SIAM J. CONTROL OPTIM.
Vol. 62, No. 6, pp. 2859–2883

AVERAGE COST OPTIMALITY OF PARTIALLY OBSERVED MDPs: CONTRACTION OF NONLINEAR FILTERS AND EXISTENCE OF OPTIMAL SOLUTIONS AND APPROXIMATIONS*

YUNUS EMRE DEMIRCI[†], ALI DEVRAN KARA[‡], AND SERDAR YÜKSEL[†]

Abstract. The average cost optimality is known to be a challenging problem for partially observable stochastic control, with few results available beyond the finite state, action, and measurement setup, for which somewhat restrictive conditions are available. In this paper, we present explicit and easily testable conditions for the existence of solutions to the average cost optimality equation where the state space is compact. In particular, we present a novel contraction based analysis, which, to the best of our knowledge, is new to the literature, building on recent regularity results for nonlinear filters. Beyond establishing existence, we also present several implications of our analysis that also are new to the literature: (i) robustness to incorrect priors, (ii) near optimality of policies based on quantized approximations, (iii) near optimality of policies with finite memory, and (iv) convergence in Q-learning. In addition to our main theorem, each of these represents a novel contribution for average cost criteria.

Key words. nonlinear filtering, average cost optimality equation

MSC codes. 90C40, 93E11, 93E20

DOI. 10.1137/24M1643736

1. Introduction.

We study optimal control for partially observable Markov decision processes (PODMPs) under the average cost criterion. Let $\mathbb{X}$ denote a standard Borel space, which is the state space of a partially observed controlled Markov process. Let $\mathbb{B}(\mathbb{X})$ be its Borel σ-field. Let $\mathbb{C}_b(\mathbb{X})$ be the set of all continuous, bounded functions on $\mathbb{X}$. 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 standard Borel space denoting the observation space of the model, and let the state be observed through an observation channel Q. The observation channel, Q, is defined as a stochastic kernel (regular conditional probability) from $\mathbb{X}$ to $\mathbb{Y}$, such that $Q(\,\cdot\,|x)$ is a probability measure on the power set $P(\mathbb{Y})$ of $\mathbb{Y}$ for every $x \in \mathbb{X}$, and $Q(A|\,\cdot\,) : \mathbb{X} \to [0,1]$ is a Borel measurable function for every $A \in P(\mathbb{Y})$. A decision maker (DM) is located at the output of the channel Q and hence only sees the observations $\{Y_t,\, t \in \mathbb{Z}_{\geq 0}\}$ and chooses its actions from $\mathbb{U}$, the action space which is a compact set. An *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_{\geq 0}\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables $I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \quad I_0 = \{Y_0\}$, where

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_{\geq 0}, \tag{1.1}$$

are the $\mathbb{U}$-valued control actions and $Y_{[0,t]} = \{Y_s,\, 0 \leq s \leq t\}, \quad U_{[0,t-1]} = \{U_s,\, 0 \leq s \leq t-1\}$. In the above, the dependence of control policies on the initial distribution

*Received by the editors March 4, 2024; accepted for publication (in revised form) July 2, 2024; published electronically November 4, 2024. Part of the results in this paper will be presented at the 2024 American Control Conference.
https://doi.org/10.1137/24M1643736
Funding: The research of the first and third authors was supported by the Natural Science and Engineering Research Council of Canada (NSERC).
[†]Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (21yed@queensu.ca, yuksel@queensu.ca).
[‡]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 USA (alikara@umich.edu).

π_0 is implicit. We will denote the collection of admissible control policies as Γ. The update rules of the system are determined by (1.1) and the following relationships:

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

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

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

$B\in\mathcal{B}(\mathbb{X}\times\mathbb{Y}), t\in\mathbb{N}$, where $\mathcal{T}$ is the transition kernel of the model which is a stochastic kernel from $\mathbb{X}\times\mathbb{U}$ to $\mathbb{X}$.

We may let the agent's goal be to minimize the expected discounted cost

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

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 the stagewise measurable cost function, and the expectation E_μ^γ is taken over the initial state probability measure μ under policy γ. The optimal cost for the discounted infinite horizon is defined as

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

The average cost control problem under partial observations is shown as follows involves finding an optimal policy that minimizes the average cost of the system over an infinite horizon:

$$J^*(\mu)=\inf_{\gamma\in\Gamma} J(\mu,\gamma)$$

where

$$J(\mu,\gamma)=\limsup_{n\to\infty}\frac{1}{n}E_\mu^\gamma\left[\sum_{t=0}^{n-1}c(X_t,U_t)\right].$$

Contributions.

(i) In the theory of partially observable Markov decision processes (POMDPs), very few conditions for the existence of solutions to the average cost optimality equation (ACOE)have been found in the literature, especially in scenarios beyond those with finite states, actions, and measurements, often restricted by stringent conditions, as we will discuss in detail. This paper introduces a novel result concerning the existence of a solution for the ACOE by utilizing a Wasserstein regularity result (Theorem 2.2), as detailed in Theorem 1.2. Under Assumption 1, we show the existence of a solution for the ACOE. We provide a detailed comparison to existing results in the literature and highlight the explicit nature of our conditions.

(ii) Subsequently, the paper then presents several applications and implications of the existence of a solution to the ACOE for POMDPs. To the best of our knowledge, the following lead to new contributions in this area of study:

1. Subsection 3.1 establishes robustness to incorrect priors under universal filter stability, demonstrating that an optimal policy designed for an incorrect prior remains optimal when applied for the correct prior under the average cost criteria, leading to a complete robustness property. Additionally, the subsection includes an analysis for the discounted cost setup.
2. Subsection 3.2.1 establishes near optimality of quantized approximation policies for average cost criteria.
3. Subsection 3.2.2 examines how, under conditions of filter stability, optimal policies derived from a finite window of measurements and actions for the discounted cost criteria are near-optimal solutions for average cost criteria in POMDPs.
4. Subsection 3.2.2 also highlights the use of Q-learning to derive near-optimal policies applicable for both discounted finite window and discounted quantized approximation scenarios.

1.1. Notation and preliminaries.

Belief MDP reduction for POMDPs. It is known that any POMDP can be reduced to a completely observable Markov process [35, 27], whose states are the posterior state distributions or *belief*s of the observer; that is, the state at time n is

$$z_n(\cdot) := P\{X_n \in \cdot | y_0, \ldots, y_n, u_0, \ldots, u_{n-1}\} \in \mathcal{P}(\mathbb{X}).$$

We call this equivalent process the filter process. We denote by $\mathcal{Z} := \mathcal{P}(\mathbb{X})$ the set of probability measures on $(\mathbb{X}, \mathbb{B}(\mathbb{X}))$ under the weak convergence topology, where, under this topology, $\mathcal{Z}$ is also a standard Borel space, that is, $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ is separable and completely metrizable under the weak convergence topology. The filter process has state space $\mathcal{Z}$ and action space $\mathbb{U}$. Let $\mathcal{P}(\mathcal{Z})$ denote the probability measures on $\mathcal{Z}$, equipped with the weak convergence topology.

The transition probability η of the filter process can be determined via the following equation [14, 27, 35]:

$$\eta(\cdot \mid z, u) = \int_{\mathbb{Y}} 1_{\{F(z,u,y) \in \cdot\}} P(dy \mid z, u), \tag{1.2}$$

where

$$P(\cdot \mid z, u) = \Pr\{Y_{n+1} \in \cdot \mid Z_n = z, U_n = u\}$$

from $\mathcal{Z} \times \mathbb{U}$ to $\mathbb{Y}$ and

$$F(z, u, y) := \Pr\{X_{n+1} \in \cdot \mid Z_n = z, U_n = u, Y_{n+1} = y\}$$

from $\mathcal{Z} \times \mathbb{U} \times \mathbb{Y}$ to $\mathcal{Z}$.

The one-stage cost function $\tilde{c} : \mathcal{Z} \times \mathbb{U} \to \mathbb{R}$ is a Borel measurable function and is given by

$$\tilde{c}(z, u) := \int_{\mathbb{X}} c(x, u) z(dx), \tag{1.3}$$

where $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is the stagewise cost function.

This way, we obtain a completely observable Markov decision process from the POMDP, with the components $(\mathcal{Z}, \mathbb{U}, \tilde{c}, \eta)$. The resulting MDP is often referred to as the belief MDP.

Convergence notions for probability measures. Let $\{\mu_n, n \in \mathbb{N}\}$ be a sequence in $\mathcal{P}(\mathbb{X})$. The sequence $\{\mu_n\}$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ *weakly* if

$$\int_{\mathbb{X}} f(x)\mu_n(dx) \to \int_{\mathbb{X}} f(x)\mu(dx) \tag{1.4}$$

for every continuous and bounded $f : \mathbb{X} \to \mathbb{R}$.

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

$$\begin{aligned} \|\mu - \nu\|_{TV} &:= 2 \sup_{B \in \mathcal{B}(\mathbb{X})} |\mu(B) - \nu(B)| \\ &= \sup_{f:\, \|f\|_\infty \leq 1} \left| \int f(x)\mu(dx) - \int f(x)\nu(dx) \right|, \end{aligned} \tag{1.5}$$

where the supremum is over all measurable real f such that $\|f\|_\infty = \sup_{x\in\mathbb{X}} |f(x)| \leq 1$.

Finally, the bounded-Lipschitz metric ρ_{BL} [31, p. 109] can also be used to metrize weak convergence as

$$\rho_{BL}(\mu, \nu) = \sup_{\|f\|_{BL} \leq 1} \left| \int_{\mathbb{X}} f(x)\mu(dx) - \int_{\mathbb{X}} f(x)\nu(dx) \right|, \tag{1.6}$$

where

$$\|f\|_{BL} := \|f\|_\infty + \|f\|_L, \quad \|f\|_L = \sup_{x \neq x'} \frac{f(x) - f(x')}{d_{\mathbb{X}}(x, x')},$$

and d is the metric on $\mathbb{X}$.

When $\mathbb{X}$ is compact, one way to metrize $\mathcal{Z}$ under the weak convergence topology is via the Kantorovich–Rubinstein metric (also known as the Wasserstein metric of order 1) ([1, Theorem 8.3.2]) defined as follows:

$$W_1(\mu, \nu) := \sup \left\{ \int_{\mathbb{X}} f(x)\mu(dx) - \int_{\mathbb{X}} f(x)\nu(dx) : f \in \mathrm{Lip}(\mathbb{X}, 1) \right\}, \tag{1.7}$$

$\mu, \nu \in \mathcal{Z}$, where for $k \in \mathbb{N}$,

$$\mathrm{Lip}(\mathbb{X}, k) = \{f : \mathbb{X} \to \mathbb{R}, \|f\|_L \leq k\}.$$

Definition 1.1 ([8, eq. (1.16)]). *For a kernel operator $K : S_1 \to \mathcal{P}(S_2)$ we define the Dobrushin coefficient as*

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

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

1.2. Average cost optimality equation. The average cost optimality equation (ACOE) plays a crucial role in the analysis and the existence results of MDPs under the infinite horizon average cost optimality criteria. In the framework of the belief MDP noted above, the triplet (h, ρ^*, γ^*), where $h : \mathcal{Z} \to \mathbb{R}$, $\gamma^* : \mathcal{Z} \to \mathbb{U}$ are measurable functions and $\rho^* \in \mathbb{R}$ is a constant, forms the ACOE if

$$\begin{aligned} h(z) + \rho^* &= \inf_{u \in \mathbb{U}} \left\{ \tilde{c}(z, u) + \int h(z_1)\eta(dz_1|z, u) \right\} \\ &= \tilde{c}(z, \gamma^*(z)) + \int h(z_1)\eta(dz_1|z, \gamma^*(z)) \end{aligned} \tag{1.8}$$

for all $z \in \mathcal{Z}$. It is well known that (see, e.g., [14, Theorem 5.2.4]) if (1.8) is satisfied with the triplet (h, ρ^*, γ^*), and, furthermore, if h satisfies

$$\sup_{\gamma \in \Gamma} \lim_{t \to \infty} \frac{E_z^\gamma [h(Z_t)]}{t} = 0 \quad \forall z \in \mathcal{Z}, \tag{1.9}$$

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

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

We will refer to the function h as the relative value function in the rest of the paper. Note that there may not be a unique relative value function h that satisfies the ACOE; however, any h that satisfies the ACOE and the condition (1.9) can be used for optimality analysis.

1.3. Statement of the main result. Now we state the main result of our paper.

Assumption 1.

1. U is a compact space, and $(\mathbb{X}, d)$ is a 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 continuous and thus bounded since $\mathbb{X}$ is assumed to be compact.
6. $$K_2 := \frac{\alpha D(3 - 2\delta(Q))}{2} < 1,$$
where $\delta(Q)$ is defined as in Definition 1.1.

We thus state the following main theorem.

THEOREM 1.2. *Under Assumption* 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.*

As can be seen, the testability/verification of these criteria is explicit. We provide three examples to illustrate this, one in a discrete setting and the others in a continuous setting.

As we will discuss in more detail in section 3, the theorem provides significant implications for robustness and approximations. By utilizing this theorem, we can derive near-optimal policies for the average cost criteria in partially observable Markov decision processes. To the best of our knowledge, obtaining near-optimal policies for the average cost criteria is a novel result in the literature on partially observable Markov decision processes (POMDPs). Additionally, under certain conditions, our findings contribute a new perspective to the literature on robustness, demonstrating that the average cost optimization problem is completely robust to initialization errors.

1.4. Examples. The first example will be for the discrete case. For the case with finite $\mathbb{X}$, consider the discrete metric d defined as follows:

$$d(x,x') = \begin{cases} 1 & \text{if } x \neq x', \\ 0 & \text{if } x = x'. \end{cases}$$

With this choice of metric, the diameter D is equal to 1.

Example 1.1. Let $\mathbb{X} = \{0,1,2,3\}$, $\mathbb{Y} = \{0,1\}$, $\mathbb{U} = \{0,1\}$, $\epsilon \in (0,1/2)$, and let c be any function from $X \times U$ to R^+. Now, consider the transition and measurement matrices given by

$$\mathcal{T}_0 = \begin{pmatrix} 1/2 & 1/3 & 1/6 & 0 \\ 0 & 1/2 & 1/6 & 1/3 \\ 1/2 & 1/6 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 3/4-\epsilon & 1/4+\epsilon \\ 3/4-\epsilon & 1/4+\epsilon \\ 1/4+\epsilon & 3/4-\epsilon \\ 1/4+\epsilon & 3/4-\epsilon \end{pmatrix},$$

$$\mathcal{T}_1 = \begin{pmatrix} 1/3 & 1/2 & 1/6 & 0 \\ 0 & 1/3 & 1/2 & 1/6 \\ 1/2 & 1/3 & 0 & 1/6 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}.$$

For this example, please note that $\delta(Q)$ is greater than 1/2, and the diameter D is equal to 1. We can choose α to be 1. Hence, according to Theorem 1.2, the average cost optimality equation (ACOE) has a solution.

Alternatively, if we were to select α to be less than 1, a more relaxed condition on $\delta(Q)$ would suffice. If we can choose α to be less than 2/3, without any other constraints on the observation kernel, we can assert that the ACOE has a solution. As we will see in the literature review section, these conditions are new.

The second example will be for the continuous state space case.

Example 1.2. Let $\mathbb{X} = [0,2]$, $\mathbb{Y} = \{0,1\}$, and $\mathbb{U} = [0,12]$; let the transition kernel $\mathcal{T}(.\,|\,x,u) = \text{Unif}(0, \min(2, 1+((x+u)/7)))$, where Unif stands for uniform distribution; let measurement kernel $Q(x) = \lfloor x \rfloor$; and let cost function $c(x,u) = x+u$.

For this example, we observe that $\delta(Q) = 0$. Furthermore, considering any control input u in the set $\mathbb{U}$ and any states x and x' within the space $\mathbb{X}$ such that $x' < x$, we can derive the following bound:

$$\begin{aligned} \|\mathcal{T}(\cdot|x,u) - \mathcal{T}(\cdot|x',u)\|_{TV} &\leq \|\text{Unif}(0,1+(x'+u)/7) - \text{Unif}(0,1+(x+u)/7)\|_{TV} \\ &= 2\left(1 - \frac{1+\frac{x'+u}{7}}{1+\frac{x+u}{7}}\right) = 2\frac{x-x'}{x+u+7} \\ &< \frac{2}{7} d(x,x'). \end{aligned}$$

We have $D = \sup d(x,x') = 2$, and by choosing $\alpha = \frac{2}{7}$ and $K_1 = 1$, we can set K_2 to 6/7. This choice ensures that we satisfy the conditions outlined in Assumption 1. Consequently, we can directly apply Theorem 1.2 to this specific example.

Example 1.3. Consider $\mathbb{X} = [0,1]$, $\mathbb{U} = [-p,p]$, and the cost function $c(x,u) = x-u$. Define the transition kernel $\mathcal{T}(.|x,u) = \bar{N}(x+u,\sigma^2)$. Here, $\bar{N}(\mu,\sigma^2)$ denotes the truncated version of $N(\mu,\sigma^2)$, where the support is restricted to $[0,1]\subset\mathbb{R}$. Its probability density function f is given by

$$f(x;\mu,\sigma)=\frac{1}{\sigma}\frac{\varphi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{1-\mu}{\sigma}\right)-\Phi\left(\frac{0-\mu}{\sigma}\right)}.$$

Here, $\varphi(\cdot)$ is the probability density function of the standard normal distribution, and $\Phi(\cdot)$ is its cumulative distribution function.

For any $0\le x<y\le 1$, $\frac{\|\mathcal{T}(.|y,u)-\mathcal{T}(.|x,u)\|_{TV}}{y-x}\le\frac{\sqrt{2}}{\sigma\sqrt{\pi}}$. This shows that the transition kernel $\mathcal{T}$ satisfies Assumption 1(3) with $\alpha=\frac{\sqrt{2}}{\sigma\sqrt{\pi}}$. If we choose $\sigma>3/\sqrt{2\pi}$, Assumption 1(6) is satisfied for every observation channel. Under this condition, by Theorem 1.2, the ACOE has a solution.

1.5. Literature review and comparison. In the following we present a detailed literature review and also provide a comparison with our results. We will in particularly highlight the fact that many of the results in the literature are not easy to verify, which also serves to demonstrate that the problem is a challenging one.

References [26, 12, 28, 16] study the average cost control problem under the assumption that the state space is finite; they provide reachability type conditions for the belief kernels. Reference [16] also provides a detailed literature review. An additional line of argument, via simulation and coupling, was introduced in [2, 3, 5, 4]. Under certain strong continuity requirements of the transition kernel, [30] extends the positivity condition to Polish spaces. We also note that, via the weak Feller property of nonlinear filters, the convex analytic methods can also be utilized [33, Theorem 1.2], though the dependence on the initial condition is a limitation.

Borkar and Budhiraja [5] consider $\mathbb{X}$, $\mathbb{Y}$, and $\mathbb{U}$ as Polish spaces, where $\mathbb{X}$ is a finite-dimensional Euclidean space (not necessarily compact, unlike ours) and $\mathbb{U}$ is a compact space. In their setup, unlike ours, the observation measurement depends on the previous time's state and action. For $(x,u)\in\mathbb{X}\times\mathbb{U}$, the transition probability measure $p(x,u,dz,dy)\in\mathcal{P}(\mathbb{X}\times\mathbb{Y})$ is defined. Reference [5] assumes that p is continuous in a strong sense as follows: Let λ denote the Lebesgue measure on $\mathbb{X}$, and let $\eta\in\mathcal{P}(\mathbb{Y})$ assume a density function $\varphi(x,u,z,y)$ on $\mathbb{X}\times\mathbb{U}\times\mathbb{X}\times\mathbb{Y}$ such that $p(x,u,dz,dy)=\varphi(x,u,z,y)\lambda(dz)\eta(dy)$, with $\varphi(\cdot)>0$. It is assumed that $\varphi(x,u,z,y)$ is continuous. Note that by Scheffé's lemma, this implies that the transition kernel, as well as the measurement kernel, is continuous in total variation. The update rule is then given as

$$Pr\left(X_{n+1}\in A,Y_{n+1}\in A'\mid X_{[0,n]},Y_{[0,n]},U_{[0,n]}\right)=\int_{A'}\int_{A}\varphi\left(X_n,U_n,z,y\right)\lambda(\mathrm{d}z)\eta(\mathrm{d}y).$$

Furthermore, to facilitate a coupling argument, [5] additionally assumes

$$\int\bar{\varphi}\left(x,u,x'\right)\left(\frac{\varphi\left(x,u,x',y\right)}{\bar{\varphi}\left(x,u,x'\right)}\right)^{1+\varepsilon_0}\lambda\left(dx'\right)\eta(dy)<\infty \tag{1.10}$$

for some $\varepsilon_0>0$, where $\bar{\varphi}(x,u,x')=\int\varphi(x,u,x',y)\eta(dy)$. Finally, the following Lyapunov type assumption is assumed (which always holds when $\mathbb{X}$ is compact).

Assumption 2. There exist $\mathcal{V}\in\mathbb{C}(\mathbb{X})$ and $\hat{\mathcal{V}}\in\mathbb{C}(\mathbb{X})$ satisfying $\lim_{\|x\|\to\infty}\mathcal{V}(x)=\infty$ and $\lim_{\|x\|\to\infty}\hat{\mathcal{V}}(x)=\infty$, and under any wide-sense admissible $\{Z_n\}$,

$$E\left[\hat{\mathcal{V}}\left(X_{n+1}\right)\mid\mathbb{F}_n\right]-\hat{\mathcal{V}}\left(X_n\right)\le-\mathcal{V}\left(X_n\right)+\hat{C}I_{\hat{B}}\left(X_n\right),$$

$$\limsup_{n\to\infty} \frac{E\left[\hat{\mathcal{V}}(X_n)\right]}{n} = 0,$$

where $\hat{C} > 0$, $\hat{B} = \{x \in S : \|x\| \leq \hat{R}\}$ for some $\hat{R} > 0$, and $\mathbb{F}_n = \sigma(X_{[0,n]}, Y_{[0,n]}, U_{[0,n]})$.

THEOREM 1.3 (Theorem 4.1 of [5]). *For the system described above and under the mentioned assumptions as well as Assumption* 2*, a solution to the average cost optimality equation (ACOE) exists. This implies the existence of an optimal control policy, and the optimal cost remains constant for every initial state.*

Compared to this result, we present complementary conditions which are explicit and testable. Additionally, for the compact setup, we do not have continuity assumptions on the measurement kernels, and the condition (1.10) is not needed. Our paper looks to be the first to utilize contraction properties of filter kernels.

In [26], Platzman addresses the finite state average cost control problem with finite state, observation, and action spaces under restrictive reachability, subrectangularity, and detectability conditions.

Runggaldier and Stettner in [28] consider $\mathbb{X}$ and $\mathbb{Y}$ as finite and $\mathbb{U}$ as compact. They prove that under the following positivity condition [28], ACOE has a bounded solution.

Assumption 3 (see [28]).

$$\inf_{i,j\in\mathbb{X}} \inf_{u,u'\in\mathbb{U}} \inf_{\{C\in\mathbb{B}(\mathbb{X}):[\mathcal{T}(C|i,u)]>0\}} \frac{\mathcal{T}(C|j,u')}{T(C|i,u)} > 0.$$

We note that Example 1.1 does not meet this condition.

Borkar [2] employs a coupling argument under the following assumption.

Assumption 4 ([2, Assumption A]). There exist constants $K_0 \in \mathbb{R}^+$ and $\delta \in (0,1)$ such that $\sup_{i,j}\sup_\gamma P(\tau > n|X'_0 = i, X_0 = j) \leq K_0\delta^n$ holds for all $n \geq 0$. This supremum is taken over all wide-sense admissible policies, and τ represents the coupling time, i.e., $\tau = \min\{n : X'_n = X_n\}$.

However, verifying this assumption is not simple due to the necessity of taking the supremum over all wide-sense admissible policies (which is a strict superset of deterministic admissible policies).

Hsu, Chuang, and Arapostathis [16] consider a finite $\mathbb{X}$ and $\mathbb{Y}$ and a compact $\mathbb{U}$. They provide two different sets of assumptions under which ACOE has a solution.

Assumption 5 (Assumption 2 of [16]). $\mathcal{Z}_\epsilon = \{\mu \in \mathcal{Z} : \mu(x) > \epsilon \text{ for all } x \in \mathbb{X}\}$. There exist constants $\epsilon > 0$, $k_0 \in \mathbb{N}$, and $\alpha < 1$ such that if $z_*(\beta) \in \arg\min_{z\in\mathcal{Z}} J^*_\beta(z)$, then for each $\beta \in [\alpha, 1)$ we have

$$\max_{1\leq k\leq k_0} \mathbb{P}^{\gamma_\beta}_{z_*(\beta)}(Z_k \in \mathcal{Z}_\epsilon) \geq \epsilon,$$

where Z_k is the filter process, and γ_β is the optimal policy for the β-discounted horizon cost problem.

Because checking this assumption can be quite complex, an alternative, easier-to-verify assumption is provided.

Assumption 6 (see [16]). There exist $k \geq 1$ and $\Delta > 0$ such that, for all admissible $\{U_t\}$,

$$\mathbb{P}(X_k = j|X_0 = i, U_{t-1}, Y_t, 1 \leq t \leq k) \geq \Delta \quad \forall i,j \in \mathbb{X}.$$

For the second set of assumptions, the conditions are relaxed to some extent when the cost function c is continuous.

Assumption 7 (see [16]). There exist $k \geq 1$ and $\Delta > 0$ such that, for all $y^k \in \mathbb{Y}^k$ and $u^k \in \mathbb{U}^k$,

$$P_{ij}\left(y^k|u^k\right) \geq \Delta \sum_{\ell \in \mathbb{X}} P_{\ell j}\left(y^k|u^k\right) \quad \forall i,j \in \mathbb{X},$$

where $P_{ij}\left(y^k \mid u^k\right) = \mathbb{P}\left(X_k = j, Y_{[1,k]} = y^k \mid X_0 = i, U_{[0,k-1]} = u^k\right)$.

THEOREM 1.4 (Theorem 11 of [16]). *Under either Assumption* 5 *or Assumption* 7*, the average cost optimality equation (ACOE) possesses a bounded solution.*

Finally, Stettner in [30] extends the results of [28] to Polish state spaces and observation spaces, encompassing both nondegenerate and degenerate observations as follows: For $u_1, u_2 \in \mathbb{U}$ and $\mu, \nu \in P(\mathbb{X})$ define

$$\lambda(u_1, u_2, \mu, \nu) := \inf_{\{C : \mathcal{T}(C|\mu, u_1) > 0\}} \frac{\mathcal{T}(C|\nu, u_2)}{\mathcal{T}(C|\mu, u_1)}.$$

Then define $\lambda(\mu, \nu) = \inf_{u \in U} \lambda(u, u, \mu, \nu)$.

Assumption 8 (see [30]). $\lambda\left(u^1, u_n^2, \mu, \nu_n\right) \to 1$ and $\lambda\left(u_n^2, u^1, \nu_n, \mu\right) \to 1$ when $u_n^2 \to u^1$ and $\nu_n \Rightarrow \mu$. Similarly, assume that both $\lambda(\nu_n, \mu)$ and $\lambda(\mu, \nu_n)$ converge to 1 as $\nu_n \Rightarrow \mu$ (where $\Rightarrow$ denotes weak convergence of probability measures).

THEOREM 1.5 (Theorem 5.6 of [30]). *Let* $\mathbb{X}$ *and* $\mathbb{Y}$ *be Polish spaces,* $\mathbb{U}$ *be a compact space, and* c *be a continuous and bounded function. Under Assumptions* 3 *and* 8*, the average cost optimality equation (ACOE) admits a bounded solution.*

In [32], the authors study near optimality of finite window policies for average cost problems where 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 [11] (which becomes consequential by the equality assumption) and the finiteness of the state space is crucial. The paper shows that for any given $\epsilon > 0$, there exists an ϵ-optimal finite window policy.

With this review, we have both summarized some existing key studies and highlighted the fact that our paper presents accessible and testable conditions, compared with most of the literature reviewed above.

In our paper, we present a contraction based analysis for nonlinear filters, which is a novel contribution and which we expect will have significant consequences for learning theoretic and approximation results.

2. Proof of the main theorem.

Recently, [22] presented the following regularity results for controlled filter processes.

THEOREM 2.1 ([22, Theorem 7-i, Theorem 7-iv]). *Assume that* $\mathbb{X}$ *and* $\mathbb{Y}$ *are Polish spaces.*

i. *If Assumption* 1(3) *is fulfilled, then we have*

$$\rho_{BL}\left(\eta(\cdot \mid z, u), \eta(\cdot \mid z', u)\right) \leq 3(1+\alpha)\,\rho_{BL}(z, z')$$

for any $z, z' \in \mathcal{P}(\mathbb{X})$ *and* $u \in \mathbb{U}$.

ii. *Without any assumptions,*

$$\rho_{BL}\left(\eta(\cdot|z,u),\eta(\cdot|z',u)\right) \leq (3-2\delta(Q))(1-\tilde{\delta}(\mathcal{T}))\|z-z'\|_{TV}$$

for any $z,z' \in \mathcal{P}(\mathbb{X})$ *and* $u \in \mathbb{U}$*, where* $\tilde{\delta}(\mathcal{T}) := \inf_{u\in\mathbb{U}} \delta(\mathcal{T}(\cdot\mid\cdot,u))$.

Building on these results and their proof method, the next result follows from [7, Theorem 5.1], which considered the control-free setup. A proof sketch is given in Appendix A.

THEOREM 2.2. *Assume that* $\mathbb{X}$ *and* $\mathbb{Y}$ *are Polish spaces. If Assumption* 1(1) *and Assumption* 1(3) *are fulfilled, then we have*

$$W_1\left(\eta(\cdot\mid z_0,u),\eta\left(\cdot\mid z_0',u\right)\right) \leq \left(\frac{\alpha D(3-2\delta(Q))}{2}\right) W_1\left(z_0,z_0'\right)$$

for all $z_0,{z_0}' \in \mathcal{Z}$*,* $u\in\mathbb{U}$.

As we assume that $\mathbb{X}$ is compact, under the W_1 (that is, the 1-Wasserstein or Kantorovich–Rubinstein) metric, $\mathcal{Z}$ is compact.

Under Assumption 1(4), we have that $\tilde{c}$ defined in (1.3) is Lipschitz continuous, since

$$|\tilde{c}(z,u)-\tilde{c}(z',u)| = \left|\int_{\mathbb{X}} c(x,u)z(dx) - \int_{\mathbb{X}} c(x,u)z'(dx)\right| \leq K_1 W_1\left(z,z'\right). \tag{2.1}$$

THEOREM 2.3 ([17, Theorem 2]). *Under Assumption* 1(2)*, the transition probability* $\eta(\cdot\mid z,u)$ *of the filter process is weakly continuous in* (z,u).

We also highlight that additional recent results are available on the weak Feller property, such as those in [10]; however, as there are no restrictions on Q, this result [17, Theorem 2] is relevant here.

LEMMA 2.4. *Under Assumption* 1*, for any* β *the value function* J_β^* *is Lipschitz continuous with coefficient* K*, where* $K=\frac{K_1}{1-\beta K_2}$.

Proof. We adopt the approach in the proof of [29, Theorem 4.37]. Let $f \in \mathrm{Lip}(\mathcal{Z},k)$ for some $k>0$. Then $g=\frac{f}{k}\in \mathrm{Lip}(\mathcal{Z},1)$, and therefore for all $u\in \mathrm{U}$ and $z,y\in\mathcal{Z}$ we have

$$\left|\int_{\mathcal{Z}} f(x)\eta(dx\mid z,u) - \int_{\mathcal{Z}} f(x)\eta(dx\mid y,u)\right| \tag{2.2}$$

$$= k\left|\int_{\mathcal{Z}} g(x)\eta(dx\mid z,u) - \int_{\mathcal{Z}} g(x)\eta(dx\mid y,u)\right| \tag{2.3}$$

$$\leq kW_1(\eta(\cdot\mid z,u),\eta(\cdot\mid y,u)) \leq kK_2W_1(z,y) \tag{2.4}$$

by Theorem 2.2. Let T be the Bellman optimality operator,

$$(Tf)(x) = \min_{u\in U}\left\{c(x,u)+\beta\int_{y\in\mathcal{Z}} \eta(dy|x,u)f(y)\right\}.$$

We know that J_β^* satisfies $TJ_\beta^* = J_\beta^*$ [14, Lemma 4.2.6]. T is a contraction, so by the Banach fixed point theorem, T^nf converges to J_β^* as follows:

$$\begin{aligned}
&|\,Tf(z)-Tf(y)\,|\\
&\quad\leq \max_{u\in\mathrm{U}}\left\{|\tilde{c}(z,u)-\tilde{c}(y,u)| + \beta\left|\int_{\mathbf{Z}} f(x)\eta(dx\mid z,u) - \int_{\mathbf{Z}} f(x)\eta(dx\mid y,u)\right|\right\}\\
&\quad\leq K_1W_1(z,y)+\beta kK_2W_1(z,y) = (K_1+\beta kK_2)\,W_1(z,y) =: M_1W_1(z,y).
\end{aligned}$$

By induction we have, for all $n \geq 2$,

$$T^n f \in \mathrm{Lip}\left(\mathcal{Z}, M_n\right),$$

where $M_n = K_1 + \beta K_2 M_{n-1}$ and thus $M_n = K_1 \sum_{i=0}^{n-1} \left(\beta K_2\right)^i + k\left(\beta K_2\right)^n$. Then, the sequence M_n monotonically converges to $\frac{K_1}{1-\beta K_2}$ since $K_2 < 1$. Hence, $T^n f \in \mathrm{Lip}(\mathcal{Z}, \frac{K_1}{1-\beta K_2})$ for all n, and therefore, $J_\beta^* \in \mathrm{Lip}(\mathcal{Z}, \frac{K_1}{1-\beta K_2})$ since it is closed with respect to the sup-norm. Taking $k \leq \frac{K_1}{1-\beta K_2}$, we certify that the fixed point satisfies the desired Lipschitz continuity. ◻

For any $\beta \in (0,1)$, $J_\beta^* \in \mathrm{Lip}(\mathcal{Z}, \frac{K_1}{1-K_2})$. Therefore, for any $z_0 \in \mathcal{Z}$,

$$h_\beta(z) = J_\beta^*(z) - J_\beta^*(z_0) \leq \frac{K_1}{1-K_2} W_1(z, z_0) \leq \frac{K_1 \cdot D}{1-K_2}. \tag{2.5}$$

In view of the results above, we now introduce a crucial auxiliary result that will play a pivotal role in establishing our main result.

Assumption 9.

1. The one-stage cost function $\tilde{c}$ is bounded and continuous.
2. The stochastic kernel $\eta(\cdot \mid x, u)$ is weakly continuous in $(x,u) \in \mathcal{Z} \times \mathbb{U}$, i.e., if $(x_k, u_k) \to (x, u)$, then $\eta\left(\cdot \mid x_k, u_k\right) \to \eta(\cdot \mid x, u)$ weakly.
3. $\mathbb{U}$ is compact.
4. $\mathcal{Z}$ is σ-compact, that is, $\mathcal{Z} = \cup_n S_n$ where $S_n \subset S_{n+1}$ and each S_n is compact.

There exist $\alpha \in (0,1)$ and $N \geq 0$ and a state $z_0 \in \mathcal{Z}$ such that

5. $-N \leq h_\beta(z) \leq N$ for all $z \in \mathcal{Z}$ and $\beta \in [\alpha, 1)$, where

$$h_\beta(z) = J_\beta^*(z) - J_\beta^*\left(z_0\right). \tag{2.6}$$

6. The sequence $\left\{h_{\beta(k)}\right\}$ is equicontinuous, where $\{\beta(k)\}$ is a sequence of discount factors converging to 1, which satisfies $\lim_{k\to\infty}(1-\beta(k)) J_{\beta(k)}^*(z) = \rho^*$ for all $z \in \mathcal{Z}$ for some $\rho^* \in [0, L]$.

LEMMA 2.5 ([34, Theorem 7.3.3]). *Under Assumption* 9*, a solution to the average cost optimality equation (ACOE) exists, leading to the existence of an optimal control policy, and optimal cost is constant for every initial state.*

Proof. See Appendix B. ◻

We note that a similar result, with slightly stronger continuity conditions (not applicable to our setup) under [14, Assumption 4.2.1], can be found in [14, Theorem 5.5.4]. Using this auxiliary result, we are now ready to present the proof of the main theorem.

Proof of Theorem 1.2. First, we will show that if a partially observable Markov decision process (POMDP) satisfies Assumption 1, then the corresponding belief MDP satisfies Assumption 9.

Assumption 9(1) is valid due to equation (2.1). Assumption 9(3)(4) hold because $\mathbb{U}$ and $\mathcal{Z}$ are compact. Assumption 9(2) follows from Theorem 2.3, and Assumption 9(5) follows from (2.5).

By inequality (2.5), we know that

$$h_\beta(z) = J_\beta^*(z) - J_\beta^*(z_0) \leq \frac{K_1}{1-K_2} W_1(z, z_0)$$

for all $\beta \in (0,1)$. Therefore, h_β is equicontinuous for all $\beta \in (0,1)$. Since this condition holds for all subsequences, Assumption 9(6) is satisfied.

Thus, if a POMDP satisfies Assumption 1, then the belief MDP satisfies Assumption 9. All conditions of Lemma 2.5 are satisfied, and the proof is completed using Lemma 2.5. ∎

3. Implications for approximations, robustness, and learning.

In this section, we discuss several implications of our results on the existence of a solution to the ACOE and Wasserstein regularity.

First, we establish robustness by studying how an optimal policy developed for an incorrect initial prior performs when applied to the correct initial distribution in the context of the average cost criteria. Under certain conditions, we find that the average cost optimization problem is robust to errors in initialization.

We then study approximate optimality through the quantization of the belief space, or $\mathcal{P}(\mathbb{X})$. This method yields a near-optimal policy for the original problem. We then present an alternate method to construct a finite Markov decision process (MDP). This is achieved by replacing the complete observable Markov process with the most recent N observations and actions. Within this finite state MDP framework, we apply Q-learning to obtain a near optimal policy.

3.1. Robustness to incorrect priors for discounted and average cost criteria.

In this section, we discuss the implications of our results on robustness. Let us first formally define the robustness problem. We have already introduced J^* and J^*_β in the first section.

An optimal control policy γ^μ for a given prior μ is a policy that achieves the lowest expected cost over all admissible control policies as follows:

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

Consider a scenario where a controller incorrectly assumes that the system's prior is ν, while in reality it is μ. In this case, the controller would implement the policy γ^ν, optimal for ν, but this results in an expected cost of $J(\mu, \gamma^\nu)$. If the controller had used the correct policy, the cost could have been $J^*(\mu)$. In studying robustness, we are interested in the differences

$$J_\beta\left(\mu, \gamma^\nu_\beta\right) - J^*_\beta(\mu),$$
$$J(\mu, \gamma^\nu) - J^*(\mu)$$

for the discounted cost problem and the average cost problem. Here γ^ν_β is an optimal policy for the initial ν in the context of discounted cost criteria, while γ^ν is an optimal policy for the initial ν in the context of average cost criteria.

In the literature on POMDPs, there are upper bounds on this error cost for discounted cost problem [19], studies on its relationship with filter stability, and attempts to find uniform upper bounds for average cost problem [25]. There is also work considering how robustness is affected under different transition kernels for discounted cost problems [20]. The following builds on [19, Theorem 3.2].

THEOREM 3.1 ([25, Theorem 3.8], [19, Theorem 3.2]). *Assume the cost function c is bounded, nonnegative, and measurable. Let γ^ν be the optimal control policy designed with respect to a prior ν. Then we have*

$$J(\mu, \gamma^\nu) - J^*(\mu) \le 2\|c\|_\infty \|\mu - \nu\|_{TV}.$$

However, by utilizing filter stability, the bound can be refined, as we will observe in the following.

DEFINITION 3.1. *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 $\lim_{n\to\infty} E^{\mu,\gamma}[\|\pi_n^{\mu,\gamma} - \pi_n^{\nu,\gamma}\|_{TV}] = 0$.*

The filter is universally stable in total variation in expectation if it holds with respect to every admissible policy $\gamma \in \Gamma$.

For the average cost criteria, under Assumption 1, asymptotic filter stability is sufficient for robustness, as the following shows.

THEOREM 3.2 ([25, Theorem 3.9]). *Assume the cost function c is bounded, nonnegative, and measurable, and assume the filter is universally stable in total variation in expectation. Consider the span seminorm,*

$$\|J^*\|_{sp} := \sup_{\mu_1 \in \mathcal{P}(\mathcal{X})} J^*(\mu_1) - \inf_{\mu_2 \in \mathcal{P}(\mathcal{X})} J^*(\mu_2);$$

then we have

$$J(\mu, \gamma^{\nu}) - J^*(\mu) \le \|J^*\|_{sp}.$$

In particular, if $\|J^*\|_{sp} = 0$, then the average cost optimization problem is completely robust to initialization errors.

We note that in [25], the existence of an optimal solution was not shown; our implication also presents an existence result for optimal policies. Under Assumption 1, we show that $\|J^*\|_{sp} = 0$ (Theorem 1.2), meaning that if the filter is universally stable in total variation in expectation, then the average cost optimization problem is completely robust to initialization errors.

COROLLARY 3.3. *Under Assumption 1 and assuming universal filter stability in total variation in expectation, we have that*

$$J(\mu, \gamma^{\nu}) = J^*(\mu) \quad \forall \mu, \nu \in \mathcal{Z}.$$

Note 3.1. As mentioned in Note 3.2, [24, Theorem 4.1] demonstrates that if the condition $(1 - \tilde{\delta}(\mathcal{T}))(2 - \delta(Q)) < 1$ is met, then the filter is universally stable in total variation in expectation.

In addition to our results on the average cost optimality equation, our analysis is also consequential for the discounted cost criterion and offers an improvement on [25, Theorem 3.10]. Utilizing the span seminorm bound derived in our analysis of the average cost optimality equation, and by modifying the proof of [25, Theorem 3.10] along with individually bounding equations (1.8), (1.9), and (1.10) from the same source, we arrive at the following theorem.

THEOREM 3.4. *Under Assumption 1 and the condition $\bar{\alpha} := (1 - \tilde{\delta}(T))(2 - \delta(Q)) < 1$, we have the following robustness bound:*

$$J_\beta(\mu, \gamma_\beta^{\nu}) - J_\beta^*(\mu) \le \inf_{n\in\mathbb{N}} \left(\|c\|_\infty \frac{(1-\beta^n)}{1-\beta} + \beta^n \frac{DK_1}{1-K_2\beta} + 4\frac{\|c\|_\infty}{1-\beta}(\bar{\alpha}\beta)^n \right).$$

Proof. A brief proof is provided here.

(3.1)

$$\begin{aligned}
&J_\beta\left(\mu,\gamma_\beta^\nu\right) - J_\beta^*(\mu)\\
&\quad = E^{\mu,\gamma_\beta^\nu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right]\\
&\qquad + \beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta\left(\pi_{n-}^{\mu,\gamma_\beta^\nu},\gamma_\beta^{\nu'}\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\mu}\right)\right]\right)\\
&\quad = E^{\mu,\gamma_\beta^\nu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right]\\
&\qquad + \beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta\left(\pi_{n-}^{\mu,\gamma_\beta^\nu},\gamma_\beta^{\nu'}\right) + J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right) - J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\mu}\right)\right]\right)\\
&\quad = E^{\mu,\gamma_\beta^\nu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[\sum_{i=0}^{n-1}\beta^i c\left(X_i,U_i\right)\right]
\end{aligned}$$

(3.2)

$$\qquad + \beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\mu}\right)\right]\right)$$

(3.3)

$$\qquad + \beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta\left(\pi_{n-}^{\mu,\gamma_\beta^\nu},\gamma_\beta^{\nu'}\right) - J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right)\right]\right),$$

where $\nu' = \pi_{n-}^{\nu,\gamma_\beta^\nu}$, and $\gamma_\beta^{\nu'}$ is an optimal discounted policy for ν'. As in [25], we can refer to these three components as transient cost, strategic measure cost, and approximation cost, respectively. The transient cost (3.1) is upper bound by

$$E^{\mu,\gamma_\beta^\nu}\left[\sum_{i=0}^{n-1}\beta^i c\left(x_i,u_i\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[\sum_{i=0}^{n-1}\beta^i c\left(x_i,u_i\right)\right] \le \|c\|_\infty \sum_{i=0}^{n-1}\beta^i = \|c\|_\infty\left(\frac{1-\beta^n}{1-\beta}\right).$$

The strategic measure cost (3.2) is upper bound by

$$\beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right)\right] - E^{\mu,\gamma_\beta^\mu}\left[J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\mu}\right)\right]\right) \le \beta^n\left\|J_\beta^*\right\|_{sp} \le \beta^n\frac{K_1}{1-\beta K_2}D$$

because of Lemma 2.4. The approximation cost (3.3) is upper bound by

$$\beta^n\left(E^{\mu,\gamma_\beta^\nu}\left[J_\beta\left(\pi_{n-}^{\mu,\gamma_\beta^\nu},\gamma_\beta^{\pi_{n-}^{\nu,\gamma_\beta^\nu}}\right) - J_\beta^*\left(\pi_{n-}^{\mu,\gamma_\beta^\nu}\right)\right]\right) \le 4\frac{\|c\|_\infty}{1-\beta}(\bar\alpha\beta)^n$$

following the stability criteria from [24, Theorem 4.1]. We establish the desired result. $\square$

3.2. Approximations and learning. Throughout the rest of this section, we assume that Assumption 1 holds. Earlier, we already showed that if Assumption 1 is valid for the POMDP, then Assumption 9 is applicable for the belief MDP. The following theorems assist in finding near-optimal policies for the average cost criterion; for further analysis see [6, Theorem 5] or [34, Theorem 7.3.6]. In particular, the fact that a solution to the average cost optimality equation exists and is arrived at via the vanishing discount method in our analysis earlier is critical for the applicability of the following result.

THEOREM 3.5. (a) *Under Assumption* 9, $\lim_{\beta\uparrow 1}(1-\beta)J^*_\beta(x_0)\to\rho^*$, *where* ρ^* *is the optimal average cost. Furthermore, if* γ_β *solves the discounted cost optimality equation,*

$$J^*_\beta(x)=\min_{u\in\mathbb{U}}\left\{c(x,u)+\beta\int_{\mathbb{X}}J^*_\beta(y)\mathcal{T}(dy\mid x,u)\right\},$$

then, for every $\epsilon>0$, *there exists* $\beta_\epsilon\in(0,1)$ *such that for* $\beta\in[\beta_\epsilon,1)$,

$$J\left(x,\gamma_\beta\right)-\rho^*<\epsilon.$$

This implies that the discounted cost optimal policy is near-optimal for the average cost criterion.

(b) *Under Assumption* 9, *let* β_ϵ *be chosen as above, such that with* h_β *as given in* (2.6),

$$\left|\rho^*-(1-\beta_\epsilon)\,J_{\beta_\epsilon}(x_0)\right|\le\frac{\epsilon}{2}\quad\text{and}\quad(1-\beta_\epsilon)\left\|h_{\beta_\epsilon}\right\|_\infty\le\frac{\epsilon}{2},$$

so that γ_{β_ϵ} *is* ϵ*-optimal. Suppose that* $\gamma^\delta_{\beta_\epsilon}$ *is an* δ*-optimal policy for* β_ϵ *discounted cost criteria, satisfying*

$$J_{\beta_\epsilon}\left(x,\gamma^\delta_{\beta_\epsilon}\right)-J_{\beta_\epsilon}(x)<\delta.$$

Then,

$$J\left(x,\gamma^\delta_{\beta_\epsilon}\right)-\rho^*<\epsilon+\delta.$$

That is, a near-optimal discounted cost policy is also near-optimal for the average cost criterion.

Since Assumption 1 for the POMDP implies Assumption 9 for the belief MDP, as proven in the proof of Theorem 1.2, the above theorem also holds under Assumption 1. Therefore, under Assumption 1, a near-optimal policy for the discounted cost criteria is also near-optimal for the average cost criteria. The subsequent subsections will detail how to establish a near-optimal policy for discounted costs.

We have that $\mathbb{U}$ is a compact space. As demonstrated in [29, Theorem 3.2], if the transition kernel is weakly continuous and the cost function is bounded and continuous, then the optimal policy for a quantized action model in a discounted context is also near-optimal for the original model. This allows us to consider $\mathbb{U}$ as finite to identify a near-optimal policy.

3.2.1. Near optimality of finite models via quantization for average cost. In this subsection, we focus on achieving a near-optimal policy by quantizing the states of the belief MDP, namely $P(\mathbb{X})$. We follow the approach and results from [18].

To create a new finite MDP model, we begin by quantizing the belief states. We select disjoint sets $\{Z_i\}_{i=1}^M$ such that $\bigcup_i Z_i=\mathcal{Z}$, and each Z_i is distinct 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 $\bar{Z}:=\{z_1,\dots,z_M\}$. The quantization function maps the original state space $\mathcal{Z}$ to this finite set $\bar{Z}$ as follows:

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

To define the cost function, we select a weighting measure $\pi^*\in P(\mathcal{Z})$ over $\mathcal{Z}$ such that $\pi^*(Z_i)>0$ for all Z_i. Under Assumption 1, we know that $\mathcal{Z}$ is compact under

W_1 metric. We then define normalized measures for each quantization bin Z_i using the weighting measure as

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

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

Next, we define the stagewise 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 stagewise 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 $\mathcal{Z}$ 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

$$\bar{L} := \max_{i=1,\dots,M} \sup_{z,z' \in Z_i} W_1(z,z'). \tag{3.4}$$

Assumption 10 ([18, Assumption 4]).
1. $\mathcal{Z}$ is compact.
2. There exists $\alpha_c > 0$ such that $|\tilde{c}(z,u) - \tilde{c}(z',u)| \le \alpha_c d(z,z')$ for all $z,z' \in \mathcal{Z}$ and for all $u \in \mathbb{U}$.
3. There exists $\alpha_\eta > 0$ such that $W_1(\eta(\cdot \mid z,u), \eta(\cdot \mid z',u)) \le \alpha_\eta d(z,z')$ for all $z,z' \in \mathcal{Z}$ and for all $u \in \mathbb{U}$.

The following theorem states that an optimal policy of the quantized model is near-optimal for the original model as $\bar{L} \to 0$.

THEOREM 3.6 ([18, Theorem 6]). *Under Assumption* 10*, we have*

$$\sup_{z \in \mathcal{Z}} \left| J_\beta(z,\hat{\gamma}) - J_\beta^*(z) \right| \le \frac{2\alpha_c}{(1-\beta)^2(1-\beta\alpha_\eta)} \bar{L},$$

where $\bar{L}$ is defined as in (3.4)*, and $\hat{\gamma}$ denotes the optimal policy of the finite state approximate model extended to the state space $\mathcal{Z}$ via the quantization function q.*

A similar result is presented in [29, Theorem 4.38], offering a slightly weaker bound.

Under Assumption 1, the belief MDP satisfies Assumption 10 because $\mathcal{Z}$ is compact under the W_1 metric. Due to inequality (2.1), we have $|\tilde{c}(z,u) - \tilde{c}(z',u)| \le K_1 W_1(z,z')$ for all $z,z' \in \mathcal{Z}$ and for all $u \in \mathbb{U}$. Theorem 2.2 implies $W_1(\eta(\cdot \mid z,u), \eta(\cdot \mid z',u)) \le K_2 W_1(z,z')$ for all $z,z' \in \mathcal{Z}$ and for all $u \in \mathbb{U}$. Thus, for the belief MDP, quantization provides the following bound:

$$\sup_{z \in \mathcal{Z}} \left| J_\beta(z,\hat{\gamma}) - J_\beta^*(z) \right| \le \frac{2K_1}{(1-\beta)^2(1-\beta K_2)} \bar{L}.$$

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

Q-learning. This section introduces the Q-iteration method for identifying the optimal policy for a quantized belief Markov decision process (MDP). We apply this method to the quantized belief MDP with finite state space, $\bar{Z}$, and a finite action space, $\mathbb{U}$. The Q-learning process updates Q-functions as follows: For each time step $t \geq 0$, if the current state-action pair is $(Z_t, U_t) = (z, u)$, the Q-value for this pair is updated in the following manner:

$$
\begin{aligned}
(3.5) \qquad Q_{t+1}(Z_t, U_t) &= (1 - \alpha_t(Z_t, U_t)) Q_t(Z_t, U_t) \\
&\quad + \alpha_t(Z_t, U_t) \left(c^*(Z_t, U_t) + \beta \min_{v \in \mathbb{U}} Q_t(Z_{t+1}, v) \right)
\end{aligned}
$$

Key assumptions for this Q-learning approach include the following.

Assumption 11 ([18, Assumption 5]).

- The learning rate $\alpha_t(z,u) = 0$ if $(Z_t, U_t) \neq (z,u)$; otherwise, it is defined as

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

- Under the exploration policy γ^*, Z_t is uniquely ergodic and thus has a unique invariant measure π_{γ^*}.
- During exploration, every possible state-action pair in $\bar{Z} \times \mathbb{U}$ is visited an infinite number of times.[1]

Under Assumption 11, the Q-learning algorithm (3.5) converges to the fixed point solution Q^*. A stationary policy, γ^N, that selects actions to minimize the Q-value at each state, i.e., $\gamma^N(z) \in \min_u Q^*(z,u)$, is optimal. This method enables determining the optimal policy for the quantized belief MDP [18].

3.2.2. Near optimality of finite window policies for average cost. In this subsection, we focus on the finite window history to obtain a near-optimal policy for the average cost criteria. Throughout this subsection, we assume $\mathbb{Y}, \mathbb{U}$ to be finite. Assuming $\mathbb{Y}, \mathbb{U}$ as finite, we use finite window information to obtain finite state MDP and demonstrate that the near-optimal policy of this finite MDP is near-optimal for POMDP. We follow the approach and results from [21].

Recall z_n is the belief distribution defined as $z_n(\cdot) = P^\mu\{X_n \in \cdot \mid y_0, \dots, y_n, u_0, \dots, u_{n-1}\} \in \mathcal{P}(\mathbb{X})$, where the initial state X_0 has a prior distribution $\mu \in \mathcal{P}(\mathbb{X})$.

DEFINITION 3.7. *z_n^- is the posterior distribution at time n before observing Y_n and is defined as $z_n^-(\cdot) := P^\mu \{X_n \in \cdot \mid y_0, \dots, y_{n-1}, u_0, \dots, u_{n-1}\} \in \mathcal{P}(\mathbb{X})$, where the initial state X_0 has a prior distribution $\mu \in \mathcal{P}(\mathbb{X})$.*

For any $N, n \geq 0$, we can determine z_{n+N} as follows:

$$
z_{n+N} = P\left\{X_{n+N} \in \cdot \mid z_n^-, y_n, \dots, y_{n+N}, u_n, \dots, u_{n+N-1}\right\}.
$$

We then consider an alternative finite window belief MDP reduction. Let us define the state variable at time $n \geq N$ as

$$
(3.6) \qquad \hat{z}_n = \left(z_{n-N}^-, I_n^N\right),
$$

where, for $N \geq 1$, the components are

[1]This can be relaxed and refined to concern those states in the support of an invariant measure under an exploration policy [23].

$$z_{n-N}^{-} = \Pr\left(X_{n-N} \in \cdot \mid y_{n-N-1}, \ldots, y_0, u_{n-N-1}, \ldots, u_0\right), \tag{3.7}$$

$$I_n^N = \{y_n, \ldots, y_{n-N}, u_{n-1}, \ldots, u_{n-N}\}, \tag{3.8}$$

and for $N = 0$, $I_n^N = y_n$ with the prior measure μ on X_0. The state space is thus $\hat{\mathcal{Z}} = \mathcal{Z} \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$.

The natural mapping between state spaces is defined by $\psi : \hat{\mathcal{Z}} \to \mathcal{Z}$, such that

$$\psi\left(\hat{z}_n\right) = \psi\left(z_{n-N}^{-}, I_n^N\right) = P^{z_{n-N}^{-}}\left(X_n \in \cdot \mid y_{n-N}, \ldots, y_n, u_{n-N-1}, \ldots, u_{n-N-1}\right) = z_n. \tag{3.9}$$

The new transition kernel and cost function are defined as

$$\hat{\eta}(\cdot \mid \hat{z}, u) = \int_{\mathbb{Y}} 1_{\{(z_{n-N+1}^{-}, I_{n+1}^N) \in \cdot\}} \hat{P}(dy \mid \hat{z}, u), \tag{3.10}$$

where

$$\hat{P}(\cdot \mid \hat{z}, u) = \Pr\left\{Y_{n+1} \in \cdot \mid Z_{n-N} = z_{n-N}^{-}, I_n^N, U_n = u\right\}$$

from $\mathcal{Z} \times \mathbb{U}$ to $\mathbb{Y}$. The cost function is

$$\begin{aligned}\hat{c}\left(\hat{z}_n, u_n\right) &= \tilde{c}\left(\psi\left(z_{n-N}^{-}, I_n^N\right), u_n\right) \\ &= \int_{\mathbb{X}} c\left(x_n, u_n\right) P^{z_{n-N}^{-}}\left(dx_n \mid y_{n-N}, \ldots, y_n, u_{n-1}, \ldots, u_{n-N}\right).\end{aligned}$$

If γ is an optimal policy of the belief MDP, then $\psi^{-1}(\gamma)$ is an optimal policy of the finite window belief MDP [21].

Next, we explain how to derive a near-optimal policy for the finite window belief MDP in the context of discounted cost. We introduce a new approximate MDP for this purpose.

For $n \geq N$, fixing z_{n-N}^{-} to a constant $z \in P(\mathbb{X})$, we obtain a new MDP with state $\hat{z}_n = (z, I_n^N)$ with state space $\hat{\mathcal{Z}}^N := \{z\} \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$, cost function $\hat{c}^N(\hat{z}_n^N, u_n) := \hat{c}((z, I_n^N), u_n)$, and transition kernel $\hat{\eta}^N\left(\hat{z}_{n+1}^N \mid \hat{z}_n^N, u_n\right) := \hat{\eta}\left(\mathcal{Z} \times I_{n+1}^N \mid (z, I_n^N), u_n\right)$ [21].

This process is finite state and fully observable and allows for finding an optimal policy φ^N through Q-iteration. This policy can be extended to $\hat{\mathcal{Z}}$ as $\tilde{\varphi}^N(\hat{z}_n^N) = \tilde{\varphi}^N(z_{n-N}^{-}, I_n^N) := \varphi^N(z, I_n^N)$.

The following theorem indicates that this policy is also nearly optimal for $\hat{z}_n$.

Theorem 3.8 ([21, Theorem 3]). *For $\hat{z}_0 = P^{z_0^{-}}\left(X_n \in \cdot \mid I_0^N\right)$, with a policy $\hat{\gamma}$ acting on the first N steps which produces $I_0^N = Y[0, N], U_{[0,N-1]}$, the following holds:*

$$E_{z_0^{-}}^{\hat{\gamma}}\left[\left|\tilde{J}_\beta^N\left(\hat{z}_0, \tilde{\varphi}^N\right) - J_\beta^*\left(\hat{z}_0\right)\right| \mid I_0^N\right] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t^N,$$

where

$$\begin{aligned}L_t^N := \sup_{\hat{\gamma} \in \hat{\Gamma}} E_{z_0^{-}}^{\hat{\gamma}} [\, \| \, & P^{z_t^{-}}\left(X_{t+N} \in \cdot \mid Y_{[t,t+N]}, U_{[t,t+N-1]}\right) \\ & - P^{z}\left(X_{t+N} \in \cdot \mid Y_{[t,t+N]}, U_{[t,t+N-1]}\right) \|_{TV} \,].\end{aligned}$$

Here, $\hat{\gamma} \in \hat{\Gamma}$, where $\hat{\Gamma}$ can be taken to be Markov control policies.

With filter stability [21], as N approaches infinity, the upper bound tends to zero. Given filter stability and a sufficiently large N, the policy $\tilde{\varphi}^N$ becomes near-optimal for the finite window belief MDP for discounted cost criteria. If β is sufficiently large, this policy also becomes near-optimal for the average cost criterion.

Q-learning. We now outline how to find all optimal policies using Q-learning for the approximate finite belief MDP in the context of discounted cost. As the posterior distribution z is fixed, we track observations and actions. We assume tracking of the last $N+1$ observations and the last N control actions after at least $N+1$ time steps. Thus, at time n, we monitor the information variables I_n^N.

The Q-value iteration is constructed using these information variables. For these new approximate states, we follow the standard Q-learning algorithm. For any $I \in \mathbb{Y}^{N+1} \times \mathbb{U}^N$ and $u \in \mathbb{U}$, the Q-value update is

$$Q_{t+1}(I,u) = (1-\alpha_t(I,u))\,Q_t(I,u) + \alpha_t(I,u)\left(\hat{c}^N(I,u) + \beta \min_v Q_t\left(I_1^t, v\right)\right), \tag{3.11}$$

where $I_1^t = \{Y_{t+1}, y_t, \ldots, y_{t-N+1}, u_t, \ldots, u_{t-N+1}\}$. Exploration policies are employed, randomly choosing control actions independently, such that at time t, the action u_t is selected with probability σ_i for each $u_i \in \mathbb{U}$, where $\sigma_i > 0$ for all i. The following assumption is a revision of Assumption 11, specifically adapted for the finite window context:

Assumption 12 ([21 Assumption 4.1]).
1. $\alpha_t(I,u) = 0$ unless $(I_t, u_t) = (I,u)$. In other cases,
$$\alpha_t(I,u) = \frac{1}{1 + \sum_{k=0}^t 1_{\{I_k = I, u_k = u\}}}.$$
2. Under the 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. Furthermore, we have that $P(Y_t = y|x) > 0$ for every $x \in \mathbb{X}$, and thus during the exploration phase, every (I,u) pair is visited infinitely often.

THEOREM 3.9 ([21, Theorem 4.1, Corollary 5.1]). *Suppose the following conditions hold:*
1. *Assumption* 12 *holds.*
2. *The POMDP is such that the filter is stable uniformly over priors in expectation under total variation, meaning $L_t \to 0$ as $N \to \infty$.*

Then, the followings are true:
- *The algorithm given in* (3.11) *converges almost surely to Q^* which satisfies the fixed point equation.*
- *A stationary policy γ^N that satisfies $\gamma^N(I) \in \min_u Q^*(I,u)$ is an optimal policy.*

Note 3.2 (see [24]). Theorem 4.1 of [24] provides sufficient conditions for uniform filter stability. If we have $\bar{\alpha} = (1-\tilde{\delta}(\mathcal{T}))(2-\delta(Q)) < 1$, then the filter is exponentially stable with coefficient $\bar{\alpha}$ for any control policy. Here, $\tilde{\delta}(\mathcal{T}) = \inf_{u \in \mathcal{U}} \delta(T(\cdot|\cdot,u))$. See also an analysis via the Hilbert projective metric [13, 24].

This method allows for obtaining near-optimal policies of POMDP for the discounted cost and, consequently, under Assumption 1 for average cost. It is important to note that for the average cost optimal policy, the actions taken in the first N steps are not significant. The cost incurred during these initial steps does not impact the overall outcome, and since the optimal cost for each state is constant, applying the optimal policy after N steps will still yield an optimal policy.

4. Concluding remarks.

The average cost optimality of the partially observable Markov decision process is a challenging problem. In this paper, we presented explicit and easily testable conditions for the existence of solutions to the average cost optimality equation where the state space is compact. A comparison with the related literature and several examples are presented. Notably, our paper appears to be the first to present and utilize a contraction result for optimality analysis. Finally, we presented several implications of our analysis and existence result for approximations, learning, and robustness.

Appendix A. Proof of Theorem 2.2.

Proof. We equip $\mathcal{Z}$ with the metric W_1 to define the Lipschitz seminorm $\|f\|_L$ of any Borel measurable function $f:\mathcal{Z}\to\mathbb{R}$.

$$W_1\left(\eta(\cdot\mid z_0,u),\eta\left(\cdot\mid z_0',u\right)\right) \tag{A.1}$$
$$= \sup_{f\in \mathrm{Lip}(\mathcal{Z},1),\, \|f\|_\infty\le D/2}\left|\int_{\mathbb{Y}} f\left(z_1\left(z_0',u,y_1\right)\right)P\left(dy_1\mid z_0',u\right)\right.$$
$$\left. - \int_{\mathbb{Y}} f\left(z_1\left(z_0,u,y_1\right)\right)P\left(dy_1\mid z_0,u\right)\right|. \tag{A.2}$$

For any $f:\mathcal{Z}\to\mathbb{R}$ such that $\|f\|_L\le 1$ and $\|f\|_\infty\le D/2$, we have

$$\left|\int_{\mathbb{Y}} f\left(z_1\left(z_0',u,y_1\right)\right)P\left(dy_1\mid z_0',u\right) - \int_{\mathbb{Y}} f\left(z_1\left(z_0,u,y_1\right)\right)P\left(dy_1\mid z_0,u\right)\right| \tag{A.3}$$
$$\le \left|\int_{\mathbb{Y}} f\left(z_1\left(z_0',u,y_1\right)\right)P\left(dy_1\mid z_0',u\right) - \int_{\mathbb{Y}} f\left(z_1\left(z_0',u,y_1\right)\right)P\left(dy_1\mid z_0,u\right)\right|$$
$$+ \int_{\mathbb{Y}} \left|f\left(z_1\left(z_0',u,y_1\right)\right) - f\left(z_1\left(z_0,u,y_1\right)\right)\right| P\left(dy_1\mid z_0,u\right)$$
$$\le \frac{D}{2}\left\|P\left(\cdot\mid z_0',u\right) - P\left(\cdot\mid z_0,u\right)\right\|_{TV}$$
$$+ \int_{\mathbb{Y}} \left|f\left(z_1\left(z_0',u,y_1\right)\right) - f\left(z_1\left(z_0,u,y_1\right)\right)\right| P\left(dy_1\mid z_0,u\right). \tag{A.4}$$

For the first term,

$$\left\|P\left(\cdot\mid z_0',u\right) - P\left(\cdot\mid z_0,u\right)\right\|_{TV} \tag{A.5}$$
$$= \sup_{\|g\|_\infty\le 1}\left|\int g\left(y_1\right)P\left(dy_1\mid z_0',u\right) - \int g\left(y_1\right)P\left(dy_1\mid z_0,u\right)\right|$$
$$\le (1-\delta(Q))\left\|\mathcal{T}\left(dx_1\mid z_0',u\right) - \mathcal{T}\left(dx_1\mid z_0,u\right)\right\|_{TV}, \tag{A.6}$$

and by the Dobrushin contraction Theorem [8],

$$\left\|\mathcal{T}\left(dx_1\mid z_0',u\right) - \mathcal{T}\left(dx_1\mid z_0,u\right)\right\|_{TV} \tag{A.7}$$
$$= \sup_{\|g\|_\infty\le 1}\left(\int g\left(x_1\right)T\left(dx_1\mid z_0',u\right) - \int g\left(x_1\right)T\left(dx_1\mid z_0,u\right)\right)$$
$$= \sup_{\|g\|_\infty\le 1}\left(\int \tilde{g}_g(x_0)z_0'(dx_0) - \tilde{g}_g(x_0)z_0(dx_0)\right), \tag{A.8}$$

where

$$\tilde{g}_g(x) = \int g(x_1)\, T(dx_1 \mid x, u).$$

For all $x_0', x_0 \in \mathbb{X}$, we have

$$\|\mathcal{T}(dx_1 \mid x_0', u) - \mathcal{T}(dx_1 \mid x_0, u)\|_{TV} \leq \alpha d(x_0, x_0').$$

As a result, we get $\tilde{g}/\alpha \in Lip(\mathbb{X}, 1)$.

Then, by inequalities (A.5) and (A.7) we can write

$$\|P(\cdot \mid z_0', u) - P(\cdot \mid z_0, u)\|_{TV} \leq \alpha(1 - \delta(Q)) W_1(z_0', z_0). \tag{A.9}$$

Finally, we can analyze the second term in (A.4),

$$\begin{aligned}
&\int_{\mathbb{Y}} |f(z_1(z_0', u, y_1)) - f(z_1(z_0, u, y_1))|\, P(dy_1 \mid z_0, u) \\
&\quad \leq \int_{\mathbb{Y}} W_1(z_1(z_0', u, y_1), z_1(z_0, u, y_1)) P(dy_1 \mid z_0, u) \\
&\quad = \int_{\mathbb{Y}} \sup_{g \in \mathrm{Lip}(\mathbb{X})} \left(\int_{\mathbb{X}} g(x_1) w_{y_1}(dx_1) \right) P(dy_1 \mid z_0, u), \qquad \text{(A.10)}
\end{aligned}$$

where $w_{y_1} = (z_1(z_0', u, y_1) - z_1(z_0, u, y_1))$, which is a signed measure on $\mathbb{X}$. By the measurable selection theorem,[2] choose measurable

$$g_y \in \arg \sup_{g \in \mathrm{Lip}(\mathbb{X},1)} \left(\int_{\mathbb{X}} g(x) w_y(dx) \right).$$

After that we can continue with (A.10),

$$\begin{aligned}
&\int_{\mathbb{Y}} \sup_{g \in \mathrm{Lip}(\mathbb{X},1)} \left(\int_{\mathbb{X}} g(x_1) w_{y_1}(dx_1) \right) P(dy_1 \mid z_0, u) && \text{(A.11)} \\
&\quad = \int_{\mathbb{Y}} \int_{\mathbb{X}} g_{y_1}(x_1) z_1(z_0', u, y_1)(dx_1) P(dy_1 \mid z_0, u) \\
&\qquad - \int_{\mathbb{Y}} \int_{\mathbb{X}} g_{y_1}(x_1) z_1(z_0', u, y_1)(dx_1) P(dy_1 \mid z_0', u) && \text{(A.12)} \\
&\qquad + \int_{\mathbb{Y}} \int_{\mathbb{X}} g_{y_1}(x_1) z_1(z_0', u, y_1)(dx_1) P(dy_1 \mid z_0', u) && \text{(A.13)} \\
&\qquad - \int_{\mathbb{Y}} \int_{\mathbb{X}} g_{y_1}(x_1) z_1(z_0, u, y_1)(dx_1) P(dy_1 \mid z_0, u). && \text{(A.14)}
\end{aligned}$$

For the first term, we can write by the same argument as earlier

$$\left\| \int_{\mathbb{X}} g_{y_1}(x_1) z_1(z_0', u, y_1)(dx_1) \right\|_\infty \leq \|g_{y_1}\|_\infty \leq D/2.$$

So,

[2] See [15, Theorem 2, the Kuratowski–Ryll-Nardzewski measurable selection theorem]. Let $\mathbb{X}, \mathbb{Y}$ be Polish spaces, let $\Gamma = (x, \psi(x))$ where $\psi(x) \subset \mathbb{Y}$ be such that, $\psi(x)$ is closed for each $x \in \mathbb{X}$, and let Γ be a Borel measurable set in $\mathbb{X} \times \mathbb{Y}$. Then, there exists at least one measurable function $f : \mathbb{X} \to \mathbb{Y}$ such that $\{(x, f(x)), x \in \mathbb{X}\} \subset \Gamma$.

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} g_{y_1}(x_1) z_1\left(z_0', u, y_1\right)(dx_1) P(dy_1 \mid z_0, u) \tag{A.15}$$

$$\begin{aligned} &- \int_{\mathbb{Y}}\int_{\mathbb{X}} g_{y_1}(x_1) z_1\left(z_0', u, y_1\right)(dx_1) P(dy_1 \mid z_0', u) \qquad \text{(A.16)} \\ &\leq \frac{D}{2}\left\|P\left(\cdot \mid z_0', u\right) - P\left(\cdot \mid z_0, u\right)\right\|_{TV} \\ &\leq \alpha\frac{D}{2}(1-\delta(Q)) W_1\left(z_0', z_0\right) \end{aligned}$$

by inequality (A.9).

For the second term, we can write by smoothing

$$\int_{\mathbb{Y}}\int_{\mathbb{X}} g_{y_1}(x_1) z_1\left(z_0', u, y_1\right)(dx_1) P(dy_1 \mid z_0', u) \tag{A.17}$$

$$\begin{aligned} &- \int_{\mathbb{Y}}\int_{\mathbb{X}} g_{y_1}(x_1) z_1\left(z_0, u, y_1\right)(dx_1) P(dy_1 \mid z_0, u) \qquad \text{(A.18)} \\ &= \int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid z_0', u\right) - \int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid z_0, u\right), \end{aligned}$$

where

$$\omega(x_1) = \int_{\mathbb{Y}} g_{y_1}(x_1) Q\left(dy_1 \mid x_1\right).$$

For any $x', x'' \in \mathbb{X}$,

$$\begin{aligned} \int_{\mathbb{X}} \omega(x) \mathcal{T}\left(dx \mid x'', u\right) - \int_{\mathbb{X}} \omega(x) \mathcal{T}\left(dx \mid x', u\right) &\leq \|\omega\|_\infty \left\|\mathcal{T}(\cdot \mid x'', u) - \mathcal{T}\left(\cdot \mid x', u\right)\right\|_{TV} \\ &\leq \|\omega\|_\infty \alpha d(x'', x') \leq \alpha\frac{D}{2} d(x'', x'). \end{aligned}$$

So, by definition of the W_1 norm (1.7), we have

$$\begin{aligned} &\int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid z_0', u\right) - \int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid z_0, u\right) \qquad \text{(A.19)} \\ &\quad = \int_{\mathbb{X}}\int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid x_0, u\right) z_0'(dx_0) - \int_{\mathbb{X}}\int_{\mathbb{X}} \omega(x_1) \mathcal{T}\left(dx_1 \mid x_0, u\right) z_0(dx_0) \\ &\quad \leq \alpha\frac{D}{2} W_1\left(z_0, z_0'\right). \end{aligned}$$

So, by the inequalities (A.12), (A.14), (A.15), (A.17), (A.19) we get

$$\int_{\mathbb{Y}} \left|f\left(z_1\left(z_0', u, y_1\right)\right) - f\left(z_1\left(z_0, u, y_1\right)\right)\right| P\left(dy_1 \mid z_0, u\right) \leq \alpha\frac{D}{2}(2-\delta(Q)) W_1\left(z_0', z_0\right). \tag{A.20}$$

If we take the supremum of the equation over all $f \in Lip(\mathcal{Z})$, then by using the inequalities (A.4), (A.9), and (A.20), we can write

$$W_1\left(\eta(\cdot \mid z_0, u), \eta\left(\cdot \mid z_0', u\right)\right) \leq \left(\frac{\alpha D(3-2\delta(Q))}{2}\right) W_1\left(z_0, z_0'\right). \tag{A.21}$$

□

Appendix B. Proof of Lemma 2.5. We present a specialization of [34, Theorem 7.3.3] to the compact case, as in the paper we have that $\mathcal{Z}$ is compact.

Let us first recall the Arzelá–Ascoli theorem.

THEOREM B.1 ([9, Theorem 2.4.7]). *Let F be an equicontinuous family of functions on a compact space $\mathbb{X}$, and let h_n be a sequence in F such that the range of f_n is compact. Then, there exists a subsequence h_{n_k} which converges uniformly to a continuous function.*

Proof of Lemma 2.5. First, since the cost function $\tilde{c}$ is bounded (assume it is bounded by $M<\infty$) the expression $(1-\beta)J_\beta(z)$ is uniformly bounded by M for every $\beta\in(0,1)$ and $z\in\mathcal{Z}$. By the Bolzano–Weierstrass theorem, for a fixed z and any sequence $\beta\uparrow 1$, there exists a subsequence $\beta(k)\uparrow 1$ such that $(1-\beta(k))J_{\beta(k)}(z)\to\rho^*$ for some ρ^*. Observe that for any $z\in\mathcal{Z}$,

$$(1-\beta(k))\,J_{\beta(k)}(z)=(1-\beta(k))\left(J_{\beta(k)}(z)-J_{\beta(k)}\left(z_0\right)\right)+(1-\beta(k))\,J_{\beta(k)}\left(z_0\right),$$

which, by the uniform boundedness of $h_{\beta(k)}(z)=J_{\beta(k)}(z)-J_{\beta(k)}\left(z_0\right)$, implies that the limit ρ^* does not depend on z.

By Assumption 9(6), $h_{\beta(k)}$ is equicontinuous. By Theorem B.1, there exists a further subsequence of $h_{\beta(k)}$, $\left\{h_{\beta(k_l)}\right\}$, which converges (uniformly on compact sets) to a continuous and bounded function h. Since $\mathbb{U}$ is compact and the cost function is continuous and the transition kernel is weakly continuous, we have by the Bellman equation that

(B.1)
$$J_\beta(z)-J_\beta\left(z_0\right)=\min_{u\in\mathbb{U}}\left(c(z,u)+\beta\int\eta\left(dz'\mid z,u\right)\left(J_\beta\left(z'\right)-J_\beta\left(z_0\right)\right)-(1-\beta)J_\beta\left(z_0\right)\right).$$

Taking the limit in (B.1) along the subsequence $\beta(k_l)$, we get

$$\begin{aligned}h(z)&=\lim_l\min_{\mathbb{U}}\left[c(z,u)+\beta\left(k_l\right)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta(dy\mid z,u)-(1-\beta(k_l))J_{\beta(k_l)}\left(z_0\right)\right]\\&=\lim_l\min_{\mathbb{U}}\left[c(z,u)+\beta\left(k_l\right)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta(dy\mid z,u)\right]-\rho^*.\end{aligned}$$

From this we obtain

$$\lim_l\min_{\mathbb{U}}\left[c(z,u)+\beta\left(k_l\right)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta(dy\mid z,u)\right]=h(z)+\rho^*. \tag{B.2}$$

We now show that in the above, the order of limit and minimization can be swapped: Using the compactness of $\mathbb{U}$, the continuity of

$$c(z,u)+\beta(k_l)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta(dy|z,u)$$

on $\mathbb{U}$, and the equicontinuity of $\left\{h_{\beta(k)}\right\}$ we can define a sequence u_l such that

$$\begin{aligned}A_l&:=c\left(z,u_l\right)+\beta\left(k_l\right)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta\left(dy\mid z,u_l\right)\\&=\min_{\mathbb{U}}\left[c(z,u)+\beta\left(k_l\right)\int_{\mathcal{Z}}h_{\beta(k_l)}(y)\eta(dy\mid z,u)\right].\end{aligned}$$

By the compactness of the action space $\mathbb{U}$, there exists a further subsequence such that $u_{l_n}\to u^*$ for some u^* along this further subsequence. By weak continuity of the kernel, we then have that $\eta(dy\mid z,u_{l_n})\to\eta(dy\mid z,u^*)$. Since $h_{\beta(k_l)}$ is uniformly bounded, we have $\lim A_l=\lim A_{l_n}=c(z,u)+\int_{\mathcal{Z}}h(y)\eta(dy\mid z,u^*)$.

On the other hand, for any fixed action $\bar{u}$, we know that

$$c(z,\bar{u}) + \beta(k_l)\int_{\mathcal{Z}} h_{\beta(k_l)}(y)\eta(dy \mid z,\bar{u}) \geq A_l,$$

and taking the limit as $l \to \infty$, by the same argument,

$$c(z,\bar{u}) + \int_{\mathcal{Z}} h(y)\eta(dy \mid z,\bar{u}) \geq \lim A_l = c(z,u) + \int_{\mathcal{Z}} h(y)\eta(dy \mid z,u).$$

Therefore,

$$\lim A_l = \min_{\mathbb{U}}\left[c(z,u) + \int_{\mathcal{Z}} h(y)\eta(dy \mid z,u)\right].$$

Combining this with (B.2), we obtain

$$h(z) + \rho^* = \min_{\mathbb{U}}\left[c(z,u) + \int_{\mathcal{Z}} h(y)\eta(dy \mid z,u)\right].$$

Thus, we have found a bounded solution to the ACOE equation, completing the proof. ∎

REFERENCES

[1] V. Bogachev, *Measure Theory*, Vol. II, Springer, 2007.

[2] V. S. Borkar, *Average cost dynamic programming equations for controlled Markov chains with partial observations*, SIAM J. Control Optim., 39 (2000), pp. 673–681, https://doi.org/10.1137/S0363012998345172.

[3] V. S. Borkar, *Dynamic programming for ergodic control with partial observations*, Stochastic Process. Appl., 103 (2003), pp. 293–310.

[4] V. S. Borkar, *Dynamic programming for ergodic control of Markov chains under partial observations: A correction*, SIAM J. Control Optim., 45 (2007), pp. 2299–2304, https://doi.org/10.1137/05064610X.

[5] V. S. Borkar and A. Budhiraja, *A further remark on dynamic programming for partially observed Markov processes*, Stochastic Process. Appl., 112 (2004), pp. 79–93.

[6] 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, 2024 (also arXiv:2311.12609), 2023.

[7] Y. E. Demirci and S. Yüksel, *Geometric Ergodicity, Unique Ergodicity and Wasserstein Continuity of Non-Linear Filters with Compact State Space*, preprint, arXiv:2307.15764, 2023.

[8] R. L. Dobrushin, *Central limit theorem for nonstationary Markov chains.* , Theory Probab. Appl., 1 (1956), pp. 65–80, https://doi.org/10.1137/1101006.

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

[10] E. Feinberg, P. Kasyanov, and M. Zgurovsky, *Partially observable total-cost Markov decision process with weakly continuous transition probabilities*, Math. Oper. Res., 41 (2016), pp. 656–681.

[11] E. A. Feinberg, *Controlled Markov processes with arbitrary numerical criteria*, Theory Probab. Appl., 27 (1982), pp. 486–503, https://doi.org/10.1137/1127058.

[12] 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 Lett., 15 (1990), pp. 425–432.

[13] F. L. Gland and N. Oudjane, *Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters*, Ann. Appl. Probab., 14 (2004), pp. 144–187.

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

[15] C. J. Himmelberg, T. Parthasarathy, and F. S. V. Vleck, *Optimal plans for dynamic programming problems*, Math. Oper. Res., 1 (1976), pp. 390–394.

[16] 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 Lett., 55 (2006), pp. 165–173.
[17] A.D. Kara, N. Saldi, and S. Yüksel, *Weak Feller property of non-linear filters*, Systems Control Lett., 134 (2019), pp. 104–512.
[18] A.D. Kara, N. Saldi, and S. Yüksel, *Q-learning for mdps with General Spaces: Convergence and Near Optimality via Quantization under Weak Continuity*, J. Mach. Learn. Res., 24 (2023), pp. 1–34.
[19] A. D. Kara and S. Yüksel, *Robustness to incorrect priors in partially observed stochastic control*, SIAM J. Control Optim., 57 (2019), pp. 1929–1964, https://doi.org/10.1137/17M1157660.
[20] A. D. Kara and S. Yüksel, *Robustness to incorrect system models in stochastic control*, SIAM J. Control Optim., 58 (2020), pp. 1144–1182, https://doi.org/10.1137/18M1208058.
[21] A.D. Kara and S. Yüksel, *Convergence of finite memory Q learning for POMDPs and near optimality of learned policies under filter stability*, Math. Oper. Res. 48 (2023), pp. 2066–2093.
[22] A.D. Kara and S. Yüksel, *Near optimality of finite memory feedback policies in partially observed Markov decision processes*, J. Mach. Learn. Res., 23 (2022), pp. 1–46.
[23] A.D. Kara and S. Yüksel, *Q-learning for stochastic control under general information structures and non-Markovian environments*, Trans. Mach. Learn. Res. (TMLR), (2024), pp. 2835–8856, https://openreview.net.
[24] C. McDonald and S. Yüksel, *Exponential filter stability via Dobrushin's coefficient*, Electron. Commun. Probab., 25 (2020), pp. 1–13.
[25] C. McDonald and S. Yüksel, *Robustness to incorrect priors and controlled filter stability in partially observed stochastic control*, SIAM J. Control Optim., 60 (2022), pp. 842–870, https://doi.org/10.1137/21M1417442.
[26] L. K. Platzman, *Optimal infinite-horizon undiscounted control of finite probabilistic systems*, SIAM J. Control Optim., 18 (1980), pp. 362–380, https://doi.org/10.1137/0318028.
[27] D. Rhenius, *Incomplete information in Markovian decision models*, Ann. Statist., 2 (1974), pp. 1327–1334.
[28] W. J. Runggaldier and L. Stettner, *Approximations of Discrete Time Partially Observed Control Problems*, Giardini Editori e Stampatori, 1994.
[29] N. Saldi, T. Linder, and S. Yüksel, *Finite Approximations in Discrete-Time Stochastic Control: Quantized Models and Asymptotic Optimality*, Springer, 2018.
[30] L. Stettner, *Long run control with degenerate observation*, SIAM J. Control Optim., 57 (2019), pp. 880–899, https://doi.org/10.1137/18M1196844.
[31] C. Villani, *Optimal Transport: Old and New*, Springer, 2008.
[32] H. Yu and D. P. Bertsekas, *On near optimality of the set of finite-state controllers for average cost POMDP*, Math. Oper. Res., 33 (2008), pp. 1–11.
[33] S. Yüksel, *Another Look at Partially Observed Optimal Stochastic Control: Existence, Ergodicity, and Approximations without Belief-Reduction,* preprint, https://arxiv.org/abs/2301.11244v2, 2023.
[34] S. Yüksel, *Optimization and Control of Stochastic Systems*, lecture notes, Queen's University, 2024.
[35] 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 Probab. Appl., 21 (1976), pp. 153–158, https://doi.org/10.1137/1121014.