount factor. On the other hand, in Kara and Yüksel [14] one can relax filter stability to be under weak convergence; in our current paper, we consider filter stability under total variation.

Similar to Kara and Yüksel [14], our analysis here also makes explicit connections with filter stability; that is, how fast the controlled process forgets its initial distribution as it observes the information variables from the system. In the literature, there are various sets of assumptions to achieve filter stability. Two main approaches are:

- The transition kernel is sufficiently ergodic, forgetting the initial measure and, therefore, passing this insensitivity (to incorrect initializations) on to the filter process. This condition is often tailored toward control-free models.
- The measurement channel provides sufficient information about the underlying state, allowing the filter to track the true state process. This approach is typically based on martingale methods and, accordingly, does not often lead to rates of convergence for the filter stability problem but only asymptotic filter stability.

We use the recent results in the controlled filter stability literature presented in McDonald and Yüksel [23] for exponential filter stability and McDonald and Yüksel [22, 24] for asymptotic filter stability.

In a recent study (Golowich et al. [7]), a finite memory–based approximate planning method is studied for POMDPs, and the relation between the performance of the approximation and the filter stability is established similar to this paper and Kara and Yüksel [14]. To achieve filter stability, a restrictive rank condition is used for the observation channel, and a polynomial convergence rate is achieved as opposed to the general filter stability setup we consider here, which includes exponential or asymptotic filter stability conditions. The approach in Golowich et al. [7] to deal with the filter stability is specifically tailored toward finite state spaces, whereas we present results for possibly continuous state spaces, and our analysis in approximation is explicit for any filter stability error of the form given in L_t (see Equation (16)). The setup in Golowich et al. [7] can be viewed as a particular instance in which the state and action spaces are finite and the measurement channel has a restrictive invertibility condition. Furthermore, here, we also present a reinforcement learning algorithm using finite memory variables. We also emphasize that explicit filter stability conditions are provided in McDonald and Yüksel [23, theorem 3.3] for exponential filter stability and McDonald and Yüksel [24, theorem 3.6] for asymptotic filter stability (the latter, via the examples in McDonald and Yüksel [21, section 3], also includes a rank condition for finite models).

We highlight that one key contribution of the paper is the construction of an alternative belief MDP reduction introduced in Section 3, which provides a structure to the finite memory approximations. The alternative reduction technique leads to an explicit and rigorous error analysis by changing the topology and the construction of the state space of the reduced model with no restrictions on the discount parameter $\beta \in (0,1)$ unlike Kara and Yüksel [14].

1.4. Contributions

i. In Section 3.1, we provide an alternative belief MDP reduction method that is tailored toward finite memory approaches. In Section 3.2, we construct an approximate model using the alternative belief MDP reduction (see Figure 1). In particular, in Theorems 2 and 3, we establish bounds for the difference between the value functions of the original POMDP model and the approximate model and for the performance loss of the policy obtained using the approximate belief-MDP when it is used in the original model. We show that the policy obtained using the approximate model uses finite memory feedback variables to choose the control actions. Furthermore, Theorems 2 and 3 reveal the close connection between the finite memory approximation method and the controlled filter stability problem through a filter stability term L_t defined in (16).

ii. In Section 4, we present a Q learning algorithm that uses finite memory feedback variables. In Theorem 4, we show that the Q iterations constructed using finite history variables converge under mild ergodicity assumptions on the hidden state process, and the limit fixed point equation corresponds to the optimal solution for the approximate belief-MDP model introduced in Section 3.2.

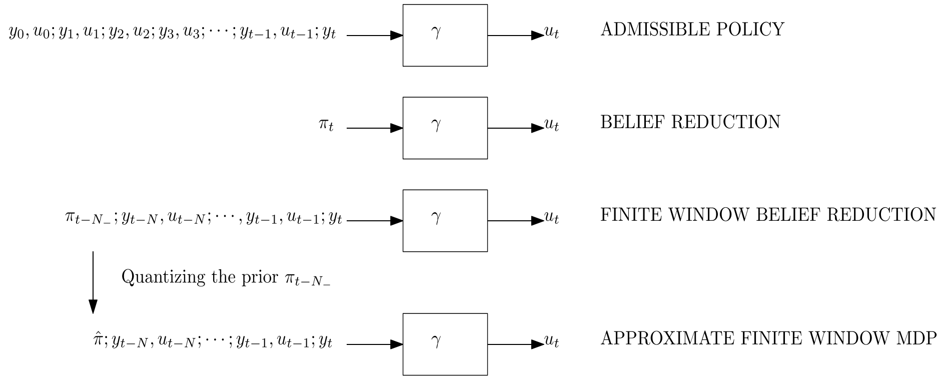
iii. We, finally, in Section 5, provide a particular result to guarantee exponential stability for controlled filter problems, which, in turn, implies that the error resulting from the finite memory approximation and learning methods decays to zero exponentially fast as the memory size increases under explicit filter stability conditions to be presented (Corollaries 2 and 3).

In Section 6, we provide numerical examples that verify both the Q learning convergence and near optimality results.

2. Partially Observed Markov Decision Processes and Belief MDP Reduction

Let $\mathbb{X} \subset \mathbb{R}^m$ denote a Borel set that is the state space of a partially observed controlled Markov process for some $m \in \mathbb{N}$. Here and throughout the paper, $\mathbb{Z}_+$ denotes the set of nonnegative integers, and $\mathbb{N}$ denotes the set of positive integers. Let $\mathbb{Y}$ be a finite set denoting the observation space of the model, and let the state be observed

Figure 1. Construction of the finite-window approximate MDP from the finite-window belief-MDP. The quantization of the finite window MDP model leads to the collapse of the first coordinate to a fixed measure.



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 the power set $P(\mathbb{Y})$ of $\mathbb{Y}$ for every $x \in \mathbb{X}$, and $O(A|\cdot): \mathbb{X} \to [0,1]$ is a Borel measurable function for every $A \in P(\mathbb{Y})$. A decision maker is located at the output of the channel O, and hence, it only sees the observations $\{Y_t, t \in \mathbb{Z}_+\}$ and chooses its actions from $\mathbb{U}$, the action space that 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

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+, \tag{2}$$

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 (2) and the following relationships:

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

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

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

$B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}), t \in \mathbb{N}$, where $\mathcal{T}$ is the transition kernel of the model that is a stochastic kernel from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{X}$. Note that, although $\mathbb{Y}$ is finite, we here use the integral sign instead of the summation sign for notational convenience by letting the measure to be sum of Dirac-delta measures (and, as we discuss later in the paper, our analysis also holds for continuous measurement spaces). We let the objective of the agent (decision maker) be the minimization of the infinite horizon discounted cost,

$$J_\beta(\mu, \mathcal{T}, \gamma) = E_\mu^{\mathcal{T}, \gamma}\left[\sum_{t=0}^{\infty} \beta^t c(X_t, U_t)\right] \tag{3}$$

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 $E_\mu^{\mathcal{T},\gamma}$ denotes the expectation with initial state probability measure μ and transition kernel $\mathcal{T}$ 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 the optimal cost for the discounted infinite horizon setup as a function of the priors and the transition kernels as

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

For the analysis of partially observed MDPs, a common approach is to reformulate the problem as a fully observed MDP, in which 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.

2.1. Reduction to Fully Observed Models Using Belief States

2.1.1. Convergence Notions for Probability Measures. For the analysis of the technical results, we use different notions of convergence for sequences of probability measures.

Two important notions of convergence for sequences of probability measures are weak convergence and convergence under total variation. For a complete, separable, and metric space $\mathbb{X}$, for a sequence $\{\mu_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathbb{X})$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ weakly if $\int_{\mathbb{X}} c(x)\mu_n(dx) \to \int_{\mathbb{X}} c(x)\mu(dx)$ for every continuous and bounded $c: \mathbb{X} \to \mathbb{R}$. One important property of weak convergence is that the space of probability measures on a complete, separable, metric (Polish) space endowed with the topology of weak convergence is itself complete, separable, and metric (Parthasarathy [25]). One such metric is the bounded Lipschitz metric (Villani [35]), which is defined for $\mu, \nu \in \mathcal{P}(\mathbb{X})$ as

$$\rho_{BL}(\mu, \nu) := \sup_{\|f\|_{BL} \le 1} \left| \int f d\mu - \int f d\nu \right|, \tag{4}$$

where

$$\|f\|_{BL} := \|f\|_\infty + \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}$$

and $\|f\|_\infty = \sup_{x \in \mathbb{X}} |f(x)|$.

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

$$\|\mu - \nu\|_{TV} = 2 \sup_{B \in \mathcal{B}(\mathbb{X})} |\mu(B) - \nu(B)| = \sup_{f:\|f\|_\infty \leq 1} \left| \int f(x)\mu(\mathrm{d}x) - \int f(x)\nu(\mathrm{d}x) \right|,$$

where the supremum is taken over all measurable real f such that $\|f\|_\infty = \sup_{x \in \mathbb{X}} |f(x)| \leq 1$. A sequence μ_n is said to converge in total variation to $\mu \in \mathcal{P}(\mathbb{X})$ if $\|\mu_n - \mu\|_{TV} \to 0$.

2.1.2. Construction of the Belief MDP and Some Regularity Properties. It is by now a standard result that, for optimality analysis, any POMDP can be reduced to a completely observable Markov decision process (Rhenius [28], Yushkevich [41]), whose states are the posterior state distributions or beliefs of the observer or the filter process; that is, the state at time t is

$$z_t := Pr\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\} \in \mathcal{P}(\mathbb{X}). \tag{5}$$

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 (Billingsley [1]). Under this topology, $\mathcal{Z}$ is a standard Borel space (Parthasarathy [25]). Then, the transition probability η of the filter process can be constructed as follows (see also Hernández-Lerma [9]). If we define the measurable function

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

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

$$\eta(\cdot|z,u) = \int_{\mathbb{Y}} 1_{\{F(z,u,y) \in \cdot\}} P(dy|z,u). \tag{6}$$

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

$$\tilde{c}(z,u) := \int_{\mathbb{X}} c(x,u) z(dx), \tag{7}$$

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

For the filter process, the information variables are defined as

$$\tilde{I}_t = \{Z_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \qquad \tilde{I}_0 = \{Z_0\}.$$

It is well-known that an optimal control policy of the original POMDP can use the belief Z_t as a sufficient statistic for optimal policies (see Rhenius [28], Yushkevich [41]), provided they exist. More precisely, the filter process is equivalent to the original POMDP in the sense that, for any optimal policy for the filter process, one can construct a policy for the original POMDP that is optimal. On existence, we note the following.

By the recent results in Feinberg et al. [6] and Kara et al. [15], the transition model of the belief MDP can be shown to satisfy weak continuity conditions on the belief state and action variables, and accordingly, we have that the measurable selection conditions (Hernández-Lerma and Lasserre [10, chapter 3]) apply. Notably, we state the following.

Assumption 1.

i. *The transition probability $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u), that is, for any $(x_n,u_n) \to (x,u)$, $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly.*

ii. *The observation channel $O(\cdot|x)$ is continuous in total variation, that is, for any $x_n \to x$, $O(\cdot|x_n) \to O(\cdot|x)$ in total variation.*

Assumption 2. *The transition probability $\mathcal{T}(\cdot|x,u)$ is continuous in total variation in (x, u), that is, for any $(x_n,u_n) \to (x,u)$, $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ in total variation.*

Theorem 1.

i. *(Feinberg et al. [6]) Under Assumption 1, the transition probability $\eta(\cdot|z,u)$ of the filter process is weakly continuous in (z,u).*

ii. *(Kara et al. [15]) Under Assumption 2, the transition probability $\eta(\cdot|z,u)$ of the filter process is weakly continuous in (z,u).*

Under these weak continuity conditions and appropriate conditions on the stage-wise cost function (e.g., bounded and continuous c with Assumption 1 or bounded c with Assumption 1), the measurable selection conditions (Hernández-Lerma and Lasserre [10, chapter 3]) apply and a solution to the discounted cost optimality equation exists, and accordingly, an optimal control policy exists.

This policy is stationary (in the belief state). If we denote this optimal belief policy by $\phi : \mathcal{P}(\mathbb{X}) \to \mathbb{U}$, we can then find a policy γ on the partially observed setup such that

$$\gamma(y_{[0,n]}) := \phi(P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n]} = y_{[0,n]})) = \phi(\pi_n^{\mu,\gamma}).$$

Hence, the policy γ can be used as an optimal policy for the partially observed MDP.

Even though, the belief MDP approach provides a strong tool for the analysis of POMDPs, it is usually too complicated computationally. The belief space $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ is always uncountable even when $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{U}$ are finite. Furthermore, the information variables I_t grows with time, and the computation of the belief state $Pr(X_t \in \cdot | I_t)$ can become intractable. Therefore, approximation of the belief MDP is usually needed. In the following section, we provide an alternative fully observed MDP approach and present approximation results that only make use of a finite history of the information variables.

3. An Alternative Finite Window Belief MDP Reduction and Its Approximation

3.1. An Alternative Finite Window Belief MDP Reduction

In this section, we construct an alternative fully observed MDP reduction with the condition that the controller has observed at least N information variables, 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 as a probability measure valued variable. In what follows, we sometimes consider the case with $n = 1$ for some of the proofs to make the presentation less complicated. The general case follows from identical arguments.

For the remainder of the paper, to emphasize the prior distribution of the starting state variable, we use the following notation for conditional probabilities on state and observation variables.

Definition 1. Assume that the initial state X_0 has a prior distribution $\mu \in \mathcal{P}(\mathbb{X})$. Then, for the conditional distribution of X_t given the past observation and action variables $\{y_t, \ldots, y_0\}$, $\{u_{t-1}, \ldots, u_0\}$, we define

$$P^{\mu}(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0) := Pr(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0).$$

Given that x_0 has a prior distribution $\mu \in \mathcal{P}(\mathbb{X})$, we define the following for the conditional distribution of Y_t given the past observation and action variables $\{y_{t-1}, \ldots, y_0\}$, $\{u_{t-1}, \ldots, u_0\}$:

$$P^{\mu}(Y_t \in \cdot | y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0) := Pr(Y_t \in \cdot | y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0).$$

Consider the following state variable at time t:

$$\hat{z}_t = (\pi_{t-N}^-, I_t^N), \tag{8}$$

where, for $N \geq 1$,

$$\pi_{t-N}^- = Pr(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0),$$
$$I_t^N = \{y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\},$$

and $I_t^N = y_t$ for $n = 0$ with μ being the prior probability measure on X_0. 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 in which we consider the weak convergence topology on the $\mathcal{P}(\mathbb{X})$ coordinate and the usual (coordinate) 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 (5). Consider the map $\psi : \hat{\mathcal{Z}} \to \mathcal{P}(\mathbb{X})$, for some $\hat{z}_t = (\pi_{t-N}^-, I_t^N)$:

$$\psi(\hat{z}_t) = \psi(\pi_{t-N}^-, I_t^N) = P^{\pi_{t-N}^-}(X_t \in \cdot | I_t^N) = P^{\pi_{t-N}^-}(X_t \in \cdot | y_t, \ldots, y_{t-N}, u_{t-N-1}, \ldots, u_{t-N-1})$$
$$= P^{\mu}(X_t \in \cdot | y_t, \ldots, y_0, u_{t-1}, \ldots, u_0) = z_t$$

such that the map ψ acts as a Bayesian update of π_{t-N}^- using I_t^N. 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 (7)) such that

$$\hat{c}(\hat{z}_t, u_t) = \hat{c}(\pi_{t-N}^-, I_t^N, u_t) = \tilde{c}(\psi(\pi_{t-N}^-, I_t^N), u_t)$$
$$= \int_{\mathbb{X}} c(x_t, u_t) P^{\pi_{t-N}^-}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}). \tag{9}$$

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

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

we write

$$Pr(\hat{z}_{t+1} \in A | \hat{z}_t, \ldots, \hat{z}_0, u_t, \ldots, u_0)$$
$$= Pr(\pi_{t-N+1}^- \in B, \hat{y}_{t+1}, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1} | \pi_{t-N}^-, \ldots, \pi_0^-, y_t, \ldots, y_0, u_t, \ldots, u_0)$$
$$= \mathbb{1}_{\{(y_t, \ldots, y_{t-N+1}, u_t, \ldots, u_{t-N+1}) = (\hat{y}_t, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1})\}}$$
$$\times \mathbb{1}_{\{G(\pi_{t-N}^-, y_{t-N}, u_{t-N}) \in B\}} P^{\pi_{t-N}^-}(\hat{y}_{t+1} | y_t, \ldots, y_{t-N}, u_t, \ldots, u_{t-N})$$
$$= Pr(\pi_{t-N+1}^- \in B, \hat{y}_{t+1}, \ldots, \hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1} | \pi_{t-N}^-, y_t, \ldots, y_{t-N}, u_t, \ldots, u_{t-N})$$
$$= Pr(\hat{z}_{t+1} \in A | \hat{z}_t, u_t)$$
$$=: \int_A \hat{\eta}(d\hat{z}_{t+1} | \hat{z}_t, u_t),$$

where the map G is defined as

$$G(\pi_{t-N}^-, y_{t-N}, u_{t-N}) = G(P^{\mu}(X_{t-N} \in \cdot | y_{t-N-1}, \ldots, y_0, u_{t-N-1}, \ldots, u_0), y_{t-N}, u_{t-N})$$
$$= P^{\mu}(X_{t-N+1} \in \cdot | y_{t-N}, \ldots, y_0, u_{t-N}, \ldots, u_0).$$

Hence, $\hat{\eta}$ defines a controlled transition model for the new states $\hat{z}_{t+1} \in \hat{\mathcal{Z}}$. Then, we have a proper fully observed MDP with the cost function $\hat{c}$, transition kernel $\hat{\eta}$, and 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 $\hat{\phi} := \phi \circ \psi$ such that

$$\hat{\phi}(\hat{z}) = \phi(\psi(z)).$$

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. Note that, with the alternative approach, the state $\hat{z}$ can only be written if we have at least N information variables. Therefore, given that the decision maker observed at least N information variables, we write the following fixed-point equation:

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

We can rewrite this fixed-point equation in a different form; for notational ease, assume $n = 1$. If $\hat{z}$ has the form (π_0^-, y_1, y_0, u_0), then we can rewrite

$$J_\beta^*(\pi_0^-, y_1, y_0, u_0) = \min_{u_1 \in \mathbb{U}} \left(\hat{c}(\pi_0^-, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in \mathbb{Y}} J_\beta^*(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1) P^{\pi_0^-}(y_2 | y_1, y_0, u_1, u_0) \right). \quad (10)$$

This representation plays an important role in the analysis of the problem. Note that the policy $\hat{\phi}^* = \phi^* \circ \psi$ satisfies this fixed-point equation.

The following fixed-point equation can also be defined for any policy $\hat{\phi} : \hat{\mathcal{Z}} \to \mathbb{U}$:

$$J_\beta(\hat{z}, \hat{\phi}) = \hat{c}(\hat{z}, \hat{\phi}(\hat{z})) + \beta \int J_\beta(\hat{z}_1, \hat{\phi}) \hat{\eta}(d\hat{z}_1 | \hat{z}, \hat{\phi}(\hat{z})),$$

where $J_\beta(\hat{z}, \hat{\phi})$ denotes the value function under the policy $\hat{\phi}$ for the initial point $\hat{z}$.

3.2. Approximation of the Finite Window Belief MDP

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

$$\hat{\mathcal{Z}}_{\pi^*}^N = \left\{ (\pi^*, y_{[0,N]}, u_{[0,N-1]}) : y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N \right\} \quad (11)$$

such that the state at time t is $\hat{z}_t^N = (\pi^*, I_t^N)$. Compared with the state $\hat{z}_t = (\pi_{t-N}^-, I_t^N)$ defined in (8), this approximate model uses π^* as the predictor no matter what the real predictor at time $t - N$ is.

The cost function is defined in the usual manner so that

$$\hat{c}(\hat{z}_t^N, u_t) = \hat{c}(\pi^*, I_t^N, u_t) = \tilde{c}(\phi(\pi^*, I_t^N), u_t) = \int_{\mathbb{X}} c(x_t, u_t) P^{\pi^*}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}).$$

We define the controlled transition model by

$$\hat{\eta}^N(\hat{z}_{t+1}^N | \hat{z}_t^N, u_t) = \hat{\eta}^N(\pi^*, I_{t+1}^N | \pi^*, I_t^N, u_t) := \hat{\eta}\left(\mathcal{P}(\mathbb{X}), I_{t+1}^N | \pi^*, I_t^N, u_t\right). \quad (12)$$

For simplicity, if we assume $n = 1$, then the transitions can be rewritten for some $I_{t+1}^N = (\hat{y}_{t+1}, \hat{y}_t, \hat{u}_t)$ and $I_t^N = (y_t, y_{t-1}, u_{t-1})$:

$$\hat{\eta}^N(\pi^*, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \pi^*, y_t, y_{t-1}, u_{t-1}, u_t) = \hat{\eta}(\mathcal{P}(\mathbb{X}), \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \pi^*, y_t, y_{t-1}, u_{t-1}, u_t) = \mathbb{1}_{\{y_t = \hat{y}_t, u_t = \hat{u}_t\}} P^{\pi^*}(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}). \quad (13)$$

Denoting the optimal value function for the approximate model by J_β^N, we can write the following fixed-point equation:

$$J_\beta^N(\hat{z}^N) = \min_{u \in \mathbb{U}} \left(\hat{c}(\hat{z}^N, u) + \beta \sum_{\hat{z}_1^N \in \hat{\mathcal{Z}}_{\pi^*}^N} J_\beta^N(\hat{z}_1^N) \hat{\eta}^N(\hat{z}_1^N | \hat{z}^N, u) \right). \quad (14)$$

By assuming $n = 1$ again, we can rewrite the fixed-point equation for some $\hat{z}_0^N = (\pi^*, y_1, y_0, u_0)$ as

$$J_\beta^N(\pi^*, y_1, y_0, u_0) = \min_{u_1 \in \mathbb{U}} \left(\hat{c}(\pi^*, y_1, y_0, u_0, u_1) + \beta \sum_{y_2 \in \mathbb{Y}} J_\beta^N(\pi^*, y_2, y_1, u_1) P^{\pi^*}(y_2 | y_1, y_0, u_1, u_0) \right). \quad (15)$$

Because everything is finite in this setup, we can assume the existence of an optimal policy ϕ^N that satisfies this fixed-point equation. Note that both J_β^N and ϕ^N are defined on the finite set $\hat{\mathcal{Z}}_{\pi^*}^N$. However, we can simply extend them to the set $\hat{\mathcal{Z}}$ by defining

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

for any $\hat{z} = (\pi, y_1, y_0, u_0) \in \hat{\mathcal{Z}}$.

We 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. For $n=1$, for example, Equation (15) is significant for the Q value iteration.

Another point to note is that the policy ϕ^N only uses the most recent N information variables to choose the control actions.

In what follows, we investigate the following differences:

$$|\tilde{J}_\beta^N(\hat{z}) - J_\beta^*(\hat{z})|,$$

$$J_\beta(\hat{z}, \tilde{\phi}^N) - J_\beta^*(\hat{z}).$$

The first one is the difference between the optimal value function of the original model and that for the approximate model. The second term is the performance loss resulting from the policy calculated for the approximate model being applied to the true model.

Remark 1. We note that, in Saldi et al. [29], the authors study approximation methods for MDPs with continuous state spaces by quantizing the state space and constructing a finite state MDP. In this section, we also construct a finite state space, $\hat{\mathcal{Z}}_{\pi^*}^N$, by quantizing $\hat{\mathcal{Z}}$. In Saldi et al. [29], continuity properties of the transition kernel $\hat{\eta}$ are used. However, establishing regularity properties for $\hat{\eta}$ is challenging. Therefore, we follow a different approach, and instead of working directly with $\hat{\eta}$, we analyze the components of a partially observed MDP for the following approximation results. We note that our quantization method is tailored toward filter stability and corresponds to a uniform quantization when we endow the finite window belief MDP space $\hat{\mathcal{Z}} = \mathcal{P}(\mathbb{X}) \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$ with the product topology of the weak convergence topology on $\mathcal{P}(\mathbb{X})$ and the usual (coordinate) topologies on $\mathbb{Y}$ and $\mathbb{U}$. We also note that our approach here then naturally applies to continuous (such as finite dimensional real valued) but compact space valued measurement and action spaces as well as a uniform quantization can be applied for all finite window belief MDP realizations.

3.2.1. Difference in the Value Functions in Terms of a Uniform Filter Stability Error.

In this section, we study the difference $|\tilde{J}_\beta^N(\hat{z}) - J_\beta^*(\hat{z})|$.

Before the result, we introduce some notation. We first define the measurable policies with respect to the new state space $\hat{\mathcal{Z}} = \mathcal{P}(\mathbb{X}) \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$ by $\hat{\Gamma}$. That is, a policy $\hat{\gamma} \in \hat{\Gamma}$ is a sequence of control functions $\{\hat{\gamma}_t, t \in \mathbb{Z}_+\}$ such that $\hat{\gamma}_t$ is measurable with respect to the σ-algebra generated by the information variables $\{\hat{z}_0, \ldots, \hat{z}_t\}$.

We now define the following bounding term:

$$L_t^N := \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}} \left[\| P^{\pi_t^-}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\pi^*}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \right], \tag{16}$$

which is the expected bound on the total variation distance between the posterior distributions of X_{t+N} conditioned on the same observation and control action variables $Y_{[t,t+N]}, U_{[t,t+N-1]}$ when the prior distributions of X_t are given by π_t^- and π^*. The expectation is with respect to the random 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^-. This constant represents the bound on the distance of two processes with different starting points when they are updated with the same observation and action processes under same policy. This is related to the filter stability problem, which is discussed in Section 5.

For the remainder of the paper, we drop the N dependence and denote the term by L_t.

Theorem 2. *For $\hat{z}_0 = (\pi_0^-, I_0^N)$, if a policy $\hat{\gamma}$ acts on the first N step of the process that produces I_0^N, we then have*

$$E_{\pi_0^-}^{\hat{\gamma}}[|\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0)| \,|\, I_0^N] \le \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t,$$

where L_t is defined as in (16).

Proof. The proof can be found in Appendix B. □

3.2.2. Performance Loss Resulting from Approximate Policy Being Applied to the True System in Terms of the Filter Stability Error.

We now study the difference $J_\beta(\hat{z}, \tilde{\phi}^N) - J_\beta^*(\hat{z})$, where $\tilde{\phi}^N$ is the optimal policy for the approximate model extended to the full space $\hat{\mathcal{Z}}$.

Theorem 3. *For $\hat{z}_0 = (\pi_0^-, I_0^N)$, with a policy $\hat{\gamma}$ acting on the first N steps,*

$$E_{\pi_0^-}^{\hat{\gamma}}[\,|J_\beta(\hat{z}_0, \tilde{\phi}^N) - J_\beta^*(\hat{z}_0)|\,|I_0^N] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t.$$

Proof. The proof can be found in Appendix C. □

Remark 2. In Kara and Yüksel [14], we construct a finite state approximate belief MDP using the state space $\hat{Z}_{\pi^*}^N$ defined in (11). However, different from the approach we use in this paper, to determine the approximate states, we used a nearest neighbor map to choose the closest element from the set $\hat{Z}_{\pi^*}^N$ to the $\mathcal{P}(\mathbb{X})$-valued belief state $z_t :=\Pr\{X_t \in \cdot | Y_0, \ldots, Y_t, U_0, \ldots, U_{t-1}\}$ under the bounded Lipschitz (BL) metric. We recall that the bounded Lipschitz metric, ρ_{BL}, for some $\mu, \nu \in \mathcal{P}(\mathbb{X})$ is given in (4). To find the closest element from $\hat{Z}_{\pi^*}^N$, one needs to know the belief state realization z_t, and to calculate/update the belief state, the system dynamics need to be known. However, as we see later, the Q learning algorithm presented here, using only the finite window information variables I_t^N, converges to the optimality equation of an approximate belief MDP that maps the belief state to an element from $\hat{Z}_{\pi^*}^N$ with matching finite window information rather than the closest element under the bounded Lipschitz metric. Hence, the alternative belief MDP construction and the approximation setup we present in this section serves better to analyze the approximate Q learning algorithm, which strictly uses the finite window memory variables. In other words, one does not need to calculate the belief state, but only needs to keep track of the information variables I_t^N for the approximation method introduced in this section. In particular, the state (π_{t-N}^-, I_t^N) is always mapped/quantized to (π^*, I_t^N), which can be done without the knowledge or computation of π_{t-N}^- as long as we have I_t^N available. Furthermore, in Kara and Yüksel [14], because we directly work with the topology and the metrics on the space of probability measures, the distinction between different realizations of history variables might be lost as we only care about the resulting posterior distribution on the hidden state variable; for example, different realizations of history variables may produce the same posterior distribution. However, for the finite memory Q learning iterations, it is key to be able to differentiate between the different realizations of finite memory feedback variables. Hence, in this paper, we put a different topology on $\hat{Z}_{\pi^*}^N$ by separating the finite memory variables rather than directly working with probability measures. This approach helps us to distinguish between different finite memory realizations.

On the other hand, one advantage of the approximation scheme used in Kara and Yüksel [14] is that, because of the nearest neighborhood map, one naturally arrives at a smaller approximation error. Furthermore, because of the continuity properties of the nearest neighbor map under the BL metric, one is able to work with the weak convergence topology; as such, we get an upper bound in terms of the BL metric, ρ_{BL}, such that the bounding term is

$$\rho_{BL}\left(P^{\pi}\left(X_t \in \cdot | y_{[t,t-N]}, u_{[t-1,t-N]}\right), P^{\pi^*}\left(X_t \in \cdot | y_{[t,t-N]}, u_{[t-1,t-N]}\right)\right),$$

which is always dominated by the total variation metric that we use in this section.

The different formulations and the approximation approach are summarized in Figure 1.

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

Assume that we start keeping track of the last $N+1$ observations and the last N control action variables after at least $N+1$ time steps. That is, at time t, we keep track of the information variables

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

We construct the Q value iteration using these information variables. In what follows, we drop the N dependence on I_t^N, and sometimes we use $n = 1$ for simplicity of the 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 $u \in \mathbb{U}$,

$$Q_{t+1}(I, u) = (1 - \alpha_t(I, u))Q_t(I, u) + \alpha_t(I, u)\left(C_t(I, u) + \beta \min_v Q_t(I_1^t, v)\right), \tag{17}$$

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

To choose the control actions, we use polices that choose the control actions randomly and independent of everything else such that, at time t,

$$u_t = u_i, \text{ w.p } \sigma_{\mathrm{i}}$$

for any $u_i \in \mathbb{U}$ with $\sigma_i > 0$ for all i.

We note that, for the convergence of the learning algorithm, it is sufficient for the hidden state process to converge to its invariant distribution under the exploration policy. Hence, any policy that leads the hidden state process to its invariant measure and visits every action with positive probability can be used for the exploration. For example, the control action can also be chosen to be a function of the most recent measurement and randomized (as long as all actions have positive probability of being selected for every measurement realization); this, again, leads to a uniquely ergodic hidden state process under our assumptions.

The algorithm differs from the usual Q value iteration:

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

In other words, the controlled transitions are time-dependent; that is, if we assume $n = 1$, then for some $I = (y_t, y_{t-1}, u_{t-1})$ and $u = u_t$,

$$Pr(I_1^t = (y'_{t+1}, y'_t, u'_t) | z = (y_t, y_{t-1}, u_{t-1}), u_t) = \mathbb{1}_{\{y_t = y'_t, u_t = u'_t\}} Pr(y_{t+1} | y_t, y_{t-1}, u_t, u_{t-1})$$

is not stationary and might change at every time step t because $Pr(y_{t+1} | y_t, y_{t-1}, u_t, u_{t-1})$ depends on the marginal distribution of x_{t-1} (x_{t-N} in the general case).

ii. Here, we only observe the cost realizations of the underlying state process $\{x_t\}_t$ and the control actions. For example, if we assume that $n = 1$, then the cost we observe is $c(x_t, u_t)$. However, $c(x_t, u_t)$ depends on (I, u) pair randomly and in a time dependent way so that, for some $I = (y_t, y_{t-1}, u_{t-1})$ and $u = u_t$,

$$C_t(I, u) = c(x_t, u_t) \in B, \quad \text{w.p. } Pr(X_t \in \{x : c(x, u_t) \in B\} | y_t, y_{t-1}, u_{t-1}),$$

where $Pr(dx_t | y_t, y_{t-1}, u_{t-1})$ can be seen as some pseudo-belief on the underlying state variable given $I = (y_t, y_{t-1}, u_{t-1})$, the most recent $n = 1$ information variables. In other words, $Pr(dx_t | y_t, y_{t-1}, u_{t-1})$ is the Bayesian update of π_{t-1}, the marginal distribution of the true state x_{t-1} at the time step $t-1$, using $I = (y_t, y_{t-1}, u_{t-1})$, and thus, it is time-dependent.

We observe that, if one assumes that the hidden state process $\{x_t\}_t$ is positive Harris recurrent or at least admits a unique invariant probability measure π^* under a stationary exploration policy γ, then the average of approximate state transitions gets closer to

$$P^*(I_{t+1} | I_t, u_t) := \hat{\eta}^N((\pi^*, I_{t+1}) | (\pi^*, I_t), u_t) \tag{18}$$

with $\hat{\eta}^N$ defined as in (12) and (13). In particular, if we assume $n = 1$, then we write

$$P^*(I_{t+1} = (y'_{t+1}, y'_t, u'_t) | I_t = (y_t, y_{t-1}, u_{t-1}), u_t) = \mathbb{1}_{\{y'_t = y_t, u'_t = u_t\}} P^{\pi^*}(y_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}), \tag{19}$$

where $P^{\pi^*}(y_{t+1} | y_t, y_{t-1}, u_t, u_{t-1})$ denotes the distribution of y_{t+1} when the marginal distribution on x_{t-1} is given by the invariant measure π^*.

We also have that the sample path averages of the random cost realizations get close to

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

where $P^*(x | I)$ is the Bayesian update of π^* using I, and $\hat{c}(\pi^*, I, u)$ is defined as in (9). If we assume $n = 1$, we can write for some $I = (y_1, y_0, u_0)$ and $u = u_1$,

$$C^*(y_1, y_0, u_0, u_1) = \hat{c}(\pi^*, (y_1, y_0, u_0), u_1) = \int_{\mathbb{X}} c(x_1, u_1) P^{\pi^*}(dx_1 | y_1, y_0, u_0). \tag{20}$$

Now, consider the following fixed-point equation:

$$Q^*(I, u) = C^*(I, u) + \beta \sum_{I'} P^*(I' | I, u) \min_v Q^*(I', v), \tag{21}$$

where P^* is defined in (18) and C^* is defined in (20).

The existence of such fixed point follows from the usual contraction arguments. The same fixed equation can also be written, for $n=1$ and for $I=(y_1,y_0,u_0)$ and $u=u_1$,

$$Q^*((y_1,y_0,u_0),u_1)=C^*((y_1,y_0,u_0),u_1)+\beta\sum_{y_2\in\mathbb{Y}}P^{\pi^*}(y_2|y_1,y_0,u_1,u_0)\min_{v\in\mathbb{U}}Q^*((y_2,y_1,u_1),v). \tag{22}$$

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

$$V^*(I):=\min_{v\in\mathbb{U}}Q^*(I,v), \tag{23}$$

$$V_t(I):=\min_{v\in\mathbb{U}}Q_t(I,v). \tag{24}$$

We note that the stationary distribution π^* does not have to be calculated by the decision maker. The Q value iterations given in (17) only use the finite memory variables I, and π^* is not used in the iterations. We show that the algorithm naturally converges to (21) if the state process is positive Harris recurrent or at least admits a unique invariant probability measure π^* under a stationary exploration policy γ, where π^* is the stationary distribution of the hidden state process x_t under the exploration policy. The performance loss depends on the stationary distribution π^* that is learned via the exploration policy; however, we establish further upper bounds that are uniform over such π^*, which decrease exponentially with the window size N (see Theorem 5, Equation (31), and Corollaries 2 and 3).

Assumption 3.

1. *We set $\alpha_t(I,u)=0$ 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 $I_k=I,u_k=u$, if k is the instant of the kth visit to (I, u) as this is crucial in the averaging of the Markov chain dynamics (see Remark 3).

2. *Under every stationary {memoryless or finite memory exploration} policy, say γ, the true state process, $\{X_t\}_t$, 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 4. *Under Assumption 3,*

i. *The algorithm given in (17) converges almost surely to Q^*, which satisfies (21).*

ii. *For any policy γ^N that satisfies $Q^*(I,\gamma^N(I))=\min_u Q^*(I,u)$, if we assume that the controller starts using γ^N at time $t=N$ (after observing at least N information variables), then denoting the prior distribution of X_N by π_N^- conditioned on the first N step information variables, we have*

$$E[J_\beta(\pi_N^-,\mathcal{T},\gamma^N)-J_\beta^*(\pi_N^-,\mathcal{T})|I_0^N]\le\frac{2\|c\|_\infty}{(1-\beta)}\sum_{t=0}^{\infty}\beta^tL_t,$$

where L_t is defined in (16) such that

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

and π^* is the invariant measure on x_t under the exploration policy γ.

Proof. For the proof of *i*, that is, for the convergence of Q learning, we separate the iterations into subiterations that are linear (as in Jaakkola et al. [11], where this superposition principle of linear systems theory is utilized in showing the convergence of standard Q learning algorithm). For the first part of the separated iterations, we use the fact that the dynamic programming equation is a contraction to prove its convergence, which is similar to the traditional Q learning algorithms. For the remaining part of the iteration, we analyze the asymptotic behavior of I_1^t, in which we distinguish our analysis from the traditional Q learning algorithms: for the usual Q iterations, one needs to study X_1 that is the consecutive state following some (x,u) pair, and we have that $X_1\sim\mathcal{T}(\cdot|x,u)$. Thus, it is distributed independently and identically given (x,u), which allows one to use Robbins–Monro type algorithms to show the convergence. However, distributions of I_1^t s are time-dependent and not controlled Markovian. To study the asymptotic behavior of I_1^t, we construct a different pair process that is Markov, and we use ergodicity properties of Markov chains.

We first prove that the process Q_t, determined by the algorithm in (17), converges almost surely to Q^*. We define

$$\Delta_t(I,u) := Q_t(I,u) - Q^*(I,u)$$
$$F_t(I,u) := C_t(I,u) + \beta V_t(I_1^t) - Q^*(I,u)$$
$$\hat{F}_t(I,u) := C^*(I,u) + \beta \sum_{I_1} V_t(I_1) P^*(I_1 | I,u) - Q^*(I,u),$$

where (V_t is defined in (23)).

Then, we can write the following iteration:

$$\Delta_{t+1}(I,u) = (1-\alpha_t(I,u))\Delta_t(I,u) + \alpha_t(I,u) F_t(I,u).$$

Now, we write $\Delta_t = \delta_t + w_t$ such that

$$\delta_{t+1}(I,u) = (1-\alpha_t(I,u))\delta_t(I,u) + \alpha_t(I,u)\hat{F}_t(I,u)$$
$$w_{t+1}(I,u) = (1-\alpha_t(I,u))w_t(I,u) + \alpha_t(I,u) r_t(I,u),$$

where $r_t := F_t - \hat{F}_t = \beta V_t(I_1^t) - \beta \sum_{I_1} V_t(I_1) P^*(I_1|I,u) + C_t(I,u) - C^*(I,u)$. Next, we define

$$r_t^*(I,u) = \beta V^*(I_1^t) - \beta \sum_{I_1} V^*(I_1) P^*(I_1 | I,u) + C_t(I,u) - C^*(I,u).$$

We further separate $w_t = u_t + v_t$ such that

$$u_{t+1}(I,u) = (1-\alpha_t(I,u))u_t(I,u) + \alpha_t(I,u) e_t(I,u)$$
$$v_{t+1}(I,u) = (1-\alpha_t(I,u))v_t(I,u) + \alpha_t(I,u) r_t^*(I,u),$$

where $e_t = r_t - r_t^*$.

In Appendix A, we show that $v_t(I,u) \to 0$ almost surely for all (I, u).

Now, we go back to the iterations:

$$\delta_{t+1}(I,u) = (1-\alpha_t(I,u))\delta_t(I,u) + \alpha_t(I,u)\hat{F}_t(I,u)$$
$$u_{t+1}(I,u) = (1-\alpha_t(I,u))u_t(I,u) + \alpha_t(I,u) e_t(I,u)$$
$$v_{t+1}(I,u) = (1-\alpha_t(I,u))v_t(I,u) + \alpha_t(I,u) r_t^*(I,u).$$

Note that we want to show $\Delta_t = \delta_t + u_t + v_t \to 0$ almost surely, and we have that $v_t(I,u) \to 0$ almost surely for all (I, u). The following analysis holds for any path that belongs to the probability one event in which $v_t(I,u) \to 0$. For any such path and any given $\epsilon > 0$, we can find an $N < \infty$ such that $\|v_t\|_\infty < \epsilon$ for all $t > N$ as (I, u) takes values from a finite set.

We now focus on the term $\delta_t + u_t$ for $t > N$:

$$(\delta_{t+1} + u_{t+1})(I,u) = (1-\alpha_t(I,u))(\delta_t + u_t)(I,u) + \alpha_t(I,u)(\hat{F}_t + e_t)(I,u). \tag{25}$$

Observe that, for $t > N$,

$$(\hat{F}_t + e_t)(I,u) = (F_t - r_t^*)(I,u) = \beta V_t(I_1^t) - \beta V^*(I_1^t) \le \beta \max_{I,u} |Q_t(I,u) - Q^*(I,u)| = \beta \|\Delta_t\|_\infty$$
$$\le \beta \|\delta_t + u_t\|_\infty + \beta\epsilon,$$

where the last step follows from the fact that $v_t \to 0$ almost surely. By choosing $C < \infty$ such that $\hat{\beta} := \beta(C+1)/C < 1$ for $\|\delta_t + u_t\|_\infty > C\epsilon$, we can write that

$$\beta \|\delta_t + u_t + \epsilon\|_\infty \le \hat{\beta} \|\delta_t + u_t\|_\infty.$$

Now, we rewrite (25):

$$\begin{aligned}(\delta_{t+1}+u_{t+1})(I,u) &= (1-\alpha_t(I,u))(\delta_t+u_t)(I,u)+\alpha_t(I,u)(\hat{F}_t+e_t)(I,u)\\ &\le (1-\alpha_t(I,u))(\delta_t+u_t)(I,u)+\alpha_t(I,u)\hat{\beta}\|\delta_t+u_t\|_\infty\\ &< \|\delta_t+u_t\|_\infty. \end{aligned} \tag{26}$$

Hence, $\max_{I,u}((\delta_{t+1}+u_{t+1})(I,u))$ monotonically decreases for $\|\delta_t+u_t\|_\infty > C\epsilon$, and hence, there are two possibilities: it either gets below $C\epsilon$ or it never gets below $C\epsilon$, in which case, by the monotone nondecreasing property, it converges to some number, say M_1 with $M_1 \ge C\epsilon$.

First, we show that, once the process hits below $C\epsilon$, it always stays there. Suppose $\|\delta_t+u_t\|_\infty < C\epsilon$,

$$\begin{aligned}(\delta_{t+1}+u_{t+1})(I,u) &\le (1-\alpha_t(I,u))(\delta_t+u_t)(I,u)+\alpha_t(I,u)\beta(\|\delta_t+u_t\|_\infty+\epsilon)\\ &\le (1-\alpha_t(I,u))C\epsilon+\alpha_t(I,u)\beta(C\epsilon+\epsilon)\\ &= (1-\alpha_t(I,u))C\epsilon+\alpha_t(I,u)\beta(C+1)\epsilon\\ &\le (1-\alpha_t(I,u))C\epsilon+\alpha_t(I,u)C\epsilon, \quad (\beta(C+1)\le C)\\ &= C\epsilon.\end{aligned}$$

We now show that the latter, that is, with the limit being $M_1 \ge C\epsilon$, is not possible. By (26), we have that, for all (I,u),

$$\sum_t \alpha_t(I,u)((\delta_t+u_t)(I,u)) - \hat{\beta}\|\delta_t+u_t\|_\infty) \le (\delta_0+u_0)(I,u) - \liminf_{N\to\infty}(\delta_t+u_t)(I,u).$$

If, pointwise, $(\delta_t+u_t)(I,u)$ admits a limit, the maximum over (I,u) is attained by an individual (I,u) beyond a finite index, and thus, we arrive at

$$\sum_t \alpha_t(I,u)(\|\delta_t+u_t\|_\infty - \hat{\beta}\|\delta_t+u_t\|_\infty) \le (\delta_0+u_0)(I,u) - \liminf_{t\to\infty}(\delta_t+u_t)(I,u),$$

which is a contradiction as α_t is not summable and the difference $(\|\delta_t+u_t\|_\infty - \hat{\beta}\|\delta_t+u_t\|_\infty)$, beyond a finite number, is bounded from below as $\hat{\beta}<1$.

Thus, it suffices to show that $(\delta_t+u_t)(I,u)$ admits a limit. That this limit exists follows from Jaakkola et al. [11, lemma 3] as (26) is an instance of bounded linear iteration under the assumption that $\|\delta_t+u_t\|_\infty$ is bounded.

This shows that the condition $\|\delta_t+u_t\|_\infty > C\epsilon$ cannot be sustained indefinitely for some fixed C (independent of ϵ). Hence, the (δ_t+u_t) process converges to some value below $C\epsilon$ for any path that belongs to the probability one set. Then, we can write $\|\delta_t+u_t\|_\infty < C\epsilon$ for large enough t. Because $\epsilon>0$ is arbitrary, taking $\epsilon\to 0$, we can conclude that $\Delta_t=\delta_t+u_t+v_t\to 0$ almost surely.

Therefore, the process Q_t, determined by the algorithm in (17), converges almost surely to Q^*.

For item (ii), recall that

$$Q^*(I,u) = C^*(I,u) + \beta\sum_{I_1} P^*(I_1|I,u)\min_v Q^*(I,v).$$

This fixed-point equation coincides with the DCOEs for the approximate belief MDP defined in (14) and (15). Hence, using Theorem 3, any policy that satisfies $Q^*(I,\gamma^N(I)) = \min_u Q^*(I,u)$ we can write

$$E[J_\beta(\pi_N^-,\mathcal{T},\gamma^N) - J^*_\beta(\pi_N^-,\mathcal{T})|I_0^N] \le \frac{2\|c\|_\infty}{(1-\beta)}\sum_{t=0}^\infty \beta^t L_t$$

such that

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

and π^* is the invariant measure on x_t under the exploration policy γ. □

A few remarks about the result are now in order.

Remark 3. The learning rates for the standard Q learning algorithm require

- $\sum_k \alpha_k = \infty$.
- $\sum_k \alpha_k^2 < \infty$.

In our case, we have a particular form. To justify this, we note that, although these standard two assumptions on the learning rates may be sufficient for convergence of the algorithm, the limit fixed-point equation (if one exists) is not necessarily useful. Consider the following example in which the state space is $\mathbb{X} = \{-1, +1\}$ and transitions are deterministic such that $Pr(x_{t+1} = 1 | x_t = -1) = 1$, $Pr(x_{t+1} = -1 | x_t = +1) = 1$ (leading to a periodic Markov chain). If one chooses the learning rates as $\alpha_{2k} = 0$, $\alpha_{2k+1} = \sigma_k$ for every k such that σ_k is square summable but not summable, then even though the algorithm converges, depending on the initial point, one of the transition models always dominates the other. To avoid such examples, we choose the learning rates to be "averaging" through time.

Remark 4. We caution the reader that our result assumes that the cost starts running after time N; that is, the effective cost is

$$E\left[\sum_{k=N}^{\infty} \beta^{k-N} c(x_k, u_k)\right]. \tag{27}$$

Of course, this criterion is also applicable if the system starts running prior to time $-N$ and the costs become in effect after time 0.

If this criterion is not applicable, and the first N stages are also crucial, (i) if β is large enough, we can conclude that the first N stages are not as critical for the analysis as their contributions are minor in comparison with the future stages for the criterion, which can also be seen by considering this equivalent criterion to (3) and noting that, for large enough β, the contributions of the first N time stages become negligible:

$$(1-\beta)E\left[\sum_{k=0}^{\infty} \beta^{k} c(x_k, u_k)\right].$$

(ii) On the other hand, if β is not large and if the cost starts running at time 0, then we can first run the Q learning algorithm to find the best N-window policies that optimize (27). The remaining question is to optimize

$$E\left[\sum_{k=0}^{N-1} c(x_k, u_k) + V(I_k)\right] \tag{28}$$

as a finite-horizon optimal control problem with a terminal cost, and the terminal cost V can be estimated by (27) via Theorems 2 and 4. The question then becomes how to select the first N actions, leading to a problem with a finite search complexity for a finite horizon problem, without knowing the system dynamics. For this, one can run a Markov chain Monte Carlo algorithm in parallel simulations to find the optimal policy for the first N time stages. Because the resulting policy minimizing (28) is at least as good as the first N-window policy under the optimal (belief MDP) policy (which is not designed to optimize (28) but the original cost (3)), the bounds presented in Theorem 3 are applicable even when the cost criterion includes the first N time stages.

Remark 5. For the convergence rate of the Q learning algorithm (17), it is clear that increasing the window size increases the size of the state space, which, in turn, slows down the convergence speed of the iterations. The sample complexity in Q learning is studied extensively for various learning rates. For linear learning rates, $\alpha_k = \frac{1}{k}$, as we use in this paper, the conclusion is that (see, e.g., Even-Dar et al. [3]) the sample complexity is in the order of $\frac{(|\mathbb{Y}| \times |\mathbb{U}|)^{\frac{N}{1-\beta}}}{(1-\beta)^4 \epsilon^2}$ when one seeks ϵ optimality in expectation. Thus, the number of samples needed to get an ϵ-near estimate in expectation increases exponentially in the window size. However, we note that this is the number of samples needed to get ϵ-near to the limit Q value for the specific window size N, which gets closer to the optimal Q value of the original POMDP exponentially fast with increasing N under suitable filter stability assumptions (see Corollaries 2 and 3). We also note that, for different system parameters, it is shown that better convergence speeds can be achieved with carefully chosen learning rates (see, e.g., Li et al. [17], Wainwright [37]). Hence, a learning rate that is adaptive to the window size N and target total approximation error (in near optimality as well as the error from sample complexity) can be used for a faster learning.

5. On the Filter Stability and Convergence to Near Optimality Under Filter Stability

In this section, we discuss the (uniform filter stability) term L_t defined in (16):

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

Before we introduce related definitions and notation, we again emphasize that, for the performance of the approximate model introduce in Section 3.2, L_t plays a crucial role. As we note earlier, L_t is a term related to controlled filter stability, a general problem in which one is interested in how fast a process forgets its incorrect initial prior with increasing observation and control variables over time. In particular, any quantitative result for controlled filter stability bounding the errors given by L_t can be used to study the performance of the approximation and the learning algorithm introduced in this paper.

We state this formally as follows.

Corollary 1. *Suppose the following assumptions hold:*

- *Assumption 3 holds.*
- *Under the exploring policy, γ, the state process $\{x_t\}_t$ is irreducible.*
- *The POMDP is such that the filter is stable uniformly over priors in expectation under total variation, that is, $L_t \to 0$ as $N \to \infty$.*

Then, for any policy γ^N that satisfies $Q^*(I,\gamma^N(I)) = \min_u Q^*(I,u)$, if we assume that the controller starts using γ^N at time $t = N$ (after observing at least N information variables), then denoting the prior distribution of X_N by π_N^- conditioned on the first N step information variables, we have

$$E[J_\beta(\pi_N^-,\mathcal{T},\gamma^N) - J^*_\beta(\pi_N^-,\mathcal{T})|I_0^N] \to 0$$

as $N \to \infty$.

Proof. The result follows directly from Theorem 4 and the definition of L_t (see (16)). □

Notably, directly related to L_t, recent results in the literature, in particular McDonald and Yüksel [24, theorem 3.6; 23, theorem 3.3], present explicit and sufficient conditions on the controlled filter stability problem under the total variation metric in expectation.

We first focus on McDonald and Yüksel [23, theorem 3.3] and recall the Dobrushin coefficient for Markov kernels, which provide exponential convergence rates for L_t.

Definition 2. For a given prior measure μ on X_0 and a policy γ, the one-step predictor process is defined as the sequence of conditional probability measures

$$\pi_{t-}^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_t \in \cdot|Y_{[0,t-1]}, U_{[t-1]} = \gamma(Y_{[0,t-1]})),$$

where $P^{\mu,\gamma}$ is the probability measure induced by the prior μ and the policy γ, when μ is the probability measure on X_0.

Definition 3. The filter process is defined as the sequence of conditional probability measures

$$\pi_t^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_t \in \cdot|Y_{[0,t]}, U_{[t-1]} = \gamma(Y_{[0,t-1]})), \tag{29}$$

where $P^{\mu,\gamma}$ is the probability measure induced by the prior μ and the policy γ.

Definition 4 (Dobrushin [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

$$\delta(K) = \inf \sum_{i=1}^{n} \min(K(x,A_i),K(y,A_i)), \tag{30}$$

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

We note that this definition holds for continuous or finite/countable spaces S_1 and S_2 and $0 \le \delta(K) \le 1$ for any kernel operator.

Example 1. Assume, for a finite setup, we have 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 in which 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

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

Theorem 5 (McDonald and Yüksel [23, theorem 3.3]). *Assume that, for $\mu, \nu \in \mathcal{P}(\mathbb{X})$, we have $\mu \ll \nu$. Then, we have*

$$E^{\mu,\gamma}[\|\pi_{n+1}^{\mu,\gamma} - \pi_{n+1}^{\nu,\gamma}\|_{TV}] \leq (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q))E^{\mu,\gamma}[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}].$$

In particular, defining $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q))$, we have

$$E^{\mu,\gamma}[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}] \leq 2\alpha^n.$$

The absolute continuity assumption, that is, $\mu \ll \nu$, can be interpreted as follows: assume that the true starting distribution is μ, but we start the update from an incorrect prior ν. The error can be fixed with the information $y_{[0,t]}, u_{[0,t-1]}$, eventually, as long as the incorrect starting distribution ν puts on a positive measure to every event that the real starting distribution μ puts on a positive measure. However, if it is not the case, that is, if the incorrect starting distribution ν puts zero measure to some event that μ puts positive measure to, information variables are not sufficient to fix the starting error occurring from that zero measure event. Of course, this is not feasible as the prior is not compatible with the measured data. In any case, in our setup, the incorrect prior serves as an approximation, and this can be made to satisfy the absolute continuity condition by design: this is the invariant measure on the state under the exploration policy.

Recall that the Q learning iteration that uses finite window information variables learns the Q values for approximate states of the form (π^*, I_t^N) instead of the true states (π_{t-N}^-, I_t^N). Theorem 5 suggests that the approximation error arising from using the stationary distribution, π^*, instead of π_{t-N}^-, can be fixed with the information variables I_t^N if π^* captures the nonzero events of π_{t-N}^-, that is, if $\pi_{t-N}^- \ll \pi^*$.

In particular, because $\tilde{\delta}(\mathcal{T})$ is a uniform Dobrushin coefficient over all control actions, the above bound is valid under any control action process. Thus, if $\pi_t^- \ll \pi^*$ for all t, then we can write

$$\begin{aligned} L_t &= \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[\|P^{\pi_t^-}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\hat{\pi}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV}] \\ &\leq 2\alpha^N \end{aligned} \tag{31}$$

for all t.

Corollary 2 (To Theorems 4 and 5). *Assume the following hold:*

- *Assumption 3 holds.*
- *The state space, $\mathbb{X}$, is finite.*
- *Under the exploring policy, γ, the state process $\{x_t\}_t$ is irreducible.*
- $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(O)) < 1$.

Then, for any policy γ^N that satisfies $Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u)$, if we assume that the controller starts using γ^N at time $t = N$ (after observing at least N information variables), then denoting the prior distribution of X_N by π_N^- conditioned on the first N step information variables, we have

$$E[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N] \leq \frac{4\|c\|_\infty}{(1-\beta)^2}\alpha^N.$$

Proof. Note that, by Theorem 4,

$$E[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N] \le \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t.$$

If the state process x_t is irreducible under the exploring policy, then by Kac's [13] lemma, we have that

$$\pi^*(x) > 0, \qquad \forall x \in \mathbb{X}.$$

Hence, using Inequality (31), we complete the proof. □

Corollary 3 (To Theorems 4 and 5). *Assume the following hold:*

- *Assumption 3 holds.*
- *$\mathbb{X} \subset \mathbb{R}^m$ for some $m < \infty$.*
- *The transition kernel $\mathcal{T}(\cdot|x_0, u_0)$ admits a density function f with respect to a measure ϕ such that $\mathcal{T}(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$ and $f(x_1, x_0, u_0) > 0$ for all x_1, x_0, u_0.*
- *$\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(O)) < 1$.*

Then, for any policy γ^N that satisfies $Q^*(I, \gamma^N(I)) = \min_u Q^*(I, u)$, if we assume that the controller starts using γ^N at time $t = N$ (after observing at least N information variables), then denoting the prior distribution of X_N by π_N^- conditioned on the first N step information variables, we have

$$E[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N] \le \frac{4\|c\|_\infty}{(1-\beta)^2} \alpha^N.$$

Proof. Note that, by assumption, $\mathcal{T}(dx_1|x_0, u_0) = f(x_1, x_0, u_0)\phi(dx_1)$ and $f(x_1, x_0, u_0) > 0$ for all x_1, x_0, u_0, and hence, under the exploration policy γ, the state process x_t is ϕ-irreducible and admits a unique invariant measure, say π^*. Using the assumptions, we can also write that, for any $A \in \mathcal{B}(\mathbb{X})$ with $\phi(A) > 0$,

$$\pi^*(A) = \int_{\mathbb{Z}} \int_A \int_{\mathbb{U}} f(x_1, x_0, u_0)\gamma(du_0)\phi(dx_1)\pi^*(dx_0) > 0,$$

which implies that $\phi \ll \pi^*$. Note that the transition kernel $\mathcal{T}(\cdot|x, u)$ is absolutely continuous with respect to ϕ for every (x, u), and thus, for the predictor π_t^- at any time step t, we can write that $\pi_t^- \ll \phi \ll \pi^*$.

Hence, Inequality (31) and Theorem 4 concludes the proof. □

6. Numerical Study

In this section, we present a numerical study for the proven results.

The example we use is a machine repair problem. In this model, we have $\mathbb{X}, Y, U = \{0, 1\}$ with

$$x_t = \begin{cases} 1 & \text{machine is working at time } t \\ 0 & \text{machine is not working at time } t. \end{cases}$$

$$u_t = \begin{cases} 1 & \text{machine is being repaired at time } t \\ 0 & \text{machine is not being repaired at time } t. \end{cases}$$

The probability that the repair is successful given initially the machine was not working is given by κ:

$$Pr(x_{t+1} = 1 | x_t = 0, u_t = 1) = \kappa.$$

The probability that the machine breaks down when in a working state is given by θ:

$$Pr(x_t = 0 | x_t = 1, u_t = 0) = \theta.$$

The probability that the channel gives an incorrect measurement is given by ϵ:

$$Pr(y_t = 1 | x_t = 0) = Pr(y_t = 0 | x_t = 1) = \epsilon.$$

The one-stage cost function is given by

$$c(x,u) = \begin{cases} R+E & x=0, u=1 \\ E & x=0, u=0 \\ 0 & x=1, u=0 \\ R & x=1, u=1, \end{cases}$$

where R is the cost of repair and E is the cost incurred by a broken machine.

We study the example with discount factor $\beta = 0.8$ and present three different results by changing the other parameters.

For the first case, we take $\epsilon = 0.3$, $\kappa = 0.8$, $\theta = 0.1$ and $R = 5, E = 1$. For the exploring policy, we use a random policy such that $Pr(\gamma(x) = 0) = \frac{1}{2}$ and $Pr(\gamma(x) = 1) = \frac{1}{2}$ for all x. Under this policy, x_t admits a stationary policy $\pi^*(\cdot) = 0.1\delta_0(\cdot) + 0.9\delta_1(\cdot)$.

We prove in Theorem 4 that the Q iteration given by (17) converges to the Q values of the approximate belief MDP defined in (14). Defining

$$V_t(I) := \min_{v \in \mathbb{U}} Q_t(I, v),$$

in Figure 2, we plot $\sup_I |V_t(I) - J_\beta^N(\pi^*, I)|$ for $n = 0, 1, 2$.

We now show the performance of γ^N that is found using the Q values for different values of N. Recall that, in Theorem 4, we show that

$$E[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t,$$

where

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

Figure 2. (Color online) Q value convergence for different window sizes.

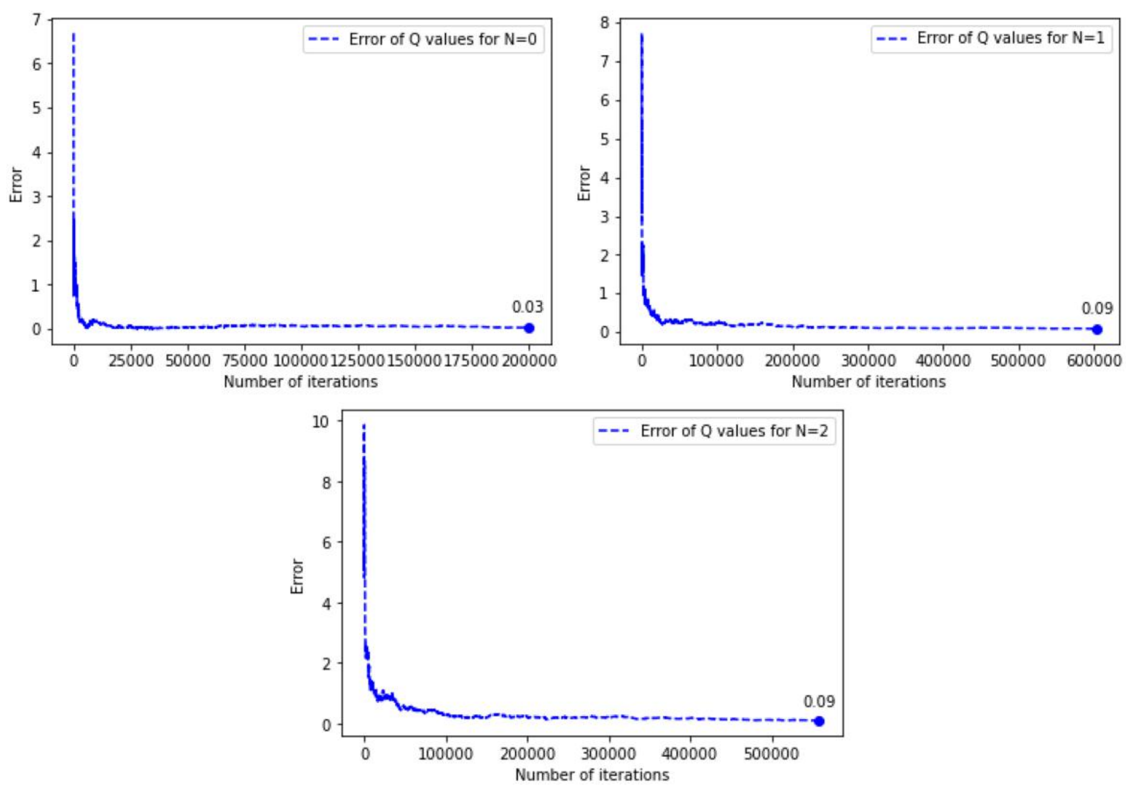
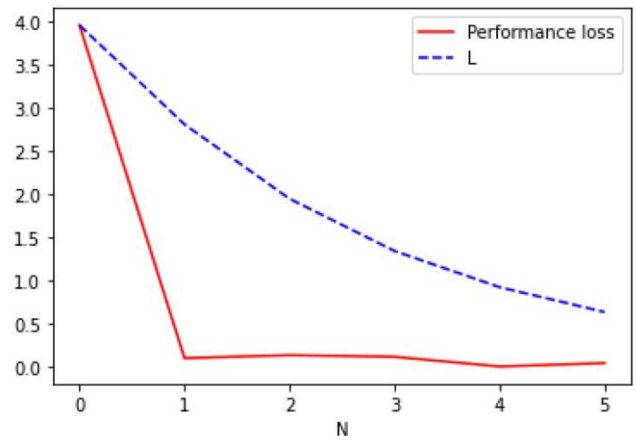


Figure 3. (Color online) Policy performance for different window sizes.



For the examples, we use the following upper bound for L_t

$$L := \sup_{\pi \in \mathcal{P}(\mathbb{X})} \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{\pi}^{\hat{\gamma}}[\|P^{\pi}(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]}) - P^{\pi^*}(X_N \in \cdot | Y_{[0,N]}, U_{[0,N-1]})\|_{TV}]$$

such that

$$E[J_\beta(\pi_N^-, \mathcal{T}, \gamma^N) - J_\beta^*(\pi_N^-, \mathcal{T}) | I_0^N] \leq \frac{2\|c\|_\infty}{(1-\beta)^2} L.$$

In Figure 3, to estimate $J_\beta^*(\mu, \mathcal{T})$, we simply use the smallest value of $J_\beta(\mu, \mathcal{T}, \gamma^N)$ among the different N values. Furthermore, we scale the L values to show the rate dependence between $J_\beta(\mu, \mathcal{T}, \gamma^N) - J_\beta^*(\mu, \mathcal{T})$ and L more clearly.

It is clearly seen that the decrease rate for L dominates the decrease rate for the error.

For the second case, we take $\epsilon = 0.1$, $\kappa = 0.9$, $\theta = 0.3$ and $R = 5, E = 1$. For the exploring policy, we again use a random policy such that $Pr(\gamma(x) = 0) = \frac{1}{2}$ and $Pr(\gamma(x) = 1) = \frac{1}{2}$ for all x. Under this policy, x_t admits a stationary policy $\pi^*(\cdot) = 0.29\delta_0(\cdot) + 0.71\delta_1(\cdot)$.

Figure 4 shows the error between $V_t(I) = \min_v Q_t(I, v)$ and $J_\beta^N(\pi^*, I)$ for $n = 0, 1, 2$.

Figure 5 shows $J_\beta(\mu, \mathcal{T}, \gamma^N) - J_\beta^*(\mu, \mathcal{T})$ and scaled L.

In the third case, note that, for the previous examples, we had $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(O)) > 1$; however, the error still decreases because the $\alpha > 1$ condition is only a sufficient condition and the error still converges to zero even when $\alpha > 1$ in some cases. For the last example, we set parameters so that $\alpha < 1$. The parameters are chosen as follows:

$$Pr(x_1 = 0 | x_0 = 0, u_0 = 0) = 0.9, \quad Pr(x_1 = 0 | x_0 = 0, u_0 = 1) = 0.6,$$
$$Pr(x_1 = 0 | x_0 = 1, u_0 = 0) = 0.4, \quad Pr(x_1 = 0 | x_0 = 1, u_0 = 1) = 0.1.$$

Notice that we manipulated some of the parameters to make the α coefficient suitable for the purpose of the example. For the measurement channel,

$$Pr(y = 0 | x = 0) = 0.7, \quad Pr(y = 1 | x = 1) = 0.7.$$

For the cost function, we choose $R = 3$ and $E = 1$.

We again use a random policy such that $Pr(\gamma(x) = 0) = \frac{1}{2}$ and $Pr(\gamma(x) = 1) = \frac{1}{2}$ for all x. Under this policy, x_t admits a stationary policy $\pi^*(\cdot) = 0.42\delta_0(\cdot) + 0.58\delta_1(\cdot)$.

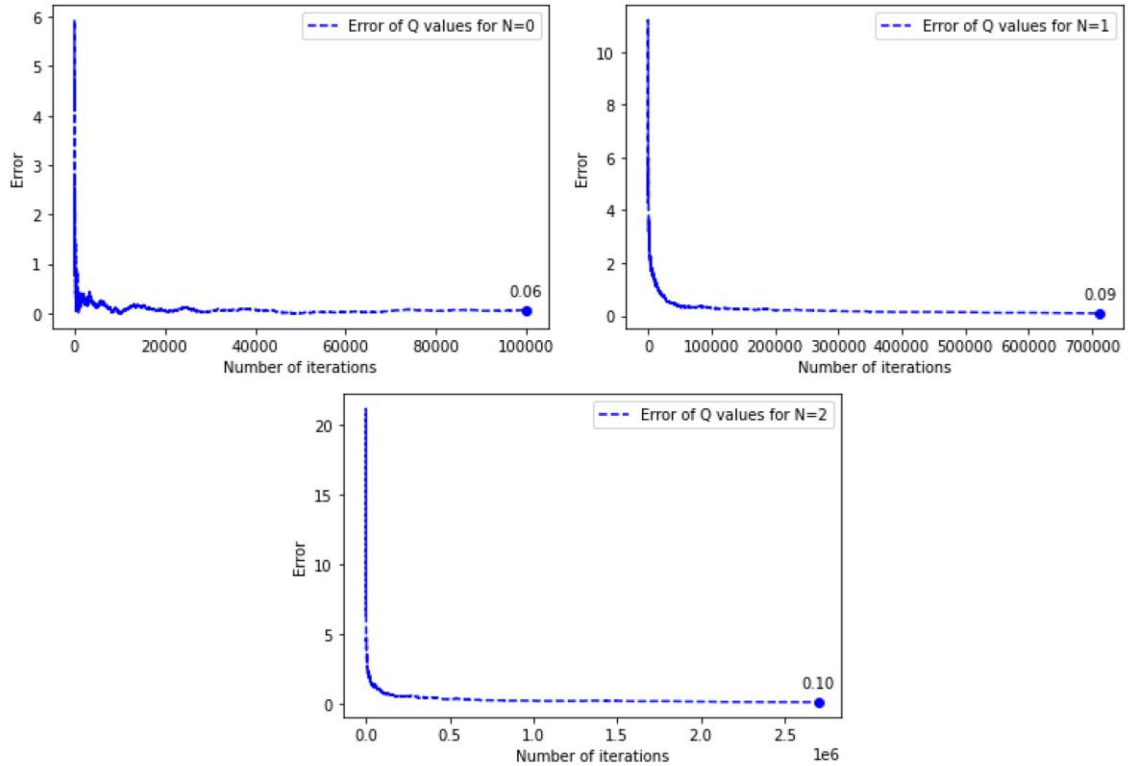
The convergence of the Q values can be seen in Figure 6.

This setup gives $\alpha = 0.7$. Figure 7 shows the error $J_\beta(\mu, \mathcal{T}, \gamma^N) - J_\beta^*(\mu, \mathcal{T})$, L, and α^N terms. We scale all of them to make them start from the same point to emphasize the decrease rates.

7. Concluding Remarks and a Discussion

We study the convergence of an approximate Q learning algorithm for partially observed stochastic control systems that uses finite window history variables. We provide sufficient conditions that guarantee the algorithm to converge and then we provide the approximate belief-MDP model to which the limit fixed equation corresponds.

Figure 4. (Color online) Q value convergence for different window sizes.



Furthermore, we provide bounds for the approximate policy that is learned with the proposed algorithm in comparison with the true optimal policy that could be designed if the system and channel were known a priori. In particular, we obtain 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.

The setup we use for this paper focuses on continuous state space and finite observation and action spaces. An immediate future direction is for continuous observation and action spaces, in which case, continuity properties of the transition model $\mathcal{T}(\cdot|x,u)$ on u and continuity properties of the channel $O(dy|x)$ on x are crucial and sufficient for consistent discretization of the observation and action spaces, leading to analogous stability results. One condition, for example, is that the channel be of the form $\int_A O(dy|x) = \int_A f(x,y)dy$ for all Borel A with f continuous in both variables.

Figure 5. (Color online) Policy performance for different window sizes.

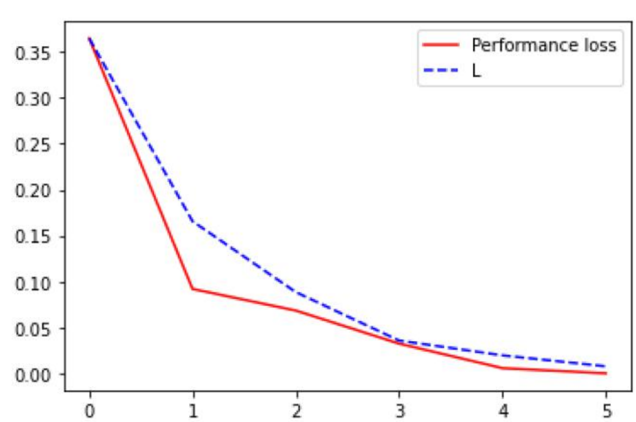
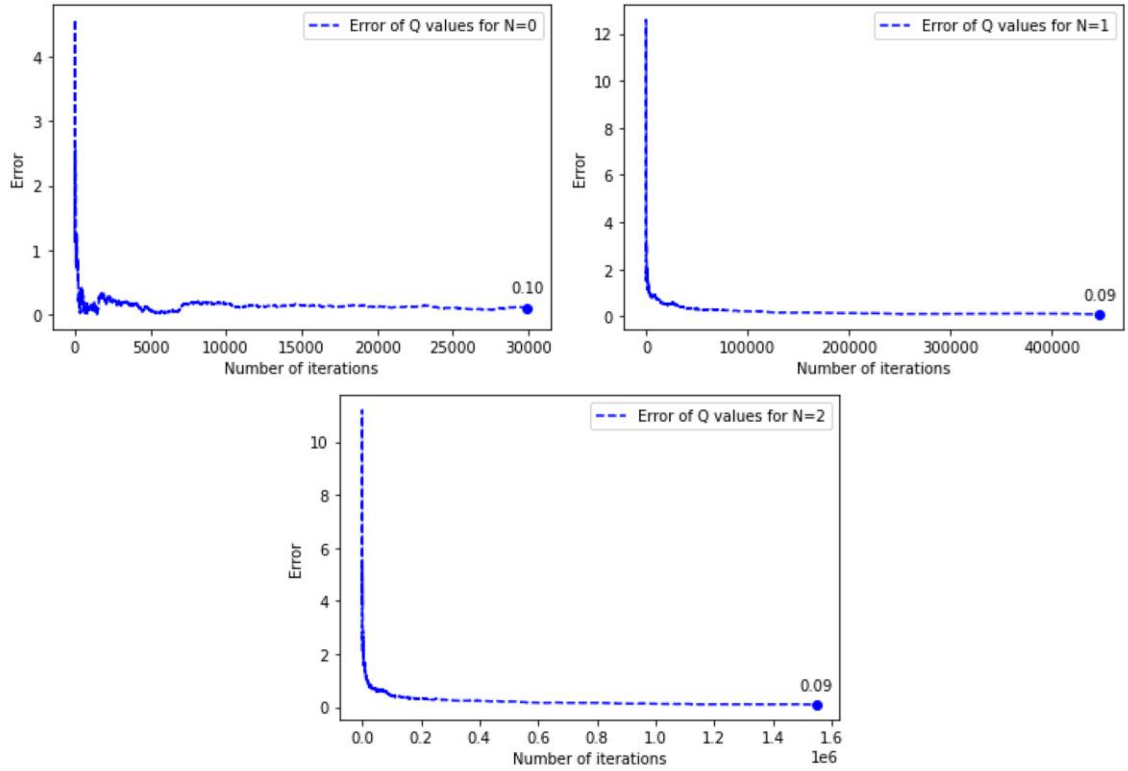


Figure 6. (Color online) Q value convergence for different window sizes.



It is also our goal to generalize such results to multiagent problems in which finite history policies likely lead to new insights toward tractable solutions in both stochastic team theory and game theory.

Appendix A. Proof of $v_t(I, u) \to 0$

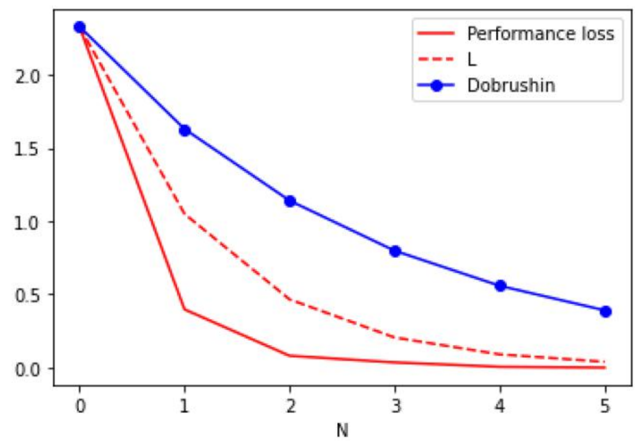
We show that $v_t(I, u) \to 0$ almost surely for all (I, u). We prove the claim only for the $N = 1$ case for simplicity and let $I = (y_1, y_0, u_0)$ and $u = u_1$ for some $(y_1, y_0, u_0, u_1) \in \mathbb{Y}^2 \times \mathbb{U}^2$. The proof for general N follows from essentially same steps. We have

$$v_{t+1}(I, u) = (1 - \alpha_t(I, u))v_t(I, u) + \alpha_t(I, u) r_t^*(I, u).$$

When the learning rates are chosen such that $\alpha_t(I, u) = 0$ unless $(I_t, U_t) = (I, u)$ and

$$\alpha_t(I, u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{I_k = I, U_k = u\}}},$$

Figure 7. (Color online) Policy performance for different window sizes.



this term reduces to

$$v_{t+1}(I,u) = \frac{\sum_{k=0}^{t-1} r_k^*(I,u)\mathbb{1}_{\{I_k,U_k=I,u\}}}{\sum_{k=0}^{t-1}\mathbb{1}_{\{I_k,U_k=I,u\}}}.$$

Recall that

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

Hence, we first analyze the asymptotic behavior of

$$I_1^k := (Y_{k+1}, Y_k, U_k),$$
$$I_k, U_k = (Y_k, Y_{k-1}, U_k, U_{k-1}).$$

To analyze the asymptotic behavior of these variables, we make use of the Markov chain theory. We show that

$$I_1^k, I_k, U_k, X_{k-1} = \{Y_{k+1}, Y_k, Y_{k-1}, U_k, U_{k-1}, X_{k-1}\}$$

form a Markov chain under the exploration policy γ. Then, we use stationary distribution of this Markov chain to analyze the asymptotic behavior of (I_1^k, I_k, U_k). We write

$$\begin{aligned}
&Pr(Y_{k+1}, Y_k, Y_{k-1}, U_k, U_{k-1}, X_{k-1}|y_k, y_{k-1}, \ldots, y_0, u_{k-1}, \ldots, u_0, x_{k-2}, \ldots, x_0)\\
&= \mathbb{1}_{\{Y_k,Y_{k-1},U_{k-1}=y_k,y_{k-1},u_{k-1}\}} Pr(Y_{k+1}|y_k, x_{k-1}, u_k, u_{k-1}) Pr(X_{k-1}|y_{k-1}, x_{k-2}, u_{k-2})\gamma(U_k)\\
&= Pr(Y_{k+1}, Y_k, U_k, X_{k-1}, U_{k-2}|y_k, y_{k-1}, x_{k-2}, u_{k-1}, u_{k-2}),
\end{aligned}$$

where we use γ as the probability measure of the exploring policy. Earlier, we used that

$$Pr(Y_{k+1}|y_k, \ldots, y_0, x_{k-1}, \ldots, x_0, \ldots, u_k, \ldots, u_0) = Pr(Y_{k+1}|y_k, x_{k-1}, u_k, u_{k-1})$$
$$Pr(X_{k-1}|y_{k-1}, \ldots, y_0, x_{k-2}, \ldots, x_0, u_{k-2} \ldots, u_0) = Pr(X_{k-1}|y_{k-1}, x_{k-2}, u_{k-2}).$$

Hence, the joint process $(Y_{k+1}, Y_k, Y_{k-1}, U_k, U_{k-1}, X_{k-1})$ is a Markov chain. One can show that it has a unique invariant measure under the assumption that the state process X_k admits a unique invariant measure under the exploration policy γ.

We denote the stationary distribution of X_k by π^* and denote the probability measure induced on the joint process and marginals of the process by this stationary distribution by P^{π^*} with an abuse of notation. Then, for any measurable function f,

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1} f(I_1^k, I_k, U_k, X_k) = \int f(y_2', y_1', y_0', u_1', u_0', x_0')P^{\pi^*}(dx_0', y_0', y_1', y_2', u_0', u_1').$$

In particular, we have that

$$\begin{aligned}
\lim_{t\to\infty}\frac{\frac{1}{t}\sum_{k=0}^{t-1} V_k^*(I,u)\mathbb{1}_{\{I_k,u_k=I,u\}}}{\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{1}_{\{I_k,u_k=I,u\}}} &= \frac{\int V^*(y_2', y_1', u_1')\mathbb{1}_{\{y_1',y_0',u_1',u_0'=y_1,y_0,u_1,u_0\}}P^{\pi^*}(y_0', y_1', y_2', u_0', u_1')}{\int \mathbb{1}_{\{y_1',y_0',u_1',u_0'=y_1,y_0,u_1,u_0\}}P^{\pi^*}(y_0', y_1', y_2', u_0', u_1')}\\
&= \frac{\int_{y_2'} V^*(y_2', y_1, u_1)P^{\pi^*}(y_2', y_1' = y_1, y_0' = y_0, u_1' = u_1, u_0' = u_0)}{P^{\pi^*}(y_1' = y_1, y_0' = y_0, u_1' = u_1, u_0' = u_0)}\\
&= \sum_{\mathbb{Y}} V^*(y_2', y_1, u_1)P^{\pi^*}(y_2'|y_1, y_0, u_1, u_0)\\
&= \sum_{I_1} V^*(I_1)P^*(I_1|I,u),
\end{aligned}$$

where $P^{\pi^*}(y_2|y_1,y_0,u_1,u_0)$ is the distribution of y_2 when the x_0's marginal distribution is given by π^* and

$$P^*(I_1=(y_2,y_1',u_1')|I,u) := \mathbb{1}_{\{y_1'=y_1,u_1'=u_1\}}P^{\pi^*}(y_2|y_1,y_0,u_1,u_0)$$

as defined in (18) and (19).

Using similar arguments, one can also show that, for $I=(y_1,y_0,u_0)$ and $u=u_1$,

$$\lim_{k\to\infty}\frac{1}{k}\sum_{k'=0}^{k-1}C_{k'}(I,u)=\int_{\mathbb{X}}c(x_1,u_1)P^{\pi^*}(dx_1|y_1,y_0,u_0)$$
$$=C^*(y_1,y_0,u_0,u_1)=C^*(I,u).$$

Thus, we have that

$$v_{t+1}(I,u)=\frac{1}{t}\sum_{k=0}^{t-1}r_k^*(I,u)\to 0$$

almost surely for all $(I,\,u)$.

Appendix B. Proof of Theorem 2

Lemma B.1. *We have that, for any $\pi,\pi^*\in\mathcal{P}(\mathbb{X})$ and for any $(y,u)_{[t,t-N]}:=\{y_t,\dots,y_{t-N},u_t,\dots,u_{[t-N]}\}\in\mathbb{Y}^N\times\mathbb{U}^N$,*

$$\|P^{\pi}(Y_{t+1}\in\cdot|(y,u)_{[t,t-N]})-P^{\pi^*}(Y_{t+1}\in\cdot|(y,u)_{[t,t-N]})\|_{TV}$$
$$\le\|P^{\pi}(X_t\in\cdot|y_{[t,t-N]},u_{[t-1,t-N]})-P^{\pi^*}(X_t\in\cdot|y_{[t,t-N]},u_{[t-1,t-N]})\|_{TV}.$$

Proof. Let f be a measurable function of $\mathbb{Y}$ such that $\|f\|_\infty\le 1$. We write

$$\int f(y_{t+1})P^{\pi}(dy_{t+1}|(y,u)_{[t,t-N]})-\int f(y_{t+1})P^{\pi^*}(dy_{t+1}|(y,u)_{[t,t-N]})$$
$$=\int f(y_{t+1})O(dy_{t+1}|x_{t+1})\mathcal{T}(dx_{t+1}|x_t,u_t)P^{\pi}(dx_t|y_{[t,t-N]},u_{[t-1,t-N]})$$
$$-\int f(y_{t+1})O(dy_{t+1}|x_{t+1})\mathcal{T}(dx_{t+1}|x_t,u_t)P^{\pi^*}(dx_t|y_{[t,t-N]},u_{[t-1,t-N]})$$
$$\le\|P^{\pi}(X_t\in\cdot|y_{[t,t-N]},u_{[t-1,t-N]})-P^{\pi^*}(X_t\in\cdot|y_{[t,t-N]},u_{[t-1,t-N]})\|_{TV}$$

at the last step, and we used the fact that $\int f(y_{t+1})O(dy_{t+1}|x_{t+1})\mathcal{T}(dx_{t+1}|x_t,u_t)$ is bounded by one as a function of x_t for $\|f\|_\infty\le 1$. Taking the supremum over all $\|f\|_\infty\le 1$ concludes the proof. □

Proof of the Main Theorem. Let $\hat{z}_0=(\pi_0^-,y_1,y_0,u_0)$. Then, we write

$$\tilde{J}_\beta^N(\hat{z}_1)=J_\beta^N(\pi^*,y_1,y_0,u_0)=\min_{u_1\in\mathbb{U}}\left(\hat{c}(\pi^*,y_1,y_0,u_0,u_1)+\beta\sum_{y_2\in\mathbb{Y}}J_\beta^N(\pi^*,y_2,y_1,u_1)P^{\pi^*}(y_2|y_1,y_0,u_1,u_0)\right).$$

Furthermore,

$$J_\beta^*(\hat{z}_1)=J_\beta^*(\pi_0^-,y_1,y_0,u_0)$$
$$=\min_{u_1\in\mathbb{U}}\left(\hat{c}(\pi_0^-,y_1,y_0,u_0,u_1)+\beta\sum_{y_2\in\mathbb{Y}}J_\beta^*(\pi_1^-(\pi_0^-,y_0,u_0),y_2,y_1,u_1)P^{\pi_0^-}(y_2|y_1,y_0,u_1,u_0)\right).$$

Note that, for any $\pi\in\mathcal{P}(\mathbb{X})$, we have

$$\tilde{J}_\beta^N(\pi,y_2,y_1,u_1)=\tilde{J}_\beta^N(\pi^*,y_2,y_1,u_1)=J_\beta^N(\pi^*,y_2,y_1,u_1).$$

In particular, we have that

$$J_\beta^N(\pi^*,y_2,y_1,u_1)=\tilde{J}_\beta^N(\pi_1^-(\pi_0^-,y_0,u_0),y_2,y_1,u_1).$$

Hence, we can write the following:

$$
\begin{aligned}
|\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0)| &\leq \max_{u_1 \in \mathbb{U}} |\hat{c}(\pi^*, y_1, y_0, u_0, u_1) - \hat{c}(\pi_0^-, y_1, y_0, u_0, u_1)| \\
&+ \max_{u_1 \in \mathbb{U}} \beta \left| \sum_{y_2 \in \mathbb{Y}} J_\beta^N(\pi^*, y_2, y_1, u_1) P^{\pi^*}(y_2|y_1, y_0, u_1, u_0) - \sum_{y_2 \in \mathbb{Y}} J_\beta^N(\pi^*, y_2, y_1, u_1) P^{\pi_0^-}(y_2|y_1, y_0, u_1, u_0) \right| \\
&+ \max_{u_1 \in \mathbb{U}} \beta \sum_{y_2 \in \mathbb{Y}} |\tilde{J}_\beta^N(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1) - J_\beta^*(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1)| P^{\pi_0^-}(y_2|y_1, y_0, u_1, u_0).
\end{aligned}
$$

Note that, by the definition of $\hat{c}$, we have

$$
|\hat{c}(\pi^*, y_1, y_0, u_0, u_1) - \hat{c}(\pi_0^-, y_1, y_0, u_0, u_1)| \leq \|c\|_\infty \|P^{\pi^*}(X_1 \in \cdot|y_1, y_0, u_0) - P^{\pi_0^-}(X_1 \in \cdot|y_1, y_0, u_0)\|_{TV}.
$$

If we denote $\hat{z}_1 = ((\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1)$, using Lemma B.1, we can write

$$
\begin{aligned}
E_{\pi_0^-}^\gamma[|\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0)|] &\leq \|c\|_\infty E_{\pi_0^-}^\gamma[\|P^{\pi^*}(X_1 \in \cdot|Y_1, Y_0, U_0) - P^{\pi_0^-}(X_1 \in \cdot|Y_1, Y_0, U_0)\|_{TV}] \\
&+ \max_{u_1 \in \mathbb{U}} \beta \|J_\beta^N\|_\infty E_{\pi_0^-}^\gamma[\|P^{\pi_0^-}(y_2|Y_1, Y_0, U_1, U_0) - P^{\pi^*}(y_2|Y_1, Y_0, U_1, U_0)\|_{TV}] \\
&+ \max_{u_1 \in \mathbb{U}} \beta E_{\pi_0^-}^\gamma \left[\sum_{y_2 \in \mathbb{Y}} |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^*(\hat{z}_1)| P^{\pi_0^-}(y_2|Y_1, Y_0, U_1, U_0) \right] \\
&\leq (\|c\|_\infty + \beta \|J_\beta^N\|_\infty) L_0 + \max_{u_1 \in \mathbb{U}} \beta E_{\pi_0^-}^\gamma \left[\sum_{y_2 \in \mathbb{Y}} |\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^*(\hat{z}_1)| P^{\pi_0^-}(y_2|Y_1, Y_0, u_1, U_0) \right] \\
&\leq (\|c\|_\infty + \beta \|J_\beta^N\|_\infty) L_0 + \sup_{\hat{\gamma} \in \hat{\Gamma}} \beta E_{\pi_0^-}^{\hat{\gamma}}[|\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^*(\hat{z}_1)|],
\end{aligned}
$$

where

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

Then, following the same steps for $E_{\pi_0^-}^{\hat{\gamma}}[|\tilde{J}_\beta^N(\hat{z}_1) - J_\beta^*(\hat{z}_1)|]$ and repeating the procedure, one can see that

$$
E_{\pi_0^-}^\gamma[|\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0)|] \leq (\|c\|_\infty + \beta \|J_\beta^N\|_\infty) \sum_{t=0}^\infty \beta^t L_t.
$$

Note that $\|J_\beta^N\|_\infty \leq \frac{\|c\|_\infty}{1-\beta}$. Hence, we can conclude

$$
E_{\pi_0^-}^\gamma[|\tilde{J}_\beta^N(\hat{z}_0) - J_\beta^*(\hat{z}_0)|] \leq \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^\infty \beta^t L_t. \quad \square
$$

Appendix C. Proof of Theorem 3

We let $\hat{z}_0 = (\pi_0^-, y_1, y_0, u_0)$. We denote the minimum selector for the approximate MDP by

$$
u_1^N := \tilde{\phi}^N(\pi_0^-, y_1, y_0, u_0) = \phi^N(\pi^*, y_1, y_0, u_0)
$$

and write

$$
\begin{aligned}
J_\beta(\hat{z}_0, \tilde{\phi}^N) &= J_\beta(\pi_0^-, y_1, y_0, u_0, \tilde{\phi}^N) \\
&= \hat{c}(\pi_0^-, y_1, y_0, u_0, u_1^N) + \beta \sum_{y_2 \in \mathbb{Y}} J_\beta(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1^N, \tilde{\phi}^N) P^{\pi_0^-}(y_2|y_1, y_0, u_1^N, u_0).
\end{aligned}
$$

Furthermore, we write the optimality equation for $\tilde{J}_\beta^N$ as follows

$$\tilde{J}_\beta^N(\hat{z}_0) = \hat{c}(\pi^*, y_1, y_0, u_0, u_1^N) + \beta \sum_{y_2 \in \mathbb{Y}} \tilde{J}_\beta^N(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1^N) P^{\pi^*}(y_2 | y_1, y_0, u_1^N, u_0).$$

Hence, denoting $\hat{z}_1 := \left(\pi_1^-(\pi_0^-, y_0, u_0), y_2, y_1, u_1^N\right)$ and using Lemma B.1, we can write that

$$\begin{aligned}
E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_0, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z}_0)|] &\le \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[|\hat{c}(\pi_0^-, Y_1, Y_0, U_0, U_1) - \hat{c}(\pi^*, Y_1, Y_0, U_0, U_1)|] \\
&+ \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}\left[\beta \sum_{y_2\in\mathbb{Y}} J_\beta(\hat{z}_1, \tilde{\phi}^N) P^{\pi_0^-}(y_2 | Y_1, Y_0, U_1, U_0) - \beta \sum_{y_2\in\mathbb{Y}} \tilde{J}_\beta^N(\hat{z}_1) P^{\pi^*}(y_2 | Y_1, Y_0, U_1, U_0)\right] \\
&\le \|c\|_\infty \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[\|P^{\pi^*}(X_1 \in \cdot | Y_1, Y_0, U_0) - P^{\pi_0^-}(X_1 \in \cdot | Y_1, Y_0, U_0)\|_{TV}] \\
&+ \beta \|\tilde{J}_\beta^N\|_\infty \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[\|P^{\pi_0^-}(y_2 | Y_1, Y_0, U_1, U_0) - P^{\pi^*}(y_2 | Y_1, Y_0, U_1, U_0)\|_{TV}] \\
&+ \beta \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z}_1)|] \\
&\le \|c\|_\infty L_0 + \beta \|\tilde{J}_\beta^N\|_\infty L_0 + \beta \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z}_1)|].
\end{aligned}$$

Following the same steps for $E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_1, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z}_1)|]$ and repeating the same procedure with $\|\tilde{J}_\beta^N\|_\infty \le \frac{\|c\|_\infty}{1-\beta}$, one can conclude that

$$E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_0, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z}_0)|] \le \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t. \tag{C.1}$$

Now, we go back to the theorem statement to write

$$\begin{aligned}
E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}_0, \tilde{\phi}^N) - J_\beta^*(\hat{z}_0)|] &\le E_{\pi_0^-}^{\hat{\gamma}}[|J_\beta(\hat{z}, \tilde{\phi}^N) - \tilde{J}_\beta^N(\hat{z})|] + E_{\pi_0^-}^{\hat{\gamma}}[|\tilde{J}_\beta^N(\hat{z}) - J_\beta^*(\hat{z})|] \\
&\le \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t.
\end{aligned}$$

The last step follows from (C.1) and Theorem 2. □

References

[1] Billingsley P (1999) *Convergence of Probability Measures*, 2nd ed. (Wiley, New York).
[2] Dobrushin R (1956) Central limit theorem for nonstationary Markov chains. I. *Theory Probab. Appl.* 1(1):65–80.
[3] Even-Dar E, Mansour Y, Bartlett P (2003) Learning rates for Q-learning. *J. Machine Learn. Res.* 5(1):1–25.
[4] Feinberg E (1982) Controlled Markov processes with arbitrary numerical criteria. *Theory Probab. Appl.* 27(3):486–503.
[5] Feinberg E, Kasyanov P, Zadioanchuk N (2012) Average cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.* 37(4):591–607.
[6] Feinberg E, Kasyanov P, Zgurovsky M (2016) Partially observable total-cost Markov decision process with weakly continuous transition probabilities. *Math. Oper. Res.* 41(2):656–681.
[7] Golowich N, Moitra A, Rohatgi D (2022) Planning in observable POMDPs in quasipolynomial time. Preprint, submitted June 7, https://arxiv.org/abs/2201.04735.
[8] Hansen EA (2013) Solving POMDPs by searching in policy space. Preprint, submitted January 30, https://arxiv.org/abs/1301.7380.
[9] Hernández-Lerma O (1989) *Adaptive Markov Control Processes* (Springer-Verlag, New York).
[10] Hernández-Lerma O, Lasserre JB (1996) *Discrete-Time Markov Control Processes: Basic Optimality Criteria* (Springer, New York).
[11] Jaakkola T, Jordan M, Singh S (1994) On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.* 6(6):1185–1201.
[12] Jaakkola T, Singh S, Jordan M (1995) Reinforcement learning algorithm for partially observable Markov decision problems. *Adv. Neural Inform. Processing Systems* 7:345–352.
[13] Kac M (1947) On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.* 53(10):1002–1010.
[14] Kara A, Yuksel S (2022) Near optimality of finite memory feedback policies in partially observed Markov decision processes. *J. Machine Learn. Res.* 23(1):1–46.
[15] Kara A, Saldi N, Yüksel S (2019) Weak Feller property of non-linear filters. *Systems Control Lett.* 134:104512.
[16] Krishnamurthy V (2016) *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing* (Cambridge University Press, Cambridge, UK).

[17] Li G, Wei Y, Chi Y, Gu Y, Chen Y (2020) Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. Preprint, submitted June 4, https://arxiv.org/abs/2006.03041.
[18] Lin LJ, Mitchell TM (1992) Memory approaches to reinforcement learning in non-Markovian domains. Report, Carnegie Mellon University, Pittsburgh.
[19] Lovejoy W (1991) A survey of algorithmic methods for partially observed Markov decision processes. *Ann. Oper. Res.* 28:47–66.
[20] McCallum A (1997) Reinforcement learning with selective perception and hidden state. Unpublished doctoral dissertation, University of Rochester, NY.
[21] McDonald C, Yüksel S (2018) Converse results on filter stability criteria and stochastic non-linear observability. Preprint, submitted December 5, https://arxiv.org/abs/1812.01772.
[22] McDonald C, Yüksel S (2019) Observability and filter stability for partially observed Markov processes. *IEEE 58th Conf. Decision and Control* (IEEE, Piscataway, NJ), 1623–1628.
[23] McDonald C, Yüksel S (2020) Exponential filter stability via Dobrushin's coefficient. *Electronic Comm. Probab.* 25.
[24] McDonald C, Yüksel S (2022) Robustness to incorrect priors and controlled filter stability in partially observed stochastic control. *SIAM J. Control Optim.* 60(2).
[25] Parthasarathy K (1967) *Probability Measures on Metric Spaces* (AMS Bookstore, Providence, RI).
[26] Pineau J, Gordon G, Thrun S (2006) Anytime point-based approximations for large POMDPs. *J. Artificial Intelligence Res.* 27(1):335–380.
[27] Porta JM, Vlassis N, Spaan MTJ, Poupart P (2006) Point-based value iteration for continuous POMDPs. *J. Machine Learn. Res.* 7:2329–2367.
[28] Rhenius D (1974) Incomplete information in Markovian decision models. *Ann. Statist.* 2(6):1327–1334.
[29] Saldi N, Yüksel S, Linder T (2017) On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *Math. Oper. Res.* 42(4):945–978.
[30] Saldi N, Yüksel S, Linder T (2020) Finite model approximations for partially observed Markov decision processes with discounted cost. *IEEE Trans. Automatic Control* 65(1):130–142.
[31] Singh S, Jaakkola T, Jordan M (1994) Learning without state-estimation in partially observable Markovian decision processes. *Machine Learn. Proc.* (Elsevier, Amsterdam), 284–292.
[32] Smith T, Simmons R (2012) Point-based POMDP algorithms: Improved analysis and implementation. Preprint, submitted July 4, https://arxiv.org/abs/1207.1412.
[33] Subramanian J, Mahajan A (2019) Approximate information state for partially observed systems. *IEEE 58th Conf. Decision Control*, 1629–1636.
[34] Tsitsiklis JN (1994) Asynchronous stochastic approximation and Q-learning. *Machine Learn.* 16:185–202.
[35] Villani C (2008) *Optimal Transport: Old and New* (Springer, New York).
[36] Vlassis N, Spaan MTJ (2005) Perseus: Randomized point-based value iteration for POMDPs. *J. Artificial Intelligence Res.* 24(1):195–220.
[37] Wainwright M (2019) Stochastic approximation with cone-contractive operators: Sharper l_∞-bounds for Q-learning. Preprint, submitted May 15, https://arxiv.org/abs/1905.06265.
[38] White CC III (1991) A survey of solution techniques for the partially observed Markov decision process. *Ann. Oper. Res.* 32:215–230.
[39] White CC III, Scherer WT (1994) Finite-memory suboptimal design for partially observed Markov decision processes. *Oper. Res.* 42(3):439–455.
[40] Yu H, Bertsekas DP (2008) On near optimality of the set of finite-state controllers for average cost POMDP. *Math. Oper. Res.* 33(1):1–11.
[41] Yushkevich A (1976) 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 Probab. Appl.* 21(1):153–158.
[42] Zhou E, Fu MC, Marcus SI (2008) A density projection approach to dimension reduction for continuous-state POMDPs. *47th IEEE Conf. Decision Control*, 5576–5581.
[43] Zhou E, Fu MC, Marcus SI (2010) Solving continuous-state POMDPs via density projection. *IEEE Trans. Automatic Control* 55(5):1101–1116.
[44] Zhou R, Hansen E (2001) An improved grid-based approximation algorithm for POMDPs. *Internat. J. Conf. Artificial Intelligence*, 707–714.