2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

2024 IEEE 63rd Conference on Decision and Control (CDC) | 979-8-3503-1633-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/CDC56724.2024.10886324

Partially Observed Optimal Stochastic Control: Regularity, Optimality, Approximations, and Learning

Ali Devran Kara[1] and Serdar Yüksel[2]

Abstract— **In this review/tutorial article, we present recent progress on optimal control of partially observed Markov Decision Processes (POMDPs). We first present regularity and continuity conditions for POMDPs and their belief-MDP reductions, where these constitute weak Feller and Wasserstein regularity and controlled filter stability. These are then utilized to arrive at existence results on optimal policies for both discounted and average cost problems, and regularity of value functions. Then, we study rigorous approximation results involving quantization based finite model approximations as well as finite window approximations under controlled filter stability. Finally, we present several recent reinforcement learning theoretic results which rigorously establish convergence to near optimality under both criteria.**

I. Partially Observed Markov Decision Processes: Introduction and Preliminaries

Partially observed Markov Decision processes (POMDPs) present challenging mathematical problems with significant applied relevance.

Consider a stochastic process $\{X_k, k \in \mathbb{Z}_+\}$, where each element X_k takes values in some standard Borel space $\mathbb{X}$, with dynamics described by

$$X_{k+1} = F(X_k, U_k, W_k) \quad (1)$$
$$Y_k = G(X_k, V_k) \quad (2)$$

where Y_k is an $\mathbb{Y}$-valued measurement sequence; we take $\mathbb{Y}$ also to be some standard Borel space. Suppose further that $X_0 \sim \mu$ for some $\mu \in \mathcal{P}(\mathbb{X})$, where $\mathcal{P}(\mathbb{X})$ represents the set of all probability measures on $\mathbb{X}$. Here, W_k, V_k are mutually independent i.i.d. noise processes. This system is subjected to a control/decision process where the control/decision at time n, U_k, incurs a cost $c(X_k, U_k)$. The decision maker only has access to the measurement process Y_k and U_k causally: An *admissible policy* γ is a sequence of control/decision functions $\{\gamma_k,\, k \in \mathbb{Z}_+\}$ such that γ_k is measurable with respect to the σ-algebra generated by the information variables

$$I_k = \{Y_{[0,k]}, U_{[0,k-1]}\}, \quad k \in \mathbb{N}, \qquad I_0 = \{Y_0\}.$$

so that

$$U_k = \gamma_k(I_k), \quad k \in \mathbb{Z}_+ \quad (3)$$

are the $\mathbb{U}$-valued control/decision actions and we use the notation

$$Y_{[0,k]} = \{Y_s,\, 0 \leq s \leq k\}, \quad U_{[0,k-1]} = \{U_s,\, 0 \leq s \leq k-1\}.$$

We define Γ to be the set of all such (strong-sense) admissible policies. We emphasize the implicit assumption that the control policy also depends on the prior probability measure μ.

We assume that all of the random variables are defined on a common probability space $(\Omega, \mathcal{F}, P)$ given the initial distribution on the state, and a policy, on the infinite product space consistent with finite dimensional distributions, by the Ionescu Tulcea Theorem [30]. We will sometimes write the probability measure on this space as P_μ^γ to emphasize the policy γ and the initialization μ. We note that (1)-(2) can also, equivalently (via stochastic realization results [27, Lemma 1.2] [5, Lemma 3.1], [1, Lemma F]), be represented with transition kernels: the state transition kernel is denoted with $\mathcal{T}$ so that for Borel $B \subset \mathbb{X}$

$$\mathcal{T}(B|x,u) := P(X_1 \in B | X_0 = x, U_0 = u), \quad .$$

We will denote the measurement kernel with Q so that for Borel $B \subset \mathbb{Y}$:

$$Q(B|x) := P(Y_0 \in B | X_0 = x).$$

For (1)-(2), we are interested in minimizing either the average-cost optimization criterion

$$J_\infty(\mu, \gamma) := \limsup_{N \to \infty} \frac{1}{N} E_\mu^\gamma[\sum_{k=0}^{N-1} c(X_k, U_k)] \quad (4)$$

or the discounted cost criterion (for some $\beta \in (0,1)$

$$J_\beta(\mu, \gamma) := E_\mu^\gamma[\sum_{k=0}^{\infty} \beta^k c(X_k, U_k)] \quad (5)$$

over all admissible control policies $\gamma = \{\gamma_0, \gamma_1, \cdots,\} \in \Gamma$ with $X_0 \sim \mu$. With $\mathcal{P}(\mathbb{U})$ denoting the set of probability measures on $\mathbb{U}$ endowed with the weak convergence topology, we will also, when needed, allow for independent randomizations so that $\gamma_k(I_k)$ is $\mathcal{P}(\mathbb{U})$-valued for each realization of I_k. Here $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}_+$ is the cost function.

One may also consider the control-free case where the system equation (1) does not have control dependence; in this case only a decision is to be made at every time stage; U is present only in the cost expression in (4). This important special case has been studied extensively in the theory of non-linear filtering.

A. Literature review and preliminaries

In the following, we present a brief literature review on optimal control of POMDPs, before presenting the main results of the article.

[1] Department of Mathematics at Florida State University akara@fsu.edu

[2] Department of Mathematics and Statistics at Queen's University, Kingston ON, Canada. yuksel@queensu.ca. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

979-8-3503-1633-9/24/$31.00 ©2024 IEEE

POMDPs, separated policies and belief-MDPs. It is well-known that any POMDP can be reduced to a (completely observable) MDP [76], [52], whose states are the posterior state probabilities, or beliefs, of the observer; that is, the state at time k is

$$\pi_k(\,\cdot\,) := P\{X_k \in \,\cdot\,|Y_0, \ldots, Y_k, U_0, \ldots, U_{k-1}\} \in \mathcal{P}(\mathbb{X}).$$

We call this equivalent MDP the belief-MDP. The belief-MDP has state space $\mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Here, $\mathcal{P}(\mathbb{X})$ is equipped with the Borel σ-algebra generated by the topology of weak convergence [3]. Since $\mathbb{X}$ is a Borel space, $\mathcal{P}(\mathbb{X})$ is metrizable with the Prokhorov metric which makes $\mathcal{P}(\mathbb{X})$ into a Borel space [50]. The transition probability η of the belief-MDP can be constructed as follows (see also [29]). If we define the measurable function

$$F(\pi, a, y) := Pr\{X_{k+1} \in \,\cdot\,|\pi_k = \pi, U_k = u, Y_{k+1} = y\}$$

from $\mathcal{P}(\mathbb{X}) \times \mathbb{U} \times \mathbb{Y}$ to $\mathcal{P}(\mathbb{X})$ and the stochastic kernel $H(\,\cdot\,|\pi, u) := Pr\{Y_{k+1} \in \,\cdot\,|\pi_k = \pi, U_k = u\}$ on $\mathbb{Y}$ given $\mathcal{P}(\mathbb{X}) \times \mathbb{U}$, then η can be written as

$$\eta(\,\cdot\,|\pi, u) = \int_{\mathbb{Y}} 1_{\{F(\pi,u,y)\in\,\cdot\,\}} H(dy|\pi, u). \qquad (6)$$

The one-stage cost function c of the belief-MDP is given by

$$\tilde{c}(\pi, u) := \int_{\mathbb{X}} c(x, u)\pi(dx). \qquad (7)$$

With cost function $c(x, u)$ is continuous and bounded on $\mathbb{X} \times \mathbb{U}$, an application of the generalized dominated convergence theorem [45, Theorem 3.5] [58, Theorem 3.5], we have that $\tilde{c}(\pi, u) = E^{\pi}[c(x, u)] := \int \pi(dx)c(x, u) : \mathcal{P}(\mathbb{X}) \times \mathbb{U} \to \mathbb{R}$ is also continuous and bounded, and thus Borel measurable.

In particular, the belief-MDP is a (fully observed) Markov decision process with the components $(\mathcal{P}(\mathbb{X}), \mathbb{U}, \eta, \tilde{c})$.

For finite horizon problems and a large class of infinite horizon discounted cost problems, it is a standard result that an optimal control policy will use the belief π_k as a sufficient statistic for optimal policies (see [76], [52], [4]).

Approximations and Learning for POMDPs. Studies on POMDPs had primarily been algorithmic and numerical, with rigorous studies applicable to a particular set of problems until recently. In particular, the regularity properties of POMDPs as pioneered in [13], [23] and later generalized to further conditions and criteria in [37], [24], [22], [41], [16] have paved the way for rigorous and explicit performance bounds. For approximate optimality on POMDPs, we refer the reader to the detailed review in [57] for discounted cost problems and [14] for the average cost problem.

If the agent does not know the underlying dynamics of the observations (transitions and/or channel), then the learning of the solutions to the optimal control problem from observed data is necessary. However, learning for POMDPs has been a challenging problem. Various approaches and studies are available in the literature starting with [60], see e.g. [21], [32], [70], [44], [2] for some of the learning approaches for POMDPs. In particular, we also cite the recent comprehensive studies [10] and [19] which study learning in non-Markov environments. [63], [59] present a general framework on approximation states and their induced optimality and near optimality properties under several uniformity bounds. Some of our explicit analysis here can also be seen in view of these bounds. We also refer the reader to the second tutorial paper [61] for complementary approaches to the planning and learning problem.

Our technical approach on learning builds on the mathematical analysis developed in [42] and generalized in [43] (see also [38], [41]). We note that more general, non-uniform, bounds are also considered in [38], [41].

Control-free setup. For the special case without control, the belief process is known as the (non-linear) filter process, and by the discussion above, this itself is a Markov process. For our paper, this setup will be useful to study the convergence and uniqueness properties involving invariant probability measures for the filter process: The stability properties of such processes has been studied, where the existence of an invariant probability measure for the belief process, as well as the uniqueness of such a measure (i.e., the unique ergodicity property) has been investigated under various conditions, see. e.g. [8] and [11] which provide a comprehensive discussion on both the ergodicity of the filter process as well as filter stability. [46, Theorem 2] and [67, Prop 2.1] assume that the hidden state process is ergodic and the filter is stable (almost surely or in expectation under total variation); these papers crucially embed the stationary state in the joint process (x_k, π_k) and note that when x_k is stationary, the Markov chain defined by this process admits an invariant probability measure. See also [33], [26], [11], [34], [64] for further filter stability and unique ergodicity results and a recent review in [72].

B. Convergence Notions for Probability Measures

For the analysis of the technical results, we will use different convergence and distance notions for probability measures. Two important notions of convergences for sequences of probability measures are weak convergence, and convergence under total variation. For some $N \in \mathbb{N}$ 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 $\int_{\mathbb{X}} f(x)\mu_n(dx) \to \int_{\mathbb{X}} f(x)\mu(dx)$ 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, and metric (Polish) space endowed with the topology of weak convergence is itself complete, separable, and metric [50]. One such metric is the bounded Lipschitz metric [69, p.109], which is defined for $\mu, \nu \in \mathcal{P}(\mathbb{X})$ as

$$\rho_{BL}(\mu, \nu) := \sup_{\|f\|_{BL}\leq 1} |\int f d\mu - \int f d\nu| \qquad (8)$$

where

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

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

We next introduce the first order Wasserstein metric. The *Wasserstein metric* of order 1 for two measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ is defined as

$$W_1(\mu, \nu) = \inf_{\eta \in \mathcal{H}(\mu,\nu)} \int_{\mathbb{X}\times\mathbb{X}} \eta(dx, dy)|x-y|,$$

where $\mathcal{H}(\mu, \nu)$ denotes the set of probability measures on $\mathbb{X} \times \mathbb{X}$ with first marginal μ and second marginal ν. Furthermore, using the dual representation of the first order Wasserstein metric, we equivalently have

$$W_1(\mu, \nu) = \sup_{Lip(f) \leq 1} \left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right|$$

where $Lip(f)$ is the minimal Lipschitz constant of f.

A sequence $\{\mu_n\}$ is said to converge in W_1 to $\mu \in \mathcal{P}(\mathbb{X})$ if $W_1(\mu_n, \mu) \to 0$. For compact $\mathbb{X}$, the Wasserstein distance of order 1 metrizes the weak topology on the set of probability measures on $\mathbb{X}$ (see [69, Theorem 6.9]). For non-compact $\mathbb{X}$ convergence in the W_1 metric implies weak convergence (in particular this metric bounds from above the Bounded-Lipschitz metric [69, p.109], which metrizes the weak convergence).

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

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

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 $\|\mu_n - \mu\|_{TV} \to 0$.

II. Regularity Results: Weak Continuity, Wasserstein Continuity, Wasserstein Contraction and Filter Stability

A. Weak Feller Continuity of the Belief-MDP

Building on [37] and [23], this section establishes the weak Feller property of the filter process; that is, the weak Feller property of the kernel defined in (6) under two different sets of assumptions.

Assumption 2.1: (i) The transition probability $\mathcal{T}(\cdot|x, u)$ is weakly continuous in (x, u), i.e., 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 $Q(\cdot|x, u)$ is continuous in total variation, i.e., for any $(x_n, u_n) \to (x, u)$, $Q(\cdot|x_n, u_n) \to Q(\cdot|x, u)$ in total variation.

Assumption 2.2: (i) The transition probability $\mathcal{T}(\cdot|x, u)$ is continuous in total variation in (x, u), i.e., 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.

(ii) The observation channel $Q(\cdot|x)$ is independent of the control variable.

Theorem 2.1: [23] Under Assumption 2.1, the transition probability $\eta(\cdot|z, u)$ of the filter process is weakly continuous in (z, u).

Theorem 2.2: [37] Under Assumption 2.2, the transition probability $\eta(\cdot|z, u)$ of the filter process is weakly continuous in (z, u).

As examples, taken from [37], suppose that the system dynamics and the observation channel are represented as follows:

$$\begin{aligned} x_{t+1} &= H(x_t, u_t, w_t), \\ y_t &= G(x_t, u_{t-1}, v_t), \end{aligned}$$

where w_t and v_t are i.i.d. noise processes.

(i) Suppose that $H(x, u, w)$ is a continuous function in x and u. Then, the corresponding transition kernel is weakly continuous.

(ii) Suppose that $G(x, u, v) = g(x, u) + v$, where g is a continuous function and V_t admits a continuous density function φ with respect to some reference measure ν. Then, the channel is continuous in total variation.

(iii) Suppose that we have $H(x, u, w) = h(x, u) + w$, where f is continuous and w_t admits a continuous density function φ with respect to some reference measure ν. Then, the transition probability is continuous in total variation.

B. Wasserstein Continuity and Contraction Properties of the Belief-MDP

Recently, [41] presented the following regularity results for controlled filter processes. Let us first recall the following:

Definition 2.1: [18, Equation 1.16][Dobrushin coefficient] 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:

$$\delta(K) = \inf \sum_{i=1}^{n} \min(K(x, A_i), K(y, A_i)) \tag{9}$$

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

Assumption 2.3:

1) $(\mathbb{X}, d)$ is a bounded compact metric space with diameter D (where $D = \sup_{x,y\in\mathbb{X}} d(x, y)$).
2) The transition probability $\mathcal{T}(\cdot \mid x, u)$ is continuous in total variation in (x, u), i.e., for any $(x_n, u_n) \to (x, u), \mathcal{T}(\cdot \mid x_n, u_n) \to \mathcal{T}(\cdot \mid x, u)$ in total variation.
3) There exists $\alpha \in R^+$such that
$$\|\mathcal{T}(\cdot \mid x, u) - \mathcal{T}(\cdot \mid x', u)\|_{TV} \leq \alpha d(x, x')$$
for every $x, x' \in \mathbb{X}$, $u \in \mathbb{U}$.
4) There exists $K_1 \in \mathbb{R}^+$ such that
$$|c(x, u) - c(x', u)| \leq K_1 d(x, x').$$
for every $x, x' \in \mathbb{X}$, $u \in \mathbb{U}$.
5) The cost function c is bounded and continuous.

Theorem 2.3: [14] Assume that $\mathbb{X}$ and $\mathbb{Y}$ are Polish spaces. If Assumptions 2.3-1,3 are fulfilled, then we have

$$W_1(\eta(\cdot \mid z_0, u), \eta(\cdot \mid z_0', u)) \leq K_2 W_1(z_0, z_0'),$$

with $K_2 := \frac{\alpha D(3-2\delta(Q))}{2}$ for all $z_0, z_0' \in \mathcal{P}(\mathbb{X})$, $u \in \mathbb{U}$.

Remark 2.1: A recent paper [72] has presented an alternative approach, without belief-separation, and has arrived further conditions for the existence of optimal policies for discounted and average cost problems as well as the unique ergodicity property for both controlled and control-free setups. Such an approach leads to complementary conditions on the weak Feller property on the state, which considers the entire past as the state endowed with the product topology.

C. Filter Stability

The filter stability problem refers to the correction of an incorrectly initialized non-linear filter for a partially observed stochastic dynamical system (controlled or control-free) with increasing measurements. As we will see, this property has significant implications on robustness as well as near optimality of sliding finite window policies with explicit approximation bounds, to be presented in the paper.

Let us describe this property more explicitly: Given a prior $\mu \in \mathcal{P}(\mathbb{X})$ and a policy $\gamma \in \Gamma$ we can define the filter and predictor for a POMDP using the (strategic) measure $P^{\mu,\gamma}$.

Definition 2.2: (i) We define the one step predictor process as the sequence of conditional probability measures

$$\pi_{n-}^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_n \in \cdot | Y_{[0,n-1]}, U_{[0,n-1]}) \qquad n \in \mathbb{N}$$

(ii) We define the filter process as the sequence of conditional probability measures

$$\pi_k^{\mu,\gamma}(\cdot) = P^{\mu,\gamma}(X_k \in \cdot | Y_{[0,k]}, U_{[0,k-1]}), \quad n \in \mathbb{Z}_+ \tag{10}$$

Remark 2.2: Recall that the $U_{[0,k-1]}$ are all functions of the $Y_{[0,k-1]}$, so conditioning on the control actions is not necessary in the above definitions. Yet this conditional probability would be *policy dependent*; if we condition on the past actions, this conditioning would be *policy-independent*. Say a prior $\mu \in \mathcal{P}(\mathbb{X})$ and a policy $\gamma \in \Gamma$ are chosen, an observer sees measurements $Y_{[0,\infty)}$ generated via the strategic measure $P^{\mu,\gamma}$. The observer is aware that the policy applied is γ, but incorrectly thinks the prior is $\nu \neq \mu$. The observer will then compute the incorrectly initialized filter $\pi_k^{\nu,\gamma}$ while the true filter is $\pi_k^{\mu,\gamma}$. The filter stability problem is concerned with the merging of $\pi_k^{\nu,\gamma}$ and $\pi_k^{\mu,\gamma}$ as k goes to infinity.

In the literature, there are a number of merging notions when one considers stability which we enumerate here. Let $C_b(\mathbb{X})$ represent the set of continuous and bounded functions from $\mathbb{X} \to \mathbb{R}$. We define here the different notions of stability for the filter.

Definition 2.3: (i) A filter process is said to be stable in the sense of weak merging with respect to a policy γ, $P^{\mu,\gamma}$ almost surely (a.s.) if there exists a set of measurement sequences $A \subset \mathcal{Y}^{\mathbb{Z}_+}$ with $P^{\mu,\gamma}$ probability 1 such that for any sequence in A; for any $f \in C_b(\mathcal{X})$ and any prior ν with $\mu \ll \nu$ (i.e., for all Borel B $\nu(B) = 0 \implies \mu(B) = 0$) we have $\lim_{n\to\infty} \left| \int f d\pi_n^{\mu,\gamma} - \int f d\pi_n^{\nu,\gamma} \right| = 0$.

(ii) A filter process is said to be stable in the sense of total variation in expectation with respect to a policy γ if for any measure ν with $\mu \ll \nu$ we have $\lim_{n\to\infty} E^{\mu,\gamma}[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}] = 0$.

(iii) A filter process is said to be stable in the sense of total variation with respect to a policy γ, $P^{\mu,\gamma}$ a.s. if there exists a set of measurement sequences $A \subset \mathcal{Y}^{\mathbb{Z}_+}$ with $P^{\mu,\gamma}$ probability 1 such that for any sequence in A; for any measure ν with $\mu \ll \nu$ we have $\lim_{n\to\infty} \|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV} = 0 \;\; P^{\mu,\gamma}\ a.s.$.

(iv) The filter is said to be *universally* stable in one of the above notions if the notion holds with respect to every admissible policy $\gamma \in \Gamma$.

One of the main differences between control-free and controlled partially observed Markov chains is that the filter is always Markovian under the former, whereas under a controlled model the filter process may not be Markovian since the control policy may depend on past measurements in an arbitrary (measurable) fashion. This complicates the dependency structure, and therefore results from the control-free case do not directly apply to the controlled setup.

Recall (9) and let us define $\tilde{\delta}(\mathcal{T}) := \inf_{u\in\mathbb{U}} \delta(\mathcal{T}(\cdot|\cdot,u))$.

Theorem 2.4: [47, Theorem 3.3] Assume that for $\mu, \nu \in \mathcal{P}(\mathbb{X})$, we have $\mu \ll \nu$. Then we have

$$E^{\mu,\gamma}\left[\|\pi_{n+1}^{\mu,\gamma} - \pi_{n+1}^{\nu,\gamma}\|_{TV}\right] \leq \alpha E^{\mu,\gamma}\left[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}\right]. \tag{11}$$

where $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q))$.

If $\alpha < 1$, by applying the Borel-Cantelli lemma and Markov's inequality, we have that exponential stability in expectation implies the same result in an almost sure sense as well; see [47, Remark 3.10]. This also establishes that the rate of convergence is uniform over all priors ν as long as $\mu \ll \nu$.

A further method, and one which leads to complementary conditions given the above, for filter stability is via the Hilbert projective metric.

Definition 2.4: Two non-negative measures μ, ν on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ are comparable, if there exist positive constants $0 < a \leq b$, such that

$$a\nu(A) \leq \mu(A) \leq b\nu(A)$$

for any Borel subset $A \subset \mathbb{X}$.

Definition 2.5 (Mixing kernel): The non-negative kernel K defined on $\mathbb{X}$ is mixing, if there exists a constant $0 < \varepsilon \leq 1$, and a non-negative measure λ on $\mathbb{X}$, such that

$$\varepsilon\lambda(A) \leq K(x, A) \leq \frac{1}{\varepsilon}\lambda(A)$$

for any $x \in \mathbb{X}$, and any Borel subset $A \subset \mathbb{X}$.

Definition 2.6: (Hilbert metric). Let μ, ν be two non-negative finite measures. We define the Hilbert metric on such measures as

$$h(\mu, \nu) = \begin{cases} \log\left(\dfrac{\sup_{A|\nu(A)>0} \frac{\mu(A)}{\nu(A)}}{\inf_{A|\nu(A)>0} \frac{\mu(A)}{\nu(A)}}\right) & \text{if } \mu, \nu \text{ are comparable} \\ 0 & \text{if } \mu = \nu = 0 \\ \infty & \text{else} \end{cases} \tag{12}$$

Note that $h(a\mu, b\nu) = h(\mu, \nu)$ for any positive scalars a, b. Therefore, the Hilbert metric is a useful metric for nonlinear filters since it is invariant under normalization, and the following lemma demonstrates that it bounds the total-variation distance.

Lemma 2.1: [28, Lemma 3.4] Let μ, ν be two non-negative finite measures,

i. $\|\mu - \nu\|_{TV} \leq \frac{2}{\log 3} h(\mu, \nu)$.

ii. If the nonnegative kernel K is a mixing kernel (see Definition 2.5) with constant ϵ, then $h(K\mu, K\nu) \leq \frac{1}{\varepsilon^2} \|\mu - \nu\|_{TV}$.

Lemma 2.2 ([28], Lemma 3.8): (Birkhoff contraction coefficient). The nonnegative linear operator τ on $\mathcal{M}^+(\mathbb{X})$ (positive measures on $\mathbb{X}$) associated with a nonnegative kernel K defined on $\mathbb{X}$

$$\tau(K) := \sup_{0 < h(\mu,\nu) < \infty} \frac{h(K\mu, K\nu)}{h(\mu, \nu)} = \tanh\left[\frac{1}{4} H(K)\right]$$

where

$$H(K) := \sup_{\mu,\nu} h(K\mu, K\nu)$$

is over nonnegative measures, is a contraction (called the Birkhoff contraction coefficient) , is a contraction under the Hilbert metric if $H(K) < \infty$ (which implies $\tau(K) < 1$).

Another filter stability result which will also be useful in numerical methods for POMDPs to be considered later is via the following *stochastic non-linear observability* definition.

Definition 2.7: [Stochastic Observability for Non-Linear Systems][48] A POMDP is called one step observable (universal in admissible control policies) if for every $f \in C_b(\mathbb{X})$ and every $\epsilon > 0$ there exists a measurable and bounded function g such that $\|f(\cdot) - \int_{\mathbb{Y}} g(y) Q(dy|\cdot)\|_\infty < \epsilon$.

Theorem 2.5: [48] Assume that $\mu \ll \nu$ and that the POMDP is one step observable. Then the predictor is universally stable weakly a.s. .

The observability notion defined above only results in stability of the predictor in the weak sense $P^{\mu,\gamma}$ almost surely. However, these can be extended to filter stability and under further criteria, see [48], [49].

We now present an example for observability.

Example 2.1: [49] Consider a finite setup $\mathbb{X} = \{a_1, \cdots, a_n\}$ and let the noise space be $\mathbb{V} = \{b_1, \cdots, b_m\}$. Now, assume $y = h(x, v)$ has K distinct outputs, where $1 \leq K \leq (n)(m)$ and $\mathbb{Y} = \{c_1, \cdots, c_K\}$. We note that for such a setup, there is already a sufficient and necessary condition for filter stability provided in [68, Theorem V.2] (see also [66]). For each x, h_x can be viewed as a partition of $\mathbb{V}$, assigning each $b_i \in \mathbb{X}$ to an output level $c_j \in \mathbb{Y}$. We can track this by the matrix $H_x(i, j) = 1$ if $h_x(b_i) = y_j$ and zero else. Let Q be the $1 \times m$ vector representing the probability measure of the noise. Let $g(c_i) = \alpha_i$, with $\alpha^\intercal = [\alpha_1, \alpha_2, \ldots, \alpha_K]$ and $\int_{\mathbb{V}} g(h(x, v)) Q(dv) =: (QH_x)\alpha$. Any function $f(x)$ can be expressed as a $n \times 1$ vector and hence the question reduces to finding a vector α so that $f = QH\alpha$, and the system is one step observable if and only if the matrix $A \equiv \begin{bmatrix} QH_{a_1} \\ \vdots \\ QH_{a_n} \end{bmatrix}$ is rank n.

Further examples for measurement channels satisfying Definition 2.7 have been reported in [49, Section 3].

Applications of these will be discussed in the context of numerical methods for POMDPs. Filter stability is also related to robustness of optimal costs to incorrect initializations for controlled models [48].

III. Existence of Optimal Policies: Discounted Cost and Average Cost

A. Discounted Cost

Theorem 3.1: If the cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is continuous and bounded, and $\mathbb{U}$ is compact, under Theorems 2.1 or 2.2, for any $\beta \in (0, 1)$, there exists an optimal solution to the discounted cost optimality problem with a continuous and bounded value function. Furthermore, under Assumption 2.3, with $K_2 = \frac{\alpha D(3 - 2\delta(Q))}{2}$, if $\beta K_2 < 1$ the value function is Lipschitz continuous.

Proof: An application of the dominated convergence theorem implies that $\tilde{c}(\pi, u)$ (7) is also continuous and bounded. If the action set is compact, then under Theorems 2.1 or 2.2, which imply that η is weakly continuous, we have that the measurable selection conditions (see e.g. [30]) apply, and solutions to the Bellman or discounted cost optimality equations exist, and accordingly an optimal control policy exists. For the second result, [54, Theorem 4.37] leads to Lipschitz regularity under the Wasserstein continuity condition on the kernel. ■

B. Average Cost

The average cost is a significantly more challenging problem as the typical contraction conditions via minorization is too demanding for η. An alternative approach is based on the Section II-B. The average cost optimality equation (ACOE) plays a crucial role for the analysis and the existence results of MDPs under the infinite horizon average cost optimality criteria. The triplet (h, ρ^*, γ^*), where $h, \gamma : \mathcal{P}(\mathbb{X}) \to \mathbb{R}$ are measurable functions and $\rho* \in \mathbb{R}$ is a constant, forms the ACOE if

$$\begin{aligned} h(z) + \rho^* &= \inf_{u \in \mathbb{U}} \left\{ \tilde{c}(z, u) + \int h(z_1) \eta(dz_1 | z, u) \right\} \\ &= \tilde{c}(z, \gamma^*(z)) + \int h(z_1) \eta(dz_1 | z, \gamma^*(z)) \end{aligned} \tag{13}$$

for all $z \in \mathcal{P}(\mathbb{X})$. It is well known that (see e.g. [30, Theorem 5.2.4]) if (13) is satisfied with the triplet (h, ρ^*, γ^*), and furthermore if h satisfies

$$\sup_{\gamma \in \Gamma} \lim_{t \to \infty} \frac{E_z^\gamma[h(Z_t)]}{t} = 0, \quad \forall z \in \mathcal{P}(\mathbb{X})$$

then γ^* is an optimal policy for the POMDP under the infinite horizon average cost optimality criteria, and

$$J^*(z) = \inf_{\gamma \in \Gamma} J(z, \gamma) = \rho^* \quad \forall z \in \mathcal{P}(\mathbb{X}).$$

Theorem 3.2: (i) [14] Under Assumption 2.3, with $K_2 = \frac{\alpha D(3 - 2\delta(Q))}{2} < 1$, a solution to the average cost optimality equation (ACOE) exists. This leads to the existence of an optimal control policy, and optimal cost is constant for every initial state.

(ii) [72, Theorem 3] If the cost function $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is continuous and bounded, and $\mathbb{U}$ is compact, under weak Feller regularity of η (e.g., under either Theorem 2.1 or 2.2), there exists an optimal policy [1]

Proof: (i) follows from a vanishing discount method [14]. (ii) follows from the convex analytic method building on [7]. ■

IV. Approximations: Discounted Cost

A. State and Action Space Quantization

By combining the approximation results in [57], [54], together with the weak Feller continuity results presented earlier, we can conclude that the numerical methods for weakly continuous fully observed MDPs can also be applied to POMDPs under the conditions reported in Theorems 2.1 and 2.2. This has explicitly been demonstrated in [57], where also methods for quantizing probability measures have been studied in [57, Section 5]. Notably, one can first quantize the action space with arbitrarily small loss (see [55][54, Theorem 3.16] for discounted cost and [55],[54, Theorem 3.22] for average cost) and then approximate the probability measures, e.g. under the W_1 metric, to obtain a finite model. In the following, we follow the approach and results from [38] applied to belief-MDPs.

To construct a finite near-optimal MDP model, we begin by quantizing the belief states. We select disjoint sets $\{Z_i\}_{i=1}^M$ such that $\bigcup_i Z_i = \mathcal{P}(\mathbb{X})$, and each Z_i is disjoint from Z_j for any $i \neq j$. For each set, we choose a representative state, denoted as $z_i \in Z_i$. This results in a finite state space for our model, represented by $\bar{Z} := \{z_1, \ldots, z_M\}$. The quantization function maps the original state space $\mathcal{P}(\mathbb{X})$ to this finite set $\bar{Z}$ as follows:

$$q(z) = z_i \quad \text{if } z \in Z_i.$$

To define the approximate cost function, we select a weight measure $\pi^* \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ over $\mathcal{P}(\mathbb{X})$ such that $\pi^*(Z_i) > 0$ for all Z_i. Under Assumption 2.3, we know that $\mathcal{P}(\mathbb{X})$ is compact under W_1 metric. We then define normalized measures for each quantization bin Z_i using the weight measure as:

$$\hat{\pi}^*_{z_i}(A) := \frac{\pi^*(A)}{\pi^*(Z_i)}, \quad \forall A \subset Z_i, \quad \forall i \in \{1, \ldots, M\}.$$

This normalized measure, $\hat{\pi}^*_{z_i}$, is specific to the set Z_i containing z_i.

Next, we define the stage-wise cost and the transition kernel for the MDP with the finite state space $\bar{Z}$ using these normalized weight measures. For any $z_i, z_j \in \bar{Z}$ and $u \in \mathbb{U}$, the stage-wise cost function and the transition kernel are:

$$c^*(z_i, u) = \int_{Z_i} \tilde{c}(z, u)\hat{\pi}^*_{z_i}(dz),$$

$$\eta^*(z_j \mid z_i, u) = \int_{Z_i} \eta(Z_j \mid z, u)\,\hat{\pi}^*_{z_i}(dz).$$

After establishing the finite state space $\bar{Z}$, the cost function c^* and the transition kernel η^*, we introduce the discounted optimal value function for this finite model, denoted as $\hat{J}_\beta : \bar{Z} \to \mathbb{R}$. We extend this function to the entire original state space $\mathcal{P}(\mathbb{X})$ by keeping it constant within the quantization bins. Therefore, for any $z \in Z_i$, we define:

$$\hat{J}_\beta(z) := \hat{J}_\beta(z_i).$$

We also define the maximum loss function among the quantization bins as:

$$\bar{L} := \max_{i=1,\ldots,M} \sup_{z,z' \in Z_i} W_1(z, z'). \tag{14}$$

Assumption 4.1: [[38] Assumption 4]

1) $\mathcal{P}(\mathbb{X})$ is compact (under W_1 metric).
2) There exists $\alpha_c > 0$ such that $|\tilde{c}(z,u) - \tilde{c}(z',u)| \leq \alpha_c d(z,z')$ for all $z, z' \in \mathcal{P}(\mathbb{X})$ and for all $u \in \mathbb{U}$. It suffices that $|c(x,u) - c(x',u)| \leq \alpha_c |x - x'|$.
3) There exists $\alpha_\eta > 0$ such that $W_1(\eta(\cdot \mid z,u), \eta(\cdot \mid z',u)) \leq \alpha_\eta d(z,z')$ for all $z, z' \in \mathcal{P}(\mathbb{X})$ and for all $u \in \mathbb{U}$.

Theorem 4.1: [38, Theorem 6] Under Assumption 4.1, we have

$$\sup_{z \in \mathcal{P}(\mathbb{X})} \left| J_\beta(z, \hat{\gamma}) - J^*_\beta(z) \right| \leq \frac{2\alpha_c}{(1-\beta)^2 (1 - \beta\alpha_\eta)} \bar{L},$$

where $\bar{L}$ is defined in (14) and $\hat{\gamma}$ denotes the optimal policy of the finite-state approximate model extended to the state space $\mathcal{P}(\mathbb{X})$ via the quantization function q.
A similar result is presented in the [54], Theorem 4.38, offering a slightly weaker bound.

Under Assumption 2.3, the belief MDP satisfies Assumption 4.1 because $\mathcal{P}(\mathbb{X})$ is compact under W_1 metric. It also follows that we have $|\tilde{c}(z,u) - \tilde{c}(z',u)| \leq K_1 W_1(z,z')$ for all $z, z' \in \mathcal{P}(\mathbb{X})$ and for all $u \in \mathbb{U}$. Theorem 2.3 implies $W_1(\eta(\cdot \mid z,u), \eta(\cdot \mid z',u)) \leq K_2 W_1(z,z')$ for all $z, z' \in \mathcal{P}(\mathbb{X})$ and for all $u \in \mathbb{U}$. Thus, for belief MDP, quantization provides the following bound:

$$\sup_{z \in \mathcal{P}(\mathbb{X})} \left| J_\beta(z, \hat{\gamma}) - J^*_\beta(z) \right| \leq \frac{2K_1}{(1-\beta)^2(1-\beta K_2)} \bar{L}.$$

Furthermore, quantized model gives near-optimal policy of the original belief MDP model as $\bar{L} \to 0$.

If we only have weak Feller regularity of η (e.g., under either Theorem 2.1 or 2.2), a similar approximation result holds though only by asymptotic convergence of the approximation error to zero as the diameters of the quantization bins converge to zero; see [57, Theorem 3] as an application of [56] and [54, Theorem 4.27].

In the following, an alternative approach is presented.

B. An Alternative Finite Window Belief-MDP Reduction and Its Approximation

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

[1] Here, the optimality result may only hold for a restrictive class of initial conditions or initializations, unlike part (i).

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.

Inspired from filter stability, consider the following: For any time step $t \geq N$ and for a fixed observation realization sequence $y_{[0,t]}$ and control action sequence $u_{[0,t-1]}$, the state process can be viewed as

$$P^{\mu}(x_t \in \cdot | y_{[0,t]}, u_{[0,t-1]}) = P^{\pi_{t-N_-}}(X_t \in \cdot | y_{[t-N,t]}, u_{[t-N,t-1]})$$

where

$$\pi_{t-N_-}(\cdot) = P^{\mu}(x_{t-N} \in \cdot | y_{[0,t-N-1]}, u_{[0,t-N-1]}).$$

That is, we can view the state as the Bayesian update of π_{t-N_-}, the predictor at time $t-N$, using the observations $y_{t-N}, \ldots, y_t$. Notice that with this representation only the most recent N observation realizations are used for the update and the past information of the observations is embedded in π_{t-N_-}.

We define the new state variable as the triple $(\pi^-_{t-N}, y_{[0,t-N-1]}, u_{[0,t-N-1]})$. We place the product metric on this new space: weak convergence on the belief and usual metric on the measurements and actions.

The idea is to quantize the new state as follows: collapse all π to a fixed state $\hat{\pi}$, define an approximate finite MDP and establish performance bounds utilizing filter stability.

In the following, we will assume that $\mathbb{X}$ is $\mathbb{R}^n$ for some n and that $\mathbb{U}, \mathbb{Y}$ are finite sets.

Define the quantization map F, such that for $(\pi, y_{[0,N]}, u_{[0,N-1]})$

$$F(\pi, y_{[0,N]}, u_{[0,N-1]}) = (\hat{\pi}, y_{[0,N]}, u_{[0,N-1]}).$$

Using the map F and the finite set $\mathcal{Z}^N$, one can define a finite belief MDP, and construct a policy for this finite model, by extending it, we can use the policy, say $\tilde{\phi}^N$ for the original model.

The cost function for the approximate model is

$$\begin{aligned}\hat{c}(\hat{z}^N_t, u_t) = \hat{c}(\hat{\pi}, I^N_t, u_t) &:= \tilde{c}(\phi(\hat{\pi}, I^N_t), u_t) \\ &= \int_{\mathbb{X}} c(x_t, u_t) P^{\hat{\pi}}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}).\end{aligned}$$

We define the controlled transition model for the approximate model by

$$\begin{aligned}&\hat{\eta}^N(\hat{z}^N_{t+1} | \hat{z}^N_t, u_t) \\ &= \hat{\eta}^N(\hat{\pi}, I^N_{t+1} | \hat{\pi}, I^N_t, u_t) := \hat{\eta}\left(\mathcal{P}(\mathbb{X}), I^N_{t+1} | \hat{\pi}, I^N_t, u_t\right)\end{aligned} \tag{15}$$

We will write $\mathcal{Z}^N_{\hat{\pi}}$ to make the dependence on $\hat{\pi}$ and N more explicit.

We denote the optimal value function for the approximate model by J^N_β, and the optimal policy for the approximate model by ϕ^N.

We investigate the following approximation error terms:

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

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 due to the policy calculated for the approximate model using finite memory being applied to the true model.

Building on [40], [36], we can show that the loss is related to the term:

$$\begin{aligned}L^N_t := \sup_{\hat{\gamma} \in \hat{\Gamma}} E^{\hat{\gamma}}_{\pi^-_0} \Big[&\| P^{\pi^-_t}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \\ &- P^{\hat{\pi}}(X_{t+N} \in \cdot | Y_{[t,t+N]}, U_{[t,t+N-1]}) \|_{TV} \Big].\end{aligned} \tag{16}$$

Let us elaborate on this term further. Consider the measurable policy space 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 $\{\hat{\gamma}_t\}$ such that $\hat{\gamma}_t$ is measurable with respect to the σ-algebra generated by the information variables $\{\hat{z}_0, \ldots, \hat{z}_t\}$. L^N_t above is then 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 the given policy. This term is directly related to filter stability, with bounds to be presented in the following.

Theorem 4.2: [42] [Continuity of Value Functions] For $\hat{z}_0 = (\pi^-_0, I^N_0)$, if a policy $\hat{\gamma}$ acts on the first N steps of the process which produces I^N_0, we then have

$$E^{\hat{\gamma}}_{\pi^-_0}\left[\left|\tilde{J}^N_\beta(\hat{z}_0) - J^*_\beta(\hat{z}_0)\right| | I^N_0\right] \leq \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L^N_t$$

Theorem 4.3: [42] [Near Optimality of Approximate Finite Window Model Solution applied to Actual Model] For $\hat{z}_0 = (\pi^-_0, I^N_0)$, with a policy $\hat{\gamma}$ acting on the first N steps,

$$E^{\hat{\gamma}}_{\pi^-_0}\left[\left|J_\beta(\hat{z}_0, \tilde{\phi}^N) - J^*_\beta(\hat{z}_0)\right| | I^N_0\right] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L^N_t. \tag{17}$$

Via a somewhat different, and more direct, derivation,[41, Section 4.2 and Theorem 17] presented the following alternative condition involving sample path-wise uniform filter stability term

$$\begin{aligned}\bar{L}^N_{TV} := \sup_{z \in \mathcal{P}(\mathbb{X})} \sup_{y_{[0,N]}, u_{[0,N-1]}} \\ \left\| P^z(\cdot | y_{[0,N]}, u_{[0,N-1]}) - P^{z^*}(\cdot | y_{[0,N]}, u_{[0,N-1]}) \right\|_{TV},\end{aligned} \tag{18}$$

to show the following *uniform* error bound:

$$\sup_z \left| J_\beta(z, \gamma_N) - J^*_\beta(z) \right| \leq \frac{2(1 + (\alpha_{\mathcal{Z}} - 1)\beta)}{(1-\beta)^3 (1 - \alpha_{\mathcal{Z}}\beta)} \|c\|_\infty \bar{L}^N_{TV} \tag{19}$$

for all $\beta \in (0,1)$ under a contraction condition, for some constant $\alpha_{\mathcal{Z}}$ defined in [41]. Additionally, [41, Theorem 9] provided conditions where the error is in the bounded-Lipschitz metric (which is equivalent to the Wasserstein-1 metric when the state space $\mathbb{X}$ is compact), however these were only applicable for a restrictive subset of the discount parameter β. On the other hand, the bound in (17) is in expectation whereas the bound in (19) is uniform, and thus the results are complementary.

1) Explicit filter stability bounds on expected filter error L_t^N and Sample Path Filter Error $\bar{L}_{TV}^N$: [42] shows that the term L_t^N can be bounded via Theorem 2.4: Recall that this states that

$$E^{\mu,\gamma}\left[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}\right] \leq 2\alpha^n. \tag{20}$$

which holds uniformly for all $\mu \ll \nu$ where $\alpha := (1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q))$, and $\delta(\cdot)$ denotes the Dobrushin coefficient of its argument (stochastic kernel). Since $\tilde{\delta}(\mathcal{T})$ is a uniform Dobrushin coefficient over all control actions, the above bound is valid under any control policy. Thus we have that $L_t^N \leq 2\alpha^N$.

As a complementary condition, via the Birkhoff-Hopf theorem, a controlled version of a contraction via the Hilbert metric [28] can be utilized [15]:

Recall that

$$F(z,y,u)(\cdot) = \Pr\{X_{k+1} \in \cdot \mid Z_k = z, Y_{k+1} = y, U_k = u\}$$

Assumption 4.2:

1) $Q(y|x) \geq \epsilon > 0$ for every $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.
2) The transition kernel $\mathcal{T}(.|.,u)$ is a mixing kernel (see Definition 2.5) for every $u \in \mathbb{U}$.

Lemma 4.1: [16] Under Assumption 4.2, there exists a constant $r < 1$ such that

$$h(F(\mu,y,u), F(\nu,y,u)) \leq rh(\mu,\nu) \tag{21}$$

for every comparable $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and for every $u \in \mathbb{U}$ and $y \in \mathbb{Y}$. Here $r = \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$, ϵ_u is the mixing constant of the kernel $\mathcal{T}(.|.,u)$.

Theorem 4.4: [15] Under Assumption 4.2, there exists a constant $r < 1$ and K such that

$$\bar{L}_{TV}^N \leq r^{N-1}K. \tag{22}$$

Here, $K = \frac{2}{\log 3}\sup h(Z_1, Z_1^*)$ and $r = \sup_{u\in\mathbb{U}} \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$.

Corollary 4.1: [15] Under Assumption 4.2, there exists a constant $r < 1$ and K such that

$$E_{z_0^-}^{\hat{\gamma}}\left[\left|\tilde{J}_\beta^N\left(\hat{z}_0, \tilde{\phi}^N\right) - J_\beta^*\left(\hat{z}_0\right)\right| \mid I_0^N\right] \leq \frac{2\|c\|_\infty}{(1-\beta)^2} r^{N-1}K. \tag{23}$$

Here, $K = \frac{2}{\log 3}\sup h(Z_1, Z_1^*)$ and $r = \sup_{u\in\mathbb{U}} \frac{1-\epsilon_u^2\epsilon}{1+\epsilon_u^2\epsilon}$.

Remark 4.1: Among recent results, [9] provides an upper error bound for finite window policies, under a persistence of excitation of the optimal policy and minorization-majorization assumptions, [9] demonstrates that, as N increases, the error term converges to zero geometrically. Unlike Assumption 4.2, a persistence of excitation of the optimal policy requires that the optimal policy must be strictly non-deterministic. We note also that the state space in our setup is not necessarily finite.

Implementing the above is still tedious, though now numerically possible. Can reinforcement learning be feasible? Can we view the finite history as an approximate *state* to run a learning algorithm? Would such an algorithm convergence, and what would such a convergence operationally mean? We address these questions, building on [43] and [42], in the following section.

C. Robustness to Incorrect Models and Priors

To complete our analysis on existence, regularity, and approximations, and before proceeding with reinforcement learning, we also review recent results on robustness to incorrect priors and models.

Suppose that we represent the cost of the model given in (1)-(2) and cost criteria (4-5), so that the dependence on the prior μ, transition kernel $\mathcal{T}$, measurement channel Q, and the stage-wise cost c is explicitly given as follows:

$$J_\infty(c,\mu,\mathcal{T},Q,\gamma) := \limsup_{N\to\infty} \frac{1}{N} E_\mu^\gamma\left[\sum_{k=0}^{N-1} c(X_k,U_k)\right], \tag{24}$$

with the infimum

$$J_\infty^*(c,\mu,\mathcal{T},Q) = \inf_{\gamma\in\Gamma} J_\infty(c,\mu,\mathcal{T},Q,\gamma)$$

or the discounted cost criterion (for some $\beta \in (0,1)$

$$J_\beta(c,\mu,\mathcal{T},Q,\gamma) := E_\mu^\gamma\left[\sum_{k=0}^{\infty} \beta^k c(X_k,U_k)\right], \tag{25}$$

with the infimum being

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

The robustness question involves the following.

Problem P1: Continuity of Optimal Cost under the Convergence of Models. Let $\{\mu_n, \mathcal{T}_n, Q_n, n \in \mathbb{N}\}$ be a sequence of priors, transition kernels, and channels which converges in some sense to another model $(\mu, \mathcal{T}, Q)$ and $\{c_n, n \in \mathbb{N}\}$ be a sequence of stage-wise cost functions corresponding to $(\mu_n, \mathcal{T}_n, Q_n)$ which converge in some sense to another cost function c corresponding to $(\mu, \mathcal{T}, Q)$. Does that imply that

$$J_\beta^*(c_n,\mu_n,\mathcal{T}_n,Q_n) \to J_\beta^*(c,\mu,\mathcal{T},Q)?$$

Problem P2: Robustness to Incorrect Models. A problem of major practical importance is robustness of an optimal controller to modeling errors. Suppose that an optimal policy is constructed according to a model which is incorrect: how does the application of the control to the true model affect the system performance and does the error decrease to zero as the models become closer to each other? In particular, suppose that γ_n^* is an optimal policy designed for $\{c_n, \mu_n, \mathcal{T}_n, Q_n, n \in \mathbb{N}\}$. Is it the case that if $(c_n, \mu_n, \mathcal{T}_n, Q_n)$ converges in some appropriate sense to $(c, \mu, \mathcal{T}, Q)$, then

$$J^*_\beta(c, \mu, \mathcal{T}, Q, \gamma^*_n) \to J^*_\beta(c, \mu, \mathcal{T}, Q).$$

The case where only $\mu_n \to \mu$ while the other parameter are fixed is referred to as *robustness to incorrect priors*.

Problem P3: Empirical Consistency of Learned Probabilistic Models and Data-Driven Stochastic Control. Let $(\mathcal{T}(\cdot|x, u), Q(\cdot|x))$ be the transition and measurement kernels, which is unknown to the decision maker (DM). Suppose the DM builds a model for these kernels, $(\mathcal{T}_n(\cdot|x, u), Q_n(dy|x))$, for all possible $x \in \mathbb{X}, u \in \mathbb{U}$ by collecting training data (e.g. from an evolving system). Do we have that the optimal cost calculated under $(\mathcal{T}_n, Q_n)$ converges to the true cost (i.e., do we have that the cost obtained from applying the optimal policy for the empirical model converges to the true cost as the training length increases)?

1) Robustness to incorrect priors: We refer the reader to [39, Theorem 3.2] and with further refinements under filter stability [48, Theorem 3.8] for discounted cost and [48], Theorem 3.9] for average cost. These show that the problem is robust to uncertainty in priors under total variation, and for robustness under weak convergence, total variation continuity of the channel as in 2.1(ii) is to be imposed. Further regularity results are present in [39]. Under filter stability, stronger robustness conditions are presented in [48].

2) Robustness to incorrect models: We refer the reader to [40], [36], [35] for robustness to transition kernel, cost functions (and which also apply to that in measurement channels); and to [75], [74] for the special case of convergence of measurement channels. To present a flavour of results, we state the following.

Assumption 4.3: (i) The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\{\mathcal{T}_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $\mathcal{T}(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $(x_n, u_n) \to (x, u)$ (this is referred to as *continuous* weak convergence [40], [36], [35]).

(ii) The stochastic kernel $\mathcal{T}(\cdot|x, u)$ is weakly continuous in (x, u).

(iii) The sequence of stage-wise cost functions c_n satisfies the following: $c_n(x_n, u_n) \to c(x, u)$ for any sequence $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $(x_n, u_n) \to (x, u)$.

(iv) The stage-wise cost function $c(x, u)$ is non-negative, bounded, and continuous on $\mathbb{X} \times \mathbb{U}$.

(v) $\mathbb{U}$ is compact.

The following hold:

(a) Continuity and robustness do not hold in general under weak convergence of kernels.

(b) Under Assumptions 4.3 and 2.1(ii), continuity and robustness hold.

(c) Continuity and robustness do not hold in general under setwise convergence of the kernels.

(d) Continuity and robustness do not hold in general under total variation convergence of the kernels.

(e) Continuity and robustness hold under *continuous* total variation convergence of the kernels (i.e. if $\mathcal{T}_n(\cdot|x, u_n) \to \mathcal{T}(\cdot|x, u)$ in total variation for any $u_n \to u$ and for any x).

The above has direct implications on data-driven learning and empirical consistency, where empirical models are constructed via data, and empirical models converge weakly (and under the W_1 distance) almost surely ([20], Theorem 11.4.1), but do not so under total variation unless density conditions are present [17, Chapter 3]; see [40], [36], [35].

V. Reinforcement Learning for POMDPs: Discounted Cost

A. A General Convergence Result for Asymptotically Ergodic Processes

We summarize the following, building on [43]. Let $\{C_t\}_t$ be $\mathbb{R}$-valued, $\{S_t\}_t$ be $\mathbb{S}$-valued and $\{U_t\}_t$ be $\mathbb{U}$-valued three stochastic processes. Consider the following iteration defined for each $(s, u) \in \mathbb{S} \times \mathbb{U}$ pair

$$\begin{aligned} Q_{t+1}(s, u) = &(1 - \alpha_t(s, u))\, Q_t(s, u) \\ &+ \alpha_t(s, u)\, (C_t + \beta V_t(S_{t+1})) \end{aligned} \tag{26}$$

where $V_t(s) = \min_{u \in \mathbb{U}} Q_t(s, u)$, and $\alpha_t(s, u)$ is a sequence of constants also called the learning rates. Note also that we overwrite the notation in this section by using Q for the Q values which is different than the channel kernel $Q(\cdot|x)$.

Consider the following equation

$$Q^*(s, u) = C^*(s, u) + \beta \sum_{s_1 \in \mathbb{S}} V^*(s_1) P^*(s_1|s, u) \tag{27}$$

for some functions Q^*, C^*, to be defined explicitly, and for some regular conditional probability distribution $P^*(\cdot|s, u)$, where $V^*(u) := \min_u Q^*(s, u)$.

An umbrella sufficient condition is the following:

Assumption 5.1: $\mathbb{S}, \mathbb{U}$ are finite sets, and the joint process (S_{t+1}, S_t, U_t, C_t) is asymptotically ergodic in the sense that for the given initialization random variable S_0, for any measurable function f, we have that with probability one,

$$\begin{aligned} &\lim_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} f(S_{t+1}, S_t, U_t, C_t) \\ &\quad = \int f(s_1, s, u, c) \phi(ds_1, ds, du, dc) \end{aligned}$$

for some measure ϕ such that the marginal on the second and third coordinates $\phi(\mathbb{S} \times B \times \mathbb{R}) > 0$ for any non-empty $B \subset \mathbb{S} \times \mathbb{U}$.

We note that although we assume that the spaces $\mathbb{S}, \mathbb{U}$ are finite, we will continue using integral and differential notation for consistency.

The above implies Assumption 5.2(ii)-(iii) below:

Assumption 5.2: i. $\alpha_t(s, u) = 0$ unless $(S_t, U_t) = (s, u)$. Furthermore,

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

and with probability 1, $\sum_t \alpha_t(s,u)=\infty$.

ii. For C_t, we have

$$\frac{\sum_{k=0}^{t} C_k 1_{\{S_k=s,U_k=u\}}}{\sum_{k=0}^{t} 1_{\{S_k=s,U_k=u\}}} \to C^*(s,u),$$

almost surely for some C^*.

iii. For the S_t process, we have, for any function f,

$$\frac{\sum_{k=0}^{t} f(S_{k+1}) 1_{\{S_k=s,U_k=u\}}}{\sum_{k=0}^{t} 1_{\{S_k=s,U_k=u\}}} \to \int f(s_1)P^*(ds_1|s,u)$$

almost surely for some P^*.

Recently, [43] presented conditions for the convergence of the iterates above:

Theorem 5.1: [43] Under Assumption 5.2, $Q_t(s,u) \to Q^*(s,u)$ almost surely for each $(s,u) \in \mathbb{S} \times \mathbb{U}$ pair where Q^* satisfies (27).

It turns out that (27) is the fixed point corresponding to an approximate MDP, with implications for POMDPs noted in the following (see also [43]).

B. Finite Window Memory POMDP with Uniform Geometric Controlled Filter Stability

We will here assume that $\mathbb{X}$ is a compact subset of a Polish space and that $\mathbb{Y}$ and $\mathbb{U}$ are finite sets.

Suppose that the controller keeps a finite window of the most recent N observation and control action variables, and perceives this as the *state* variable, which is in general non-Markovian. That is we take

$$S_t = \{Y_{[t-N,t]}, U_{[t-N,t-1]}\},$$

and $C_t := c(X_t, U_t)$.

In this case, (S_t, X_t, U_t) form a controlled Markov chain, even if (S_t, U_t) does not. We state the ergodicity assumption formally next.

Assumption 5.3: (i) Under the exploration policy γ and initialization, the controlled state and control action joint process $\{X_t, U_t\}$ is asymptotically ergodic in the sense that for any measurable function f we have that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} f(X_t,U_t) = \int f(x,u)\phi^\gamma(dx,du)$$

for some $\phi^\gamma \in \mathcal{P}(\mathbb{X}\times\mathbb{U})$. Furthermore, we have that $P(Y_t = y|x) > 0$ for every $x \in \mathbb{X}$.

(ii) Assumption 5.1(i) holds with $S_t = \{Y_{[t-N,t]}, U_{[t-N,t-1]}\}$.

We note that a sufficient condition for the ergodicity assumption, for every initialization of X_0, would be positive Harris recurrence under the exploration policy.

The question then is if the limit Q values correspond to a meaningful control problem, and how 'close' this control problem to the original POMDP. [42, Theorem 4.1] shows that the limit Q values indeed correspond to an approximate control problem, and notes the following bound:

Theorem 5.2: [42, Theorem 4.1] Under Assumption 5.3, the iterations in (26) converge with $S_t = \{Y_{[t-N,t]}, U_{[t-N,t-1]}\}$ and $C_t := c(X_t, U_t)$. Furthermore, if we denote the policies constructed using these Q values by γ^N, and apply these finite memory policies in the original problem, we get the following error bound:

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

where I_0^N is the first N observation and control variables, and the expectation is taken with respect to different realizations of I_0^N under the initial distribution of the hidden state π_0 and the exploration policy γ. Furthermore, $\pi_N^- = P(X_N \in \cdot | I_0^N)$ where L_t^N is given by (16) with the fixed prior is the invariant measure on x_t under the exploration policy γ. In particular, we assume that the control starts after observing at least N- time steps of history.

As noted earlier, L_t^N is related to the filter stability problem, see (20).

C. Quantized Approximations for Weak Feller POMDPs with only Asymptotic Filter Stability

As noted earlier, any POMDP can be reduced to a completely observable Markov process ([76], [52]) (see (6)), whose states are the posterior state distributions or *belief*s of the observer; that is, the state at time t is the filter variable

$$\pi_t(\,\cdot\,) := P\{X_t \in \cdot\, | y_0, \ldots, y_t, u_0, \ldots, u_{t-1}\} \in \mathcal{P}(\mathbb{X}).$$

Recall the kernel η (6) for the filter process. Now, by combining the quantized Q-learning and the weak Feller continuity results for the non-linear filter kernel ([23] [37]), we can conclude that the setup in Section V-A and Section IV-A is applicable though with a significantly more tedious analysis involving ergodicity requirements. Additionally, one needs to quantize probability measures. Accordingly, we take $S_t = g(\pi_t)$ for some quantizer $g : \mathcal{P}(\mathbb{X}) \to \mathcal{P}(\mathbb{X})^M =: \{B_1, B_2, \cdots, B_{|\mathcal{P}(\mathbb{X})^M|}\}$ with $|\mathcal{P}(\mathbb{X})^M| < \infty$, and $C_t := c(X_t, U_t)$.

We state the ergodicity condition formally:

Assumption 5.4: Under the exploration policy γ and initialization, the controlled belief state and control action joint process $\{\pi_t, U_t\}$ is asymptotically uniquely ergodic in the sense that for any measurable function f we have that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} f(\pi_t) = \int f(\pi)\eta^\gamma(d\pi)$$

for some $\eta^\gamma \in \mathcal{P}(\mathcal{P}(\mathbb{X})\times\mathbb{U})$ such that $\eta^\gamma(B) > 0$ for any quantization bin $B \subset \mathcal{P}(\mathbb{X})$.

We refer to the set

$$\mathcal{P}_\eta := \{\pi : \pi \in B_i \subset \mathcal{P}(\mathbb{X}) : \eta^\gamma(B_i) > 0\},$$

as the trained set of states; since these sets will be visited infinitely often under the exploration policy.

The condition that $\eta^\gamma(B) > 0$ requires an analysis tailored for each problem. For example, if the quantization is performed as in [41] by clustering bins based on a finite past window, then the condition is satisfied by requiring that $P(Y_t = y|x) > 0$ for every $x \in \mathbb{X}$. If the clustering is done, e.g. by quantization of the probability measures via first quantizing $\mathbb{X}$ and then quantizing the probability measures on the finite set (see [57, Section 5]), then the initialization could be done according to the invariant probability measure corresponding to the hidden Markov source.

Unique ergodicity of the dynamics follows from results in the literature, such as, [46, Theorem 2] and [67, Prop 2.1], which holds when the randomized control is memoryless under mild conditions on the process notably, the hidden variable is a uniquely ergodic Markov chain and the measurement structure satisfies filter stability in total variation in expectation (one can show that weak merging in expectation also suffices); we refer the reader to [49, Figure 1] for mild conditions leading to filter stability in this sense, which is related to stochastic observability in Definition 2.7 (see also [49, Definition II.1]). Notably, a uniform and geometric controlled filter stability is not required even though this would be sufficient. Therefore, due to the weak Feller property of controlled non-linear filters, we can apply the Q-learning algorithm to also belief-based models to arrive at near optimal control policies. Nonetheless, since positive Harris recurrence cannot typically be assumed for the filter process, the initial state may not be arbitrary. If the invariant measure under the exploration policy is the initial state, [67, Prop 2.1] implies that the time averages will converge as imposed in Assumption 5.2. A sufficient condition for unique ergodicity then is the following.

Assumption 5.5: Under the exploration policy γ the hidden process $\{X_t\}$ is uniquely ergodic (with measure ζ) and the measurement dynamics are so that the filter is stable in expectation under weak convergence.

Theorem 5.3: Suppose that Assumption 4.1 holds such that $\alpha_\eta \beta < 1$.

(a) Suppose that under the exploration policy and initialization, the controlled filter process satisfies Assumption 5.4 and 5.2(i) with $S_t = g(\pi_t)$, and $C_t = c(X_t, U_t)$. Then, the Q_t iterates converge almost surely.

(b) Let $\pi_0 \sim \kappa \ll \phi$ or $\pi_0 \in \operatorname{supp}(\phi)$ and under η^γ the boundary sets of the bins have measure zero. Then, the policy constructed using the limit Q values, say $\hat{\gamma}$, applied to the true model leads to the following bound:

$$J^*_\beta(\pi_0, \hat{\gamma}) - J^*_\beta(\pi_0) \leq \frac{2\alpha_c}{(1-\beta)^2(1-\beta\alpha_\eta)} \bar{L}.$$

where

$$\bar{L} := \sup_{\pi \in \operatorname{supp}(\eta^\gamma), \pi \in B_i : \eta^\gamma(B_i) > 0} W_1\left(\pi, g(\pi)\right).$$

(c) For asymptotic convergence (without a rate of convergence) to optimality as the quantization rate goes to ∞ (i.e., $\bar{L} \to 0$), only weak Feller property of η is sufficient for the the algorithm to be near optimal.

A sufficient condition for the assumptions of (a) and (b) above is that for exploration (i) $\pi_0 \sim \kappa \ll \phi$, or (ii) $\pi_0 = \zeta$ where ζ is the invariant measure for the hidden state process under exploration and that under η^γ the boundary sets of the bins have measure zero, or (iii) there exists $\pi \in \mathcal{P}(\mathbb{X})$ such that for all $\pi_0, P(\inf\{k > 0 : \pi_k = \pi\} < \infty) = 1$. For further discussion on the initialization for the algorithm during implementation, please see [12, Lemma 6 and Corollary 2].

Remark 5.1: We now present a comparison between the two approaches above: filter quantization vs. finite window based learning:

(i) For the filter quantization, we only need unique ergodicity of the filter process under the exploration policy for which asymptotic filter stability in expectation in weak or total variation is sufficient. The running cost can start immediately without waiting for a window of measurements. On the other hand, the controller must run the filter and quantize it in each iteration while running the Q-learning algorithm; accordingly the controller must know the model. Additionally, the initialization cannot be arbitrary (e.g. the initialization for the filter may be the invariant measure under the exploration policy): As noted earlier, one needs to ensure that the set of bin-action pairs which are visited infinitely often during exploration is so that an optimal policy is learned (visited infinitely often), and when this optimal policy (learned via the convergence of Q-learning) is implemented, the closed-loop process always remains in this set; see [43] and [12, Lemma 6 and Corollary 2].

(ii) For the finite window approach, a uniform convergence of filter stability, via L_t^N, is needed and it does not appear that only asymptotic filter stability can suffice. On the other hand, this is a universal algorithm in that the controller does not need to know the model. Furthermore, the initialization satisfaction holds under explicit conditions; notably if the hidden process is positive Harris recurrent, the ergodicity condition holds for every initialization; both the convergence of the algorithm as well as its implementation will always be well-defined.

For each setup, however, we have explicit and testable conditions.

VI. The Average Cost Case

Approximations and learning for POMDPs under the average cost criterion is significantly more challenging. In the classical MDP theory, the approaches primarily require strong ergodicity or minorization conditions which are not suitable for the belief-MDP. Several papers [51], [25], [53], [31] have studied the average-cost control problem under the assumption that the state space is finite; they provide reachability type conditions for the belief kernels. Reference [6] considers the finite model setup and [65] considers the case with finite-dimensional real-valued state spaces under several technical conditions on the controlled state process and [62] considers several conditions directly on the filter process leading to an equi-continuity condition on the relative discounted value functions. One could adopt techniques suitable for the average cost without needing minorization conditions, see [14].

Average Cost via Near Optimality of Discounted Cost Policies. Building on [14], consider the following two conditions: (i) There exists a solution to the average cost optimality equation as in (13), and (ii) that this solution is obtained via the vanishing discount method. Under these conditions, it follows that (see [12, Theorems 1 and 2] and [73, Theorem 7.3.6]) a near optimal policy for the discounted cost problem is also near optimal for the average cost problem.

Theorem 3.2(i) implies these conditions simultaneously. Consequently:

(i) Accordingly, under Assumption 2.3, with $K_2 = \frac{\alpha D(3-2\delta(Q))}{2} < 1$, by Theorem 3.2(i), a learning method would be to approximate the Q-learning algorithm by its classical discounted version by taking β sufficiently large. Therefore, the methods available for discounted cost also apply to the average cost problem for near optimality. For details, please see [14].

(ii) Another implication of Theorem 3.2(i) is that, generalizing [71], one can conclude that finite window policies are near optimal for average cost problems under controlled filter stability conditions.

(iii) A further byproduct of this approach is the complete robustness to incorrect initializations for POMDPs in average cost problems, as reported in [14, Corollary 3.1], connecting [48, Theorem 3.9] with Theorem 3.2(i).

VII. Conclusion

In this article a general review on partially observable Markov Decision Processes has been presented. The focus has been on regularity (including continuity and filter stability) and associated existence results, approximate optimality via finite approximations or finite memory policies, and a rigorous analysis on reinforcement learning to near optimality.

VIII. Acknowledgements

We gratefully acknowledge our collaborations with Naci Saldi, Yunus Emre Demirci, Curtis McDonald and Tamás Linder; and our many discussions with Prof. Eugene Feinberg, Vivek Borkar, Erhan Bayraktar, Gürdal Arslan, and Aditya Mahajan.

References

[1] R. J. Aumann. Mixed and behavior strategies in infinite extensive games. Technical report, Princeton University NJ, 1961.

[2] K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pages 193–256. PMLR, 2016.

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

[4] D. Blackwell. Memoryless strategies in finite-stage dynamic programming. *Annals of Mathematical Statistics*, 35:863–865, 1964.

[5] V. S. Borkar. White-noise representations in stochastic realization theory. *SIAM J. on Control and Optimization*, 31:1093–1102, 1993.

[6] V. S. Borkar. Average cost dynamic programming equations for controlled Markov chains with partial observations. *SIAM J. Control Optim.*, 39(3):673–681, 2000.

[7] V. S. Borkar. Convex analytic methods in Markov decision processes. In *Handbook of Markov Decision Processes, E. A. Feinberg, A. Shwartz (Eds.)*, pages 347–375. Kluwer, Boston, MA, 2001.

[8] A. Budhiraja. On invariant measures of discrete time filters in the correlated signal-noise case. *The Annals of Applied Probability*, 12(3):1096–1113, 2002.

[9] S. Cayci, N. He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *arXiv preprint arXiv:2206.00833*, 2022.

[10] S. Chandak, V.S. Borkar, and P. Dodhia. Reinforcement learning in non-markovian environments. *Systems & Control Letters*, 185:105751, 2024.

[11] P. Chigansky and R. Van Handel. A complete solution to Blackwell's unique ergodicity problem for hidden Markov chains. *The Annals of Applied Probability*, 20(6):2318–2345, 2010.

[12] L. Cregg, T. Linder, and S. Yüksel. Reinforcement learning for near-optimal design of zero-delay codes for markov sources. *IEEE Transactions on Information Theory, arXiv:2311.12609*, 2024.

[13] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, 2002.

[14] Y.E. Demirci, A.D. Kara, and S. Yüksel. Average cost optimality of partially observed mdps: Contraction of non-linear filters and existence of optimal solutions. *SIAM Journal on Control and Optimization, arXiv:2312.14111*, 2024.

[15] Y.E. Demirci, A.D. Kara, and S. Yüksel. Refined bounds on near optimality finite window policies in pomdps and their reinforcement learning. *arXiv*, 2024.

[16] Y.E. Demirci and S. Yüksel. Geometric ergodicity and wasserstein continuity of non-linear filters. *arXiv preprint arXiv:2307.15764*, 2023.

[17] L. Devroye and L. Györfi. *Non-parametric Density Estimation: The L_1 View*. John Wiley, New York, 1985.

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

[19] S. Dong, B. van Roy, and Z. Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent states. *The Journal of Machine Learning Research*, 23(1):11627–11680, 2022.

[20] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2nd edition, 2002.

[21] E. Even-Dar, S.M. Kakade, and Y. Mansour. Reinforcement learning in pomdps without resets. In *IJCAI*, pages 690–695, 2005.

[22] E.A. Feinberg and P.O. Kasyanov. Equivalent conditions for weak continuity of nonlinear filters. *Systems & Control Letters*, 173:105458, 2023.

[23] E.A. Feinberg, P.O. Kasyanov, and M.Z. Zgurovsky. Partially observable total-cost Markov decision process with weakly continuous transition probabilities. *Mathematics of Operations Research*, 41(2):656–681, 2016.

[24] E.A. Feinberg, P.O. Kasyanov, and M.Z. Zgurovsky. Markov decision processes with incomplete information and semiuniform feller transition probabilities. *SIAM Journal on Control and Optimization*, 60(4):2488–2513, 2022.

[25] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. Remarks on the existence of solutions to the average cost optimality equation in markov decision processes. *Systems & control letters*, 15(5):425–432, 1990.

[26] F.Kochman and J. Reeds. A simple proof of kaijser's unique ergodicity result for hidden markov α-chains. *The Annals of Applied Probability*, 16(4):1805–1815, 2006.

[27] I. I. Gihman and A. V. Skorohod. *Controlled Stochastic Processes*. Springer Science & Business Media, 2012.

[28] F. Le Gland and N. Oudjane. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters. *The Annals of Applied Probability*, 14(1):144–187, 2004.

[29] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, 1989.

[30] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[31] S.-P. Hsu, D.-M Chuang, and A. Arapostathis. On the existence of stationary optimal policies for partially observed mdps under the long-run average cost criterion. *Systems & control letters*, 55(2):165–173, 2006.

[32] C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

[33] T. Kaijser. A limit theorem for partially observed Markov chains. *Annals of Probability*, 4(677), 1975.

[34] T. Kaijser. On markov chains induced by partitioned transition probability matrices. *Acta Mathematica Sinica, English Series*, 27(3):441–476, 2011.

[35] A. D. Kara, M. Raginsky, and S. Yüksel. Robustness to incorrect models and data-driven learning in average-cost optimal stochastic control. *Automatica*, 139:110179, 2022.

[36] A. D. Kara and S. Yüksel. Robustness to approximations and model learning in MDPs and POMDPs. In A. B. Piunovskiy and Y. Zhang, editors, *Modern Trends in Controlled Stochastic Processes: Theory and Applications, Volume III*. Luniver Press, 2021.

[37] A.D Kara, N. Saldi, and S. Yüksel. Weak Feller property of non-linear filters. *Systems & Control Letters*, 134:104–512, 2019.

[38] A.D Kara, N. Saldi, and S. Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. *Journal of Machine Learning Research*, pages 1–34, 2023.

[39] A.D Kara and S. Yüksel. Robustness to incorrect priors in partially observed stochastic control. *SIAM Journal on Control and Optimization*, 57(3):1929–1964, 2019.

[40] A.D Kara and S. Yüksel. Robustness to incorrect system models in stochastic control. *SIAM Journal on Control and Optimization*, 58(2):1144–1182, 2020.

[41] A.D Kara and S. Yüksel. Near optimality of finite memory feedback policies in partially observed markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022.

[42] A.D Kara and S. Yüksel. Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4):2066–2093, 2023.

[43] A.D. Kara and S. Yüksel. Q-learning for stochastic control under general information structures and non-Markovian environments. *Transactions on Machine Learning Research (arXiv:2311.00123)*, 2024.

[44] J. Kwon, Y. Efroni, C. Caramanis, and S. Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34:24523–24534, 2021.

[45] H.J. Langen. Convergence of dynamic programming models. *Mathematics of Operations Research*, 6(4):493–512, Nov. 1981.

[46] G. Di Masi and L. Stettner. Ergodicity of hidden markov models. *Mathematics of Control, Signals and Systems*, 17(4):269–296, 2005.

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

[48] C. McDonald and S. Yüksel. Robustness to incorrect priors and controlled filter stability in partially observed stochastic control. *SIAM Journal on Control and Optimization*, 60(2):842–870, 2022.

[49] C. McDonald and S. Yüksel. Stochastic observability and filter stability under several criteria. *IEEE Transactions on Automatic Control*, 69(5):2931–2946, 2024.

[50] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. AMS Bookstore, 1967.

[51] L. K. Platzman. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control and Optimization*, 18(4):362–380, 1980.

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

[53] Wolfgang J Runggaldier and Lukasz Stettner. *Approximations of discrete time partially observed control problems*. Giardini Pisa, 1994.

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

[55] N. Saldi, S. Yüksel, and T. Linder. Near optimality of quantized policies in stochastic control under weak continuity conditions. *Journal of Mathematical Analysis and Applications*, 435(1):321–337, 2016.

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

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

[58] R. Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 380–402, 1982.

[59] E. Seyedsalehi, N. Akbarzadeh, A. Sinha, and A. Mahajan. Approximate information state based convergence analysis of recurrent q-learning. *arXiv preprint arXiv:2306.05991*, 2023.

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

[61] A. Sinha and A. Mahajan. Agent-state based policies in pomdps: Beyond belief-state mdps. In *IEEE Conference on Decision and Control, Tutorial Paper*, 2024.

[62] L. Stettner. Long run control with degenerate observation. *SIAM Journal on Control and Optimization*, 57(2):880–899, 2019.

[63] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *The Journal of Machine Learning Research*, 23(1):483–565, 2022.

[64] T. Szarek. Feller processes on nonlocally compact spaces. *Annals of Probability*, pages 1849–1863, 2006.

[65] A. Budhiraja V. S. Borkar. A further remark on dynamic programming for partially observed markov processes. *Stochastic processes and their applications*, 112(1):79–93, 2004.

[66] R. van Handel. Observability and nonlinear filtering. *Probability theory and related fields*, 145(1-2):35–74, 2009.

[67] R. van Handel. Uniform time average consistency of Monte Carlo particle filters. *Stochastic Processes and their Applications*, 119(11):3835–3861, 2009.

[68] R. van Handel. Nonlinear filtering and systems theory. In *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS semi-plenary paper)*, 2010.

[69] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[70] Y. Xiong, N. Chen, X. Gao, and X. Zhou. Sublinear regret for learning pomdps. *Production and Operations Management*, 31(9):3491–3504, 2022.

[71] H. Yu and D. P. Bertsekas. On near optimality of the set of finite-state controllers for average cost pomdp. *Mathematics of Operations Research*, 33(1):1–11, 2008.

[72] S. Yüksel. Another look at partially observed optimal stochastic control: Existence, ergodicity, and approximations without belief-reduction. *arXiv*, 2023.

[73] S. Yüksel. *Optimization and Control of Stochastic Systems*. Queen's University, Lecture Notes, available online, Queen's University, Lecture notes.

[74] S. Yüksel and T. Başar. *Stochastic Teams, Games, and Control under Information Constraints*. Springer, Cham, 2024.

[75] S. Yüksel and T. Linder. Optimization and convergence of observation channels in stochastic control. *SIAM J. on Control and Optimization*, 50:864–887, 2012.

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