Journal of Machine Learning Research 27 (2026) 1-50 Submitted 3/24; Revised 5/26; Published 6/26

Approximations and Learning for Continuous State and Action MDPs under Average Cost Criteria *

Ali Devran Kara and Serdar Yüksel †

Editor: Csaba Szepesvari

Abstract

In this paper, for Markov Decision Processes (MDPs) with standard Borel spaces, (i) we first provide a discretization based approximation method for MDPs with continuous spaces under average cost criteria, and provide error bounds for approximations when the dynamics are only weakly continuous (for asymptotic convergence of errors as the grid sizes vanish) or Wasserstein continuous (with a rate in approximation as the grid sizes vanish) under certain ergodicity assumptions. In particular, we relax the total variation condition given in prior work to weak continuity or Wasserstein continuity. (ii) We provide synchronous and asynchronous (quantized) Q-learning algorithms for continuous spaces via quantization (where the quantized state is taken to be the actual state in corresponding Q-learning algorithms presented in the paper), and establish their convergence. (iii) We finally show that the convergence is to the optimal Q values of a finite approximate model constructed via quantization, which implies near optimality of the arrived solution.

Keywords: Reinforcement learning, stochastic control, finite approximations, MDPs with general spaces

1. Introduction

In this paper, we study approximate solutions for Markov decision processes (MDPs) under average cost criterion. We consider problems with continuous state and action spaces and provide approximate planning and reinforcement learning results. In particular, we show near-optimality of solutions to approximate models and those obtained via reinforcement learning.

Before we present the related research in these problems and discuss our contributions in more detail, we introduce the problem formulation: A fully observed Markov control model is a tuple

$$(\mathbb{X}, \mathbb{U}, \mathcal{T}, c),$$

where $\mathbb{X}$ is the (standard Borel) state space that is a metric space with the associated metric $d_{\mathbb{X}}$. Similarly, $\mathbb{U}$ is the action space, assumed to be a metric space with the metric $d_{\mathbb{U}}$. Under $d_{\mathbb{X}}$ and $d_{\mathbb{U}}$, $\mathbb{X}$ and $\mathbb{U}$ are complete and separable. The transition kernel of the model is denoted by $\mathcal{T}$ on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{U}$. Finally, $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is the cost function.

Let $\mathbb{H}_0 := \mathbb{X}, \quad \mathbb{H}_t = \mathbb{H}_{t-1} \times \mathbb{X} \times \mathbb{U}$ for $t = 1, 2, \ldots$. We let, for $t \in \mathbb{Z}_+$, h_t denote an element of $\mathbb{H}_t$, with $h_t = \{x_{[0,t]}, u_{[0,t-1]}\}$. We use the notation $x_{[0,t]} := (x_0, x_1, \cdots, x_t)$.

An admissible control policy π is a sequence of measurable functions $(\pi_t : t = 0, 1, 2 \ldots)$ such that $\pi_t : \mathbb{H}_t \to \mathcal{P}(\mathbb{U})$ with $u_t = \pi_t(h_t)$ where $\mathcal{P}(\mathbb{U})$ denotes the set of all probability measures

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

†. Ali D. Kara is with the Department of Mathematics, Florida State University, Tallahassee, FL, USA, Email: akara@fsu.edu. S. Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada, Email: yuksel@queensu.ca

on $\mathbb{U}$. We denote the set of all admissible policies by Π_A. If an admissible policy π is such that $u_t \sim f(x_t)$ for some Borel measurable $f : \mathbb{X} \to \mathcal{P}(\mathbb{U})$, then the policy is said to be stationary; we denote the set of stationary policies by Π_S.

We assume that such an MDP model is given on the $(\mathbb{X} \times \mathbb{U})^{\mathbb{Z}_+}$-valued state-action process $(X_t, U_t)_{t\in\mathbb{Z}_+}$. For each $x \in \mathbb{X}$ and $\pi \in \Pi_A$, the kernel $\mathcal{T}$ and the policy π induce a probability measure (called a strategic measure) P_x^π on $((\mathbb{X} \times \mathbb{U})^{\mathbb{Z}_+}, \mathcal{B}((\mathbb{X} \times \mathbb{U})^{\mathbb{Z}_+}))$ (where $\mathcal{B}(\cdot)$ denotes the Borel σ-field on the space $\cdot$) with $X_0 \sim \delta_x$. We let E_x^π denote the corresponding expectation.

We consider the following average cost problem:

$$J^*(x) := \inf_\pi J(x, \pi) = \inf_{\pi\in\Pi_A} \limsup_{T\to\infty} \frac{1}{T} E_x^\pi [\sum_{t=0}^{T-1} c(x_t, u_t)]. \tag{1}$$

We also define the discounted cost for some discount factor $0 < \beta < 1$:

$$J_\beta^*(x) := \inf_\pi J_\beta(x, \pi) = \inf_{\pi\in\Pi_A} \sum_{t=0}^{\infty} \beta^t E_x^\pi [c(x_t, u_t)]. \tag{2}$$

1.1 Literature Review

Finding efficient solutions to MDPs with continuous spaces is a challenging and important problem. In this context, various approximation techniques have been presented (see e.g. Dufour and Prieto-Rumeau, 2012; Bertsekas, 1975; Chow and Tsitsiklis, 1991; Bertsekas and Tsitsiklis, 1996a; Szepesvári, 2010; Tsitsiklis and Roy, 1997; Singh et al., 1995; Melo et al., 2008; Gaskett and D. Wettergreen, 1999; Szepesvàri and Smart, 2004; Meyn, 2022; Saldi et al., 2017, 2018). Many studies in the literature have focused on either finite-horizon problems or infinite-horizon discounted cost problems, using the dynamic programming principle and the contraction properties of Bellman operators induced by the discount factor. For average cost problems, however, the same techniques are not immediately applicable.

For MDPs with continuous state spaces, existence of optimal solutions has been well studied under the infinite horizon average cost criterion, under either weak continuity of the kernel (in both state and action) or strong continuity (of the kernel in actions for every state), together with various stability/ergodicity assumptions. We refer the reader to the works by Arapostathis et al. (1993); Hernández-Lerma (2012); Hernández-Lerma and Lasserre (1996, 1999); Costa and Dufour (2012); Gordienko and Hernández-Lerma (1995); Vega-Amaya (2003a,b); Hernández-Lerma and Lasserre (1996); Borkar (2001); Feinberg et al. (2012) for comprehensive surveys on optimality results and general solution techniques for optimal control under the average cost criteria. However, research on computational and reinforcement learning methods for average cost optimality still entails open problems, especially for MDPs with continuous spaces.

The primary contributions of the paper concern finite approximations for MDPs with continuous spaces, their near optimality, and reinforcement learning for average cost problems. We will summarize the related research in these areas separately:

Approximations for MDPs with continuous spaces. For the study of continuous space MDPs, establishing regularity and continuity properties of the value functions is crucial. To this end, one usually needs continuity assumptions on the stage-wise cost functions, and continuity assumptions on the transition models under a suitable metric. Weak convergence metrics, and Wasserstein distances are used frequently for the regularity of the transition models since they are, in general,

weaker and less demanding than other metric choices, e.g., total variation and relative entropy (Kullback–Leibler divergence) type distance notions. These are used to establish Lipschitz continuity of the value functions, as well as to establish the consistency of model approximations. The papers by Pirotta et al. (2015); Asadi et al. (2018); Rachelson and Lagoudakis (2010); Maran et al. (2023) study MDPs with Wasserstein continuous transition models under the discounted cost setting. However, one drawback of these papers is that they work on a class of Lipschitz continuous policies which is a suboptimal class in general. For the discounted cost criterion, Saldi et al. (2016, 2015a,b, 2017); Kara et al. (2023) have shown that under only weak continuity conditions for an MDP with standard Borel state and action spaces, finite models obtained by the quantization of the state and action spaces lead to control policies that are asymptotically optimal as the quantization rate increases. They also established convergence rates under Lipschitz continuity assumptions on the model, without requiring the class of policies to be a priori continuous. The authors of the current paper generalized these results to general approximation schemes beyond discretization based approaches under several different continuity assumptions on the transition models such as *continuous* weak continuity, set-wise continuity as well as total variation continuity (Kara and Yüksel, 2020, 2021).

On the weak metrics for approximations and robustness to model mismatch. Related to the discussion above, a salient consideration involving approximations for MDPs with general spaces is that unless convergence in models occurs in a strong sense, such as in total variation uniform over sets and actions, as models approach one another, the induced expected costs may not asymptotically agree under every admissible policy. If convergence is to hold under the Wasserstein metric, such a uniform convergence only would hold if policies were restricted to be Lipschitz a priori (see e.g. Pirotta et al., 2015; Asadi et al., 2018; Rachelson and Lagoudakis, 2010; Maran et al., 2023), which is often too restrictive for the approximate models. A detailed study on such model convergence and robustness has been carried out by Kara and Yüksel (2020, 2021). Under *continuous convergence of kernels*, Kara and Yüksel (2020) in discrete-time and Pradhan and Yüksel (2024) in continuous-time, established robustness when a control policy designed for an approximate/incorrect model is applied to a true model. Saldi et al. (2017); Kara and Yüksel (2021); Kara et al. (2023) present a construction for the approximate models through quantizing the actual model with continuous spaces, which allows for continuity and robustness results with only a weak continuity assumption on the true transition kernel which, in turn, leads to *continuous weak convergence* of approximate models as discussed by Kara and Yüksel (2021) and thus becomes an important special case of robustness under model convergence. Closely related to such results and continuous convergence, a topology on spaces of probability measures corresponding to laws of stochastic processes, which has been used in a wide variety of contexts in stochastic analysis, is defined as follows: a sequence of stochastic processes is said to converge to another process if their finite-dimensional marginals converge weakly, and their conditional distributions of future variables given the past (viewed as measure-valued stochastic processes) also converge weakly. Aldous (1981) has termed this *extended weak convergence* and Hellwig (1996) has named it *the information topology*; these have recently been shown to be equivalent in discrete-time by Backhoff-Veraguas et al. (2020a); Pammer (2024). The *adapted Wasserstein metric* (see Bartl et al., 2024; Backhoff-Veraguas et al., 2020a; Beiglböck et al., 2022) has been shown to possess similar robustness properties in a variety of applications (Bayraktar et al., 2020; Backhoff-Veraguas et al., 2020b; Bartl and Wiesel, 2023), not unlike the continuous weak convergence notion noted above (see Saldi and Yüksel (2025) for a comprehensive review). Please see Section 3.2.1 for further discussion.

Approximations for the average cost criterion. We note that the analysis of average cost problems is typically more challenging, especially for problems with continuous spaces, as the stability (or the ergodicity) of the problem plays a crucial role. For the average cost criterion, Saldi (2019), Saldi et al. (2018, Theorem 4.14) provide error bounds for finite approximations; however, these results require total variation continuity of the transition models as well as certain mixing conditions.

In this paper, we relax the total variation continuity condition to weak or Wasserstein continuity.

Reinforcement learning for the average cost criterion. There are a number of publications that study reinforcement learning methods for MDPs under average cost criterion. To our knowledge, the majority of these studies focus on finite spaces.

The papers by Abounadi et al. (2001); Gosavi (2004) are among the earliest studies that provide convergent learning algorithms based on relative value iteration (as well as stochastic shortest path under a recurrence condition for a given state), and the convergence of these algorithms has been established via the ODE method by Borkar and Meyn (2000) for finite models.

Konda and Tsitsiklis (2003) studied policy improvement and actor-critic methods for continuous space (in particular, Polish space) MDPs under average cost criteria. The approach relies on the linear approximation of the value functions and the parametrization of the policies. The convergence of this method is shown under a uniform minorization assumption over the parametrized policy space (Konda and Tsitsiklis, 2003, Assumption 4.2) which is similar to the assumption used in our paper (see Assumption 1). However, the learned solution is only locally optimal due to the nature of policy parametrization methods.

Ormoneit and Sen (2002) also focus on reinforcement learning methods for MDPs with continuous state spaces under the average cost criteria. Their method is based on a kernel-based approximate dynamic programming, and convergence of the algorithms is shown under a minorization condition (similar to Assumption 1) for several different averaging kernel functions. However, they impose strong regularity conditions on the transition model for the kernel based methods to work. In particular, it is assumed that the transition kernel admits a density function with respect to the Lebesgue measure such that $\mathcal{T}(dx_1|x,u) = f_u(x_1,x)\lambda(dx_1)$ where $f_u(x_1,x)$ is strictly positive and uniformly continuous in both variables. This assumption is, in particular, even stronger than total variation continuity of the transition kernel.

Among the relatively more recent studies, the comprehensive paper Wan et al. (2021) provides convergent off-policy learning algorithms to stabilize the value function estimation for finite models; the convergence proof by Wan et al. (2021) builds on the ODE method (Abounadi et al., 2001; Borkar and Meyn, 2000) but relaxes some of the conditions in Abounadi et al. (2001) with regard to the reference term subtracted in each iterate. We should note that the proof method of convergence by Abounadi et al. (2001) (and thus Wan et al. (2021)) for both the synchronous and asynchronous Q-learning build on the synchronous update dynamics analysis as these are equivalent under the ODE method.

Abounadi et al. (2001, Section 5) note the need for generalizing the analysis to continuous spaces for relative Q learning methods. We also note that the appendix of Wan et al. (2021) notes the continuous state/action setup as an open problem, which our current paper addresses.

Zhang and Ross (2021) study a policy improvement method for average-cost (reward) MDPs for finite state-action spaces. Wang et al. (2023) focus on robust model-free methods for finite systems where the transition model belongs to an uncertainty set which is constructed under various metrics. Suttle et al. (2023) focus on relaxing the exponential mixing assumption for average cost criteria for finite models and provides an actor-critic method. Zhang et al. (2021a) study finite

sample guarantees for a synchronous Q-learning algorithm under the average cost criterion. We also note that Yang et al. (2019) study a convergent actor-critic method under average-cost criteria for continuous models with linear systems and additive Gaussian noise. We refer the reader to the paper by Szepesvári (2010) for a general review on the subject.

We also note more recent work on average cost MDPs that focuses on sample complexity (see Chen, 2025; Bravo and Cominetti, 2024; Jin et al., 2024; Zhang et al., 2021b).

A complementary line of work studies infinite-horizon average cost problems for continuous spaces under structural assumptions. One such direction involves linear MDPs where the transition models and the cost function are assumed to be within the linear span of known basis (feature) functions. Wei et al. (2021); Wu et al. (2022) and more recently Hong et al. (2025); Chae et al. (2025); Hong and Tewari (2025) study linear and linear mixture MDPs via the vanishing discount approach. For kernel-based function approximation, Vakili and Olkhovskaya (2024) propose an algorithm under reproducing kernel Hilbert space assumptions on the value function. In this paper, we do not make parametric assumptions on the model, and only assume weak or Wasserstein continuity of the transition kernel on a general standard Borel space, and establish almost sure convergence of (quantized) Q-learning to the optimal value of an explicit finite approximate model with near optimality guarantees.

Almost all of these papers focus on MDP problems under infinite horizon average cost criteria for finite models, i.e. where the state and action spaces are finite, or impose parametric (linear/kernel) structure on the model and thus are able to use the Markovian nature for learning.

To this end, in our paper we consider general continuous spaces for which we first construct a discretized model where we present weaker continuity conditions than currently available in the literature. After discretization, the convergence analysis for learning methods requires further adaptations as the discretized states are no longer Markovian. In addition, ergodicity conditions are required to ensure the stability of the associated stochastic iteration algorithms. For continuous models, the ergodicity conditions and the stability analysis differ significantly from those involving finite models. In particular, in addition to obtaining average cost counterparts of the analysis of Kara et al. (2023) (which in turn builds on Kara and Yüksel, 2023), the paper introduces additional technical methods for the convergence analysis that are new even for finite MDPs, whose treatment in the literature has been restricted to the ODE method as noted earlier.

Contributions. In view of the above, we address several open questions in the literature along the following directions:

(i) [Approximation Results for Average Cost Infinite Horizon Control] In Section 3, we provide a discretization-based approximation method for fully observed MDPs with continuous spaces under the average cost criteria, and we provide error bounds for the approximations when the dynamics are only weakly continuous under certain ergodicity assumptions. Theorem 5 provides error bounds for action space discretization. Theorem 11 presents near optimality of control policies obtained for models with discretized state spaces. Notably, we relax the total variation condition given by Saldi (2019); Saldi et al. (2018) to weak (Feller) continuity (in Theorem 15) or Wasserstein continuity conditions (in Theorem 11); the former leads only to asymptotic convergence, whereas the latter provides a rate of convergence.

(ii) [Reinforcement Learning Analysis for Continuous State/Action Models] In Section 4, we consider what happens if one uses Q-learning with a quantization map, and whether the learned values correspond to the approximate model constructed in Section 3. We present

quantized Q-learning algorithms and show that they indeed converge to the optimal Q-values of the approximate models constructed in Section 3 for a particular weighting measure that is given by the invariant measure of the state process. When one runs the Q-learning algorithm, it is important to note that the quantized process is not an MDP, and in fact should be viewed as a POMDP, a viewpoint used by Kara and Yüksel (2023) and by Kara et al. (2023). For the synchronous algorithm, we use the properties of the span semi-norm. For the asynchronous setup, we generalize the proof method given by Kara and Yüksel (2023) for the average cost criterion under certain ergodicity properties induced by an exploration policy, though with additional technical analysis as the Q-iterates do not satisfy the boundedness properties a priori unlike the discounted cost criterion setup. In particular, in Section 4.1 we present and study a synchronous Q-learning algorithm, and in Section 4.2, we present and study an asynchronous algorithm. Theorem 17 establishes the convergence of a synchronous Q-learning algorithm. Theorem 19 shows the convergence of an asynchronous Q-learning algorithm.

(iii) [Convergence to Near-Optimality for Continuous State/Action Models] For both the synchronous and the asynchronous quantized Q-learning algorithms, the limit is shown to be the fixed-point solution of the optimality equation of an approximate model as in (i) above, and thus the convergence is to near-optimal policies.

To give a taste of the main results, we summarize our approximation results for finite action spaces. Approximation of continuous action spaces by finite sets is discussed in Section 3.1; accordingly, we assume throughout this summary that the action space is finite.

We consider a finite partition of the state space $\mathbb{X}$, given by $\{B_i\}_{i=1}^M$. We denote by $L_{\mathbb{X}}$ the worst case distortion over the partition bins under a given weight measure (see (23) for the definition). A precise statement of the following theorem is given in Theorem 16.

Main Result. *Suppose $\hat{\pi}$ is optimal for the approximate model. For a given initial state $x_0 \in \mathbb{X}$, consider the average-cost of $\hat{\pi}$ in the original MDP denoted by $J(x_0, \hat{\pi})$, and the optimal average-cost value of the original MDP denoted by $J^*(x_0)$. Furthermore, the optimal average cost value of the approximate MDP defined in* (14)-(15) *is denoted by $\hat{J}(x_0)$ for initial state $x_0 \in \mathbb{X}$.*

Under Assumption 1, these values are independent of the initial state and we write

$$\rho(\hat{\pi}) = J(x_0, \hat{\pi}), \quad \hat{\rho} = \hat{J}(x_0), \quad \rho = J^*(x_0)$$

for all $x_0 \in \mathbb{X}$. Here, K_c denotes the Lipschitz constant of the one-stage cost function and $K_{\mathcal{T}}$ denotes the Lipschitz (contraction) constant of the transition kernel, as defined in Assumption 4.

Fix δ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$ where $\kappa := 1 - \nu(\mathbb{X}) > 0$, and let D denote the diameter of the state space.

i. Under Assumption 4 with $K_{\mathcal{T}} < 1$, we have that

$$|\rho - \hat{J}(x_0)| \leq \frac{K_c}{1 - K_{\mathcal{T}}} L_{\mathbb{X}}.$$

ii. Under Assumption 1 and Assumption 4,

$$|\rho - \hat{\rho}| \leq \frac{K_c D^{1-\delta}}{1 - K_{\mathcal{T}}^{\delta} \kappa^{1-\delta}} \left(L_{\mathbb{X}}\right)^{\delta}$$

Furthermore,

$$\rho(\hat{\pi}) - \rho \leq \frac{2K_c D^{1-\delta}}{(1-\kappa)\left(1 - K_{\mathcal{T}}^{\delta}\kappa^{1-\delta}\right)} \left(L_{\mathbb{X}}\right)^{\delta}$$

iii. If the quantization is such that $L_{\mathbb{X}} = \frac{1}{n}$ (which is possible as $\mathbb{X}$ is assumed to be compact), then by denoting the learned policy by π_n, under Assumptions 1 and 5, we have that

$$\lim_{n\to\infty} \rho(\pi_n) = \rho.$$

iv. Running Q-learning on quantized states (called Quantized Q-Learning with a synchronous Algorithm 1 and asynchronous Algorithm 2) converges and results in a policy $\hat{\pi}$. This policy corresponds to an optimal policy for a discretized approximate model, with discretization bins weighted by the stationary distribution of the state process induced by an exploration policy. As a corollary of the above, the convergence is to near optimality.

2. Average Cost Optimality Equation and Contraction Properties of Relative Value Iteration

We start our analysis by first reviewing technical tools needed throughout the paper and related results on average cost optimality.

2.1 Convergence Notions for Probability Measures and Regularity Properties of Transition Kernels

For the analysis of the technical results, we use different notions of convergence for sequences of probability measures.

Two important notions of convergence for sequences of probability measures are weak convergence and convergence under total variation. A sequence $\{\nu_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathbb{X})$ is said to converge to $\nu \in \mathcal{P}(\mathbb{X})$ *weakly* if $\int_{\mathbb{X}} c(x)\nu_n(dx) \to \int_{\mathbb{X}} c(x)\nu(dx)$ for every continuous and bounded $c : \mathbb{X} \to \mathbb{R}$.

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

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

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

Finally, for probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ with finite first-order moments (that is, $\int \|x\| \, d\mu$ and $\int \|x\| \, d\nu$ are finite), the *first order Wasserstein* distance is defined as

$$W_1(\mu, \nu) = \inf_{\pi(\mu,\nu)} E[d_{\mathbb{X}}(X, Y)] = \sup_{f:Lip(f)\leq 1} |\int f(x)\mu(dx) - \int f(x)\nu(dx)|$$

where $\pi(\mu,\nu)$ denotes all possible couplings of X and Y with marginals $X \sim \mu$ and $Y \sim \nu$, and

$$Lip(f) := \sup_{e \neq e'} \frac{|f(e) - f(e')|}{d_{\mathbb{X}}(e, e')},$$

and the second equality follows from the dual formulation of the Wasserstein distance (Villani, 2009, Remark 6.5). Note that weak convergence and Wasserstein convergence are equivalent if the underlying space is compact.

We now introduce the class of Hölder continuous functions and consider their use as test functions for studying probability measures. Let $f : \mathbb{X} \to \mathbb{R}$. The function f is said to be Hölder continuous of order $\delta \in (0, 1]$ if there exists a constant $C > 0$ such that

$$|f(x) - f(y)| \leq C\, d_{\mathbb{X}}(x, y)^{\delta}, \qquad \forall x, y \in \mathbb{X}.$$

Here, δ is called the *Hölder exponent* and C the *Hölder constant*. We denote the optimal Hölder constant of f by $[f]_{H^\delta}$ such that

$$[f]_{H^\delta} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d_{\mathbb{X}}(x, y)^{\delta}}. \tag{3}$$

Lemma 1 *(see Maran et al., 2023, Proposition 10) For any δ-Hölder f, we have*

$$\left| \int f(x) P(dx) - \int f(x) P'(dx) \right| \leq [f]_{H^\delta} W_1(P, P')^{\delta}.$$

We define the following regularity properties for the transition kernels:

- $\mathcal{T}(\cdot|x, u)$ is said to be weakly continuous in (x, u) (or weak Feller), if $\mathcal{T}(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ weakly for any $(x_n, u_n) \to (x, u)$.
- $\mathcal{T}(\cdot|x, u)$ is said to be continuous under total variation in (x, u), if $\|\mathcal{T}(\cdot|x_n, u_n) - \mathcal{T}(\cdot|x, u)\|_{TV} \to 0$ for any $(x_n, u_n) \to (x, u)$.
- $\mathcal{T}(\cdot|x, u)$ is said to be continuous under the first-order Wasserstein distance in (x, u), if

  $$W_1(\mathcal{T}(\cdot|x_n, u_n), \mathcal{T}(\cdot|x, u)) \to 0$$

  for any $(x_n, u_n) \to (x, u)$. To ensure continuity of $\mathcal{T}$ with respect to the first-order Wasserstein distance, in addition to weak continuity, we may assume that there exists a function $g : [0, \infty) \to [0, \infty)$ such that $\frac{g(t)}{t} \uparrow \infty$ as $t \to \infty$, and

  $$\sup_{(x,u) \in K \times \mathbb{U}} \int g(\|y\|)\, \mathcal{T}(dy|x, u) < \infty$$

  for any compact $K \subset \mathbb{X}$. Note that the latter condition implies uniform integrability of the collection of random variables with probability measures $\mathcal{T}(dx_1|X_0 = x_n, U_0 = u_n)$ as $(x_n, u_n) \to (x, u)$, which, coupled with weak convergence, can be shown to imply convergence under the Wasserstein distance.

Example 1 *Some example models satisfying these regularity properties are as follows:*

(i) For a model with the dynamics $x_{t+1} = f(x_t, u_t, w_t)$, the induced transition kernel $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u) if $f(x,u,w)$ is a continuous function of (x,u), since for any continuous and bounded function g

$$\int g(x_1)\mathcal{T}(dx_1|x_n,u_n) = \int g(f(x_n,u_n,w))\nu(dw)$$
$$\to \int g(f(x,u,w))\nu(dw) = \int g(x_1)\mathcal{T}(dx_1|x,u)$$

where ν denotes the probability measure of the noise process. If we also have that $\mathbb{X}$ is compact, the transition kernel $\mathcal{T}(\cdot|x,u)$ is also continuous under the first order Wasserstein distance.

(ii) For a model with the dynamics $x_{t+1} = f(x_t, u_t) + w_t$, the induced transition kernel $\mathcal{T}(\cdot|x,u)$ is continuous under total variation in (x,u) if $f(x,u)$ is a continuous function of (x,u), and w_t admits a continuous density function.

(iii) In general, if the transition kernel admits a continuous density function f such that $\mathcal{T}(dx_1|x,u) = f(x_1,x,u)\lambda(dx_1)$, then $\mathcal{T}(dx_1|x,u)$ is continuous in total variation. This follows from an application of Scheffé's Lemma (Billingsley, 1995, Theorem 16.12). In particular, we can write that

$$\|\mathcal{T}(\cdot|x_n,u_n) - \mathcal{T}(\cdot|x,u)\|_{TV} = \int_{\mathbb{X}} |f(x_1,x_n,u_n) - f(x_1,x,u)|\lambda(dx_1) \to 0.$$

(iv) For a model with the dynamics $x_{t+1} = f(x_t, u_t, w_t)$, if f is Lipschitz continuous in the (x,u) pair, that is, there exists some $\alpha < \infty$ such that

$$d_{\mathbb{X}}\left(f(x_n,u_n,w), f(x,u,w)\right) \leq \alpha\left(d_{\mathbb{X}}(x_n,x) + d_{\mathbb{U}}(u_n,u)\right), \tag{4}$$

we can then bound the first order Wasserstein distance between the corresponding kernels by α:

$$W_1\left(\mathcal{T}(\cdot|x_n,u_n), \mathcal{T}(\cdot|x,u)\right) = \sup_{Lip(g)\leq 1}\left|\int g(x_1)\mathcal{T}(dx_1|x_n,u_n) - \int g(x_1)\mathcal{T}(dx_1|x,u)\right|$$
$$= \sup_{Lip(g)\leq 1}\left|\int g(f(x_n,u_n,w))\nu(dw) - \int g(f(x,u,w))\nu(dw)\right|$$
$$\leq \int d_{\mathbb{X}}\left(f(x_n,u_n,w), f(x,u,w)\right)\nu(dw) \leq \alpha\left(d_{\mathbb{X}}(x_n,x) + d_{\mathbb{U}}(u_n,u)\right). \tag{5}$$

Observe that the Lipschitz bound in (4) may exhibit dependency on the realization of w, whose average growth under measure ν can lead to a more relaxed bound for the transition kernel deviation in (5).

2.2 The Average Cost Optimality Equation

Consider the following average cost problem:

$$J^*(x) := \inf_{\pi\in\Pi_A} J(x,\pi) = \inf_{\pi\in\Pi_A} \limsup_{T\to\infty} \frac{1}{T} E_x^{\pi}[\sum_{t=0}^{T-1} c(x_t,u_t)]. \tag{6}$$

To study the average cost problem, one common approach is to establish the existence of an average cost optimality equation (ACOE), and an associated verification theorem.

Definition 2 *The collection of measurable functions $\rho : \mathbb{X} \to \mathbb{R}, h : \mathbb{X} \to \mathbb{R}, f : \mathbb{X} \to \mathbb{U}$ is a canonical triplet if for all $x \in \mathbb{X}$,*

$$\rho(x) = \inf_{u \in \mathbb{U}} \int \rho(x') \mathcal{T}(dx'|x, u)$$

$$\rho(x) + h(x) = \inf_{u \in \mathbb{U}} \left(c(x, u) + \int h(x') \mathcal{T}(dx'|x, u) \right)$$

with

$$\rho(x) = \int \rho(x') \mathcal{T}(dx'|x, f(x))$$

$$\rho(x) + h(x) = \left(c(x, f(x)) + \int h(x') \mathcal{T}(dx'|x, f(x)) \right)$$

We will refer to these relations as the average cost optimality equation (ACOE).

The following verification theorem is a standard result (see Arapostathis et al., 1993; Hernández-Lerma and Lasserre, 1996)

Proposition 3 *Let ρ, h, f be a canonical triplet. If ρ is a constant and*

$$\limsup_{n \to \infty} \frac{1}{n} E_x^{\pi}[h(x_n)] = 0, \tag{7}$$

for all x and under every policy π, then the stationary deterministic policy $\pi^* = \{f, f, f, \cdots\}$ is optimal, such that

$$\rho = J(x, \pi^*) = \inf_{\pi \in \Pi_A} J(x, \pi)$$

where

$$J(x, \pi) = \limsup_{T \to \infty} \frac{1}{T} E_x^{\pi}[\sum_{k=0}^{T-1} c(x_t, u_t)].$$

In the following, we present two different approaches to establish existence of solutions to the average cost optimality problem. One approach is via a direct contraction argument, which relates average cost optimality to discounted cost optimality of an equivalent problem; and another using a Wasserstein contraction argument. These approaches will also be used to establish error bounds for the approximation methods presented in the paper.

2.2.1 CONTRACTION UNDER THE SUP NORM BY EQUIVALENCE WITH A DISCOUNTED COST PROBLEM

We have the following minorization condition.

Assumption 1 *There exists a positive measure ν such that*

$$\mathcal{T}(B|x, u) \geq \nu(B),$$

for all $B \in \mathcal{B}(\mathbb{X})$ and for all $(x, u) \in \mathbb{X} \times \mathbb{U}$.

Under Assumption 1, define $\mathcal{T}'(\cdot|x,u) = \mathcal{T}(\cdot|x,u) - \nu(\cdot)$, which is a positive measure. Then, the map

$$(\mathbb{T}'(f))(x) = \min_{u\in\mathbb{U}} \left(c(x,u) + \int f(x_1)\mathcal{T}'(dx_1|x,u) \right) \tag{8}$$

is a contraction with contraction constant $\kappa := 1 - \nu(\mathbb{X}) < 1$ (see (Hernández-Lerma, 2012, p.61) for a historical review on this approach). With this assumption, one can apply the standard value iteration algorithm using $\mathbb{T}'$. The limit equation

$$\begin{aligned} f(x) &= \min_{u\in\mathbb{U}}(c(x,u) + \int f(x_1)\mathcal{T}'(dx_1|x,u)) \\ &= \min_{u\in\mathbb{U}}(c(x,u) + \int f(x_1)\mathcal{T}(dx_1|x,u)) - \int f(x_1)\nu(dx_1) \end{aligned} \tag{9}$$

is the desired ACOE in Definition 2 with $\rho \equiv \int f(x_1)\nu(dx_1)$. The existence of a minimizing control policy is ensured by measurable selection conditions, assuming either weak continuity of the kernel (in both the state and action), or strong continuity (of the kernel in actions for every state) properties. We consider the former in the following:

Assumption 2 *(a) The one-stage cost function c is bounded and continuous.*

(b) The stochastic kernel $\mathcal{T}(\,\cdot\,|x,u)$ is weakly continuous in $(x,u) \in \mathbb{X}\times\mathbb{U}$ (that is, weak Feller).

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

(d) $\mathbb{X}$ is compact.

The corresponding measurable selection criteria are given by Himmelberg et al. (1976, Theorem 2), Schäl (1975), Schäl (1974) and Kuratowski and Ryll-Nardzewski (1965). We also refer the reader to the book by Hernández-Lerma and Lasserre (1996) for a comprehensive analysis and detailed literature review. One can then show (see e.g. Arapostathis et al., 1993; Hernández-Lerma and Lasserre, 1996, 1999; Gordienko and Hernández-Lerma, 1995; Vega-Amaya, 2003a; Demirci et al., 2004; Yüksel, 2025) that under Assumptions 1 and 2 there exists a solution to the average cost optimality equation.

2.2.2 CONTRACTION UNDER THE WASSERSTEIN-1 DISTANCE

The following is the main regularity assumption of the paper.

Assumption 3

- *The original cost function c is Lipschitz, such that $|c(x,u) - c(x',u')| \leq K_c(d_{\mathbb{X}}(x,x') + d_{\mathbb{U}}(u,u'))$ for some $K_c < \infty$ for all x, x', u, u'.*
- *The transition kernel $\mathcal{T}$ is Lipschitz continuous under the first order Wasserstein distance such that $W_1(\mathcal{T}(\cdot|x,u), \mathcal{T}(\cdot|x',u')) \leq K_{\mathcal{T}}(d_{\mathbb{X}}(x,x') + d_{\mathbb{U}}(u,u'))$ for some $K_{\mathcal{T}} < \infty$ for all x, x', u, u'.*
- *$\mathbb{X}$ and $\mathbb{U}$ are compact.*

Under Assumption 3 with $K_{\mathcal{T}} < 1$, Demirci et al. (2004) showed that the ACOE for the original model admits a solution, and that

$$\lim_{\beta \to 1} (1-\beta) J^*_\beta(x) = \rho.$$

where $J^*_\beta(x)$ denotes the optimal value function under the discounted cost criteria with discount factor $0 < \beta < 1$ (see (2)).

3. Near Optimality of Quantized State and Action Space Approximations

In the following, we build on the work of Saldi et al. (2017, 2018) to construct approximate MDPs with finite state and action spaces. Unlike Saldi et al. (2017, 2018), we require weaker conditions for the average cost criterion: notably, only weak convergence is sufficient for the approximation results, rather than the total variation continuity assumed by Saldi et al. (2017, 2018).

3.1 Finite Action Approximate MDP: Quantization of the Action Space

Under Assumption 2, the action space $\mathbb{U}$ is compact, and hence totally bounded. Therefore, one can find a finite set $\Lambda = \{u_1, \ldots, u_k\} \subset \mathbb{U}$ such that

$$\sup_{u \in \mathbb{U}} \min_{\hat{u} \in \Lambda} d_{\mathbb{U}}(u, \hat{u}) =: \mathcal{E}(\Lambda) < \infty. \tag{10}$$

Consider the MDP with tuple $(\mathbb{X}, \Lambda, \mathcal{T}, c)$. Note that under Assumption 1, both the original MDP and the finite action space MDP admit solutions to the ACOE. Furthermore, under Assumption 2 or Assumption 3, there exist optimal policies for the MDPs under both continuous and the finite action spaces. We denote the optimal values under the average cost criteria by $\rho_{\mathbb{U}}$ and ρ_Λ, respectively.

For the MDPs with the continuous and finite action spaces, we write the following ACOE's:

$$h(x) = \inf_{u \in \mathbb{U}} \left\{ c(x,u) + \int h(x_1) \mathcal{T}(dx_1|x,u) - \int h(x_1) \nu(dx_1) \right\}$$
$$\hat{h}(x) = \min_{u \in \Lambda} \left\{ c(x,u) + \int \hat{h}(x_1) \mathcal{T}(dx_1|x,u) - \int \hat{h}(x_1) \nu(dx_1) \right\}.$$

Note that $\rho_{\mathbb{U}} = \int h(x_1)\nu(dx_1)$ and $\rho_\Lambda = \int \hat{h}(x_1)\nu(dx_1)$. We follow the proof strategy by Maran et al. (2023), and use the Hölder continuity of the relative value functions for the near-optimality analysis.

Lemma 4 *Let $\kappa = 1 - \nu(\mathbb{X}) > 0$, and let D denote the diameter of the state space $\mathbb{X}$. Fix $\delta > 0$ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$. Under Assumptions 1 and 3*

$$[h]_{H^\delta} \leq \frac{K_c D^{1-\delta}}{1 - \kappa \left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta}.$$

Proof The proof can be found in Appendix A. ■

Then the following holds:

Theorem 5 *Let $\kappa = 1 - \nu(\mathbb{X}) > 0$, and let D denote the diameter of the state space $\mathbb{X}$. Fix $\delta > 0$ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$. Then the following hold:*

i. Under Assumptions 1 and 2, we have $\rho_\Lambda \to \rho_{\mathbb{U}}$ as $\mathcal{E}(\Lambda) \to 0$.

ii. Under Assumption 3, if $K_{\mathcal{T}} < 1$, then

$$\rho_\Lambda - \rho_{\mathbb{U}} \leq \frac{K_c}{1 - K_{\mathcal{T}}} \mathcal{E}(\Lambda).$$

iii. Under Assumptions 1 and 3, we have

$$\rho_\Lambda - \rho_{\mathbb{U}} \leq \frac{K_c D^{1-\delta}}{1 - \kappa \left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}} \mathcal{E}(\Lambda)^{\delta}.$$

Proof (i) follows from (Saldi et al., 2016),(Saldi et al., 2018, Theorem 3.22). For (ii) under Assumption 3 with $K_{\mathcal{T}} < 1$, one can show that (see Demirci et al. (2004))

$$\lim_{\beta \to 1} (1 - \beta) J^*_\beta(x) = \rho_{\mathbb{U}}$$
$$\lim_{\beta \to 1} (1 - \beta) \hat{J}_\beta(x) = \rho_\Lambda$$

for any $x \in \mathbb{X}$ where J^*_β (respectively $\hat{J}_\beta$) represents the optimal discounted value function under the original action space $\mathbb{U}$ (respectively under the action space Λ). Furthermore, we also have the following upper-bound for the discounted value function difference (see e.g. Saldi et al. (2018)):

$$\left| \hat{J}_\beta(x) - J^*_\beta(x) \right| \leq \frac{K_c}{(1 - \beta)(1 - \beta K_{\mathcal{T}})} \mathcal{E}(\Lambda).$$

Hence, combining these two bounds gives the desired bound. Finally, for (iii), we work with the following ACOE's

$$h(x) = \inf_{u \in \mathbb{U}} \left\{ c(x,u) + \int h(x_1) \mathcal{T}(dx_1|x,u) - \int h(x_1) \nu(dx_1) \right\}$$
$$\hat{h}(x) = \min_{u \in \Lambda} \left\{ c(x,u) + \int \hat{h}(x_1) \mathcal{T}(dx_1|x,u) - \int \hat{h}(x_1) \nu(dx_1) \right\}.$$

Note that we have $\rho_{\mathbb{U}} = \int h(x_1) \nu(dx_1)$ and $\rho_\Lambda = \int \hat{h}(x_1) \nu(dx_1)$. Thus, it suffices to find an upper bound on $\|h - \hat{h}\|_\infty$. Let $\hat{u}^*$ denote the minimizer of the second equation. Furthermore, let $\hat{u} = \arg\min_\Lambda d_{\mathbb{U}}(u^*, u)$ denote the closest element in the finite action set to the minimizer u^* of the first equation whose existence is guaranteed under Assumption 3.

We define $R(\cdot|x,u) := \frac{\mathcal{T}(\cdot|x,u) - \nu(\cdot)}{\kappa}$, with $\kappa := 1 - \nu(\mathbb{X})$ which is a Markov kernel so that $R(\mathbb{X}|x,u) = 1$. It is then easy to check that

$$W_1(R(\cdot|x,u^*), R(\cdot|x,\hat{u})) \leq \frac{K_{\mathcal{T}}}{\kappa} \mathcal{E}(\Lambda)$$

using that fact that $d_{\mathbb{U}}(u^*, \hat{u}) \leq \mathcal{E}(\Lambda)$ by construction. One can write the above ACOEs as

$$h(x) = c(x, u^*) + \kappa \int h(x_1) R(dx_1|x, u^*)$$
$$\hat{h}(x) = c(x, \hat{u}^*) + \kappa \int \hat{h}(x_1) R(dx_1|x, \hat{u}^*)$$

We note that $\hat{h} \geq h$. Since $\hat{u}$ is not the optimal action for the finite action space in general, we can then write:

$$\begin{aligned}
\hat{h}(x) - h(x) &\leq |c(x, \hat{u}) - c(x, u^*)| + \kappa \left| \int \hat{h}(x_1) R(dx_1|x, \hat{u}) - \int h(x_1) R(dx_1|x, u^*) \right| \\
&\leq K_c \mathcal{E}(\Lambda) + \kappa \left| \int \hat{h}(x_1) R(dx_1|x, \hat{u}) - \int h(x_1) R(dx_1|x, \hat{u}) \right| \\
&\qquad + \kappa \left| \int h(x_1) R(dx_1|x, \hat{u}) - \int h(x_1) R(dx_1|x, u^*) \right| \\
&\leq K_c \mathcal{E}(\Lambda) + \kappa \|h - \hat{h}\|_\infty + \kappa [h^*]_{H^\delta} \left(\frac{K_{\mathcal{T}}}{\kappa} \mathcal{E}(\Lambda) \right)^\delta
\end{aligned}$$

where $[h]_{H^\delta}$ denotes the Hölder continuity constant of h under δ. Above, we also used Lemma 1.

Writing $K_c \mathcal{E}(\Lambda) = K_c \mathcal{E}(\Lambda)^\delta \mathcal{E}(\Lambda)^{1-\delta}$, we can rearrange the terms above to get

$$\|h - \hat{h}\|_\infty \leq \frac{K_c \mathcal{E}(\Lambda)^{1-\delta} + [h]_{H^\delta} \kappa \left(\frac{K_{\mathcal{T}}}{\kappa} \right)^\delta}{1 - \kappa} \mathcal{E}(\Lambda)^\delta.$$

Using Lemma 4 for $[h]_{H^\delta}$, we write

$$\|h - \hat{h}\|_\infty \leq \frac{K_c D^{1-\delta}}{\left(1 - \kappa \left(\frac{K_{\mathcal{T}}}{\kappa} \right)^\delta \right) (1 - \kappa)} \mathcal{E}(\Lambda)^\delta. \tag{11}$$

The proof is concluded by noting that $\rho_\Lambda - \rho_{\mathbb{U}} \leq \|\hat{h} - h\|_\infty \nu(\mathbb{X})$.

■

In particular, ρ_Λ denotes the average cost of the best policy that uses actions from the finite set Λ, but evaluated under the dynamics of the original MDP. In view of the above, for the rest of the paper, we will assume that the action space $\mathbb{U}$ is finite. Error bounds for a continuous action space problem can be obtained by appropriately discretizing the action space and using Theorem 5. As such, we will replace Assumption 3 with the following:

Assumption 4

- $\mathbb{X}$ *is compact and* $\mathbb{U}$ *is finite.*
- *The original cost function* c *is Lipschitz, such that* $|c(x, u) - c(x', u)| \leq K_c d_{\mathbb{X}}(x, x')$ *with constant* $K_c < \infty$ *for all* x, x', u.
- *The transition kernel* $\mathcal{T}$ *is Lipschitz continuous under the first order Wasserstein distance such that* $W_1(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|x', u)) \leq K_{\mathcal{T}} d_{\mathbb{X}}(x, x')$ *for some* $K_{\mathcal{T}} < \infty$ *for all* x, x', u.

3.2 MDP Approximation Results

We construct an approximate MDP via state space discretization, following the construction of Saldi et al. (2017, 2018). We choose a collection of disjoint sets $\{B_i\}_{i=1}^M$ such that $\bigcup_i B_i = \mathbb{X}$, and $B_i \bigcap B_j = \emptyset$ for any $i \neq j$. We choose a representative state, $y_i \in B_i$, for each disjoint set. For this setting, we denote the new finite state space by $\mathbb{Y} := \{y_1, \ldots, y_M\}$, and the mapping from the original state space $\mathbb{X}$ to the finite set $\mathbb{Y}$ is done via

$$q(x) = y_i \quad \text{if } x \in B_i. \tag{12}$$

Furthermore, we choose a weight measure $\mu \in \mathcal{P}(\mathbb{X})$ on $\mathbb{X}$ such that $\mu(B_i) > 0$ for all B_i. We now define normalized measures using the weight measure on each separate quantization bin B_i as follows:

$$\mu_{y_i}(A) := \frac{\mu(A)}{\mu(B_i)}, \quad \forall \text{ Borel } A \subset B_i, \quad \forall i \in \{1, \ldots, M\}, \tag{13}$$

that is, μ_{y_i} is the normalized weight measure on the set B_i, where y_i belongs to. We then define the cost function and transition kernel for the approximate model as follows:

$$\begin{aligned} \hat{c}(x,u) &= \int_{B_i} c(z,u)\mu_{y_i}(dz) \\ \hat{\mathcal{T}}(A|x,u) &= \int_{B_i} \mathcal{T}(A|z,u)\, \mu_{y_i}(dz) \end{aligned} \tag{14}$$

for any $A \in \mathcal{B}(\mathbb{X})$ and any $x \in B_i$. We denote the optimal average cost for this approximate model by $\hat{J}(x)$ given the initial condition is x.

Note that although the approximate model is defined over the original state space $\mathbb{X}$, by construction its cost function and transition kernel are constant over the quantization bins $\{B_i\}_{i=1}^M$. Hence, the approximate model can equivalently be viewed as a finite-state MDP with state space $\mathbb{Y} = \{y_1, \ldots, y_M\}$ where y_i is the representative state of the quantization bin B_i. Indeed, for any $y_i, y_j \in \mathbb{Y}$ and $u \in \mathbb{U}$, the stage-wise cost and the transition kernel for the finite-state model are defined as

$$\begin{aligned} C(y_i,u) &:= \int_{B_i} c(x,u)\, \mu_{y_i}(dx), \\ P(y_j|y_i,u) &:= \int_{B_i} \mathcal{T}(B_j|x,u)\, \mu_{y_i}(dx). \end{aligned} \tag{15}$$

In the following, we establish approximation and performance results for control policies obtained via the approximate MDPs when applied to the true model.

3.2.1 MDP APPROXIMATION VIA WASSERSTEIN CONTINUITY WITH MODULUS OF CONTINUITY IN APPROXIMATION

The main goal of this section is to establish the near-optimality of policies designed for a finite MDP when applied to the continuous model. Denoting the optimal policy of the finite model by $\hat{\pi}$, we aim to find upper bounds on

$$J(x, \hat{\pi}) - J^*(x)$$

where $J^*(x)$ is the optimal value under the ergodic cost criterion for the original model. To analyze this term, we add and subtract the optimal value of the approximate model, i.e. $\hat{J}(x)$ yielding

$$\left(J(x,\hat{\pi}) - \hat{J}(x)\right) + \left(\hat{J}(x) - J^*(x)\right). \tag{16}$$

The second term is the difference between the value functions of the approximate and the original MDPs, and we first bound this difference as an intermediate step. The first term (which we may refer to as the robustness error term) measures the difference in the expected cost between the approximate and original MDP models under the *same* policy, namely, the optimal policy of the discretized model. Our goal is to derive error bounds in terms of the Wasserstein–1 Lipschitz continuity of the original transition kernel using the approximate MDP. Therefore, unless the optimal policy is Lipschitz continuous, it is generally impossible to bound this term uniformly over policies. Restricting policies to be Lipschitz a priori has been previously considered in the literature (see e.g. Pirotta et al., 2015; Asadi et al., 2018; Rachelson and Lagoudakis, 2010; Maran et al., 2023). Nonetheless, we will show that when the policy is an optimal policy of the approximate MDP, it is indeed possible to derive an error bound for the second term that vanishes as the discretization becomes finer (the principal reason being that the state and corresponding optimal action pairs converge to one another due to optimality which, together with continuous weak convergence of kernels, leads to the vanishing error). We note that, as is often the case for finite space MDPs, if the original model is instead approximated in total variation distance, the value difference under the same policy can be directly bounded for any policy due to the strong regularity imposed by total variation (see Kara and Yüksel (2020, Section 4.5) and Jiang (2018)). However, for general space MDPs, such a direct condition is too demanding.

We define the following loss functions induced by the quantization on the cost function and the transition kernel:

$$\begin{aligned} L_c(x,u) &:= |c(x,u) - \hat{c}(x,u)| \\ L_T(x,u) &:= W_1\left(\mathcal{T}(\cdot|x,u), \hat{\mathcal{T}}(\cdot|x,u)\right) \end{aligned} \tag{17}$$

where $\hat{c}$ and $\hat{\mathcal{T}}$ are the cost function and transition kernel of the approximate model defined in (14).

Value Function Difference. For the value difference bound, that is the second term in (16), we give two alternatives: one assuming the minorization condition Assumption 1 and one without minorization but using Wasserstein contraction of the kernel with $K_{\mathcal{T}} < 1$ under Assumption 4.

Proposition 6 *Consider the optimal average-cost values of the original and the approximate models, denoted by $J^*(x)$ and $\hat{J}(x)$, respectively, for a given initial state x. Under Assumption 1 and Assumption 4, these values are constant with $\rho = J^*(x)$ and $\hat{\rho} = \hat{J}(x)$ for all $x \in \mathbb{X}$. Fix δ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$ where $\kappa := 1 - \nu(\mathbb{X}) > 0$, and let D denote the diameter of the state space. Then, we have*

$$|\rho - \hat{\rho}| \leq \|L_c\|_\infty + \left(\frac{K_c D^{1-\delta}\kappa^{1-\delta}}{1 - \kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}}\right)\left(\|L_T\|_\infty\right)^{\delta}.$$

Proof The proof uses a similar technique to that of Theorem 5. We will work with the following ACOE's

$$h(x) = \min_u \left\{ c(x,u) + \int h(x_1)\mathcal{T}(dx_1|x,u) - \int h(x_1)\nu(dx_1) \right\}$$
$$\hat{h}(x) := \min_u \left\{ \hat{c}(x,u) + \int \hat{h}(x_1)\hat{\mathcal{T}}(dx_1|x,u) - \int \hat{h}(x_1)\nu(dx_1) \right\}$$

Note that we have $\rho = \int h(x_1)\nu(dx_1)$ and $\hat{\rho} = \int \hat{h}(x_1)\nu(dx_1)$. Thus, it suffices to find an upper bound on $\|h - \hat{h}\|_\infty$. Denoting by $\kappa = 1 - \nu(\mathbb{X})$, one can write the above as

$$h(x) = \min_u \left\{ c(x,u) + \kappa \int h(x_1)R(dx_1|x,u) \right\}$$
$$\hat{h}(x) := \min_u \left\{ \hat{c}(x,u) + \kappa \int \hat{h}(x_1)\hat{R}(dx_1|x,u) \right\}$$

where $R(\cdot|x,u) := \frac{\mathcal{T}(\cdot|x,u) - \nu(\cdot)}{\kappa}$ is a Markov kernel so that $R(\mathbb{X}|x,u) = 1$. $\hat{R}$ is defined similarly. It is then easy to check by definition that for any x, u

$$W_1(R(\cdot|x,u), \hat{R}(\cdot|x,u)) \leq \frac{L_T(x,u)}{\kappa}.$$

Using Lemma 1, we can show the following:

Lemma 7

$$\|h - \hat{h}\|_\infty \leq \frac{1}{1-\kappa} \left(\|L_c\|_\infty + \left(\frac{K_c D^{1-\delta} \kappa^{1-\delta}}{1 - \kappa \left(\frac{K_{\mathcal{T}}}{\kappa} \right)^\delta} \right) (\|L_T\|_\infty)^\delta \right).$$

Proof

$$\begin{aligned} |h(x) - \hat{h}(x)| \leq & \sup_u |c(x,u) - \hat{c}(x,u)| \\ & + \kappa \|h - \hat{h}\|_\infty + \kappa \sup_u \left\{ \int h(x_1)R(dx_1|x,u) - \int h(x_1)\hat{R}(dx_1|x,u) \right\} \\ & \leq \|L_c\|_\infty + \kappa \|h - \hat{h}\|_\infty + \kappa [h]_{H^\delta} \left(\frac{\|L_T\|_\infty}{\kappa} \right)^\delta \end{aligned}$$

where $[h]_{H^\delta}$ denotes the Hölder continuity constant of h under δ. Rearranging the terms and using the Lemma 4 to bound $[h]_{H^\delta}$ give the desired bound. ■

We conclude the proof noting $|\rho - \hat{\rho}| \leq \|h - \hat{h}\|_\infty \nu(\mathbb{X})$ and that $\nu(\mathbb{X}) = 1 - \kappa$. ■

The following proposition shows that if the transition kernel satisfies a Wasserstein contraction, the minorization condition is not needed to bound the value difference:

Proposition 8 *Under Assumption 4 with $K_{\mathcal{T}} < 1$, we have*

$$|\rho - \hat{J}(x)| \leq \|L_c\|_\infty + \frac{K_c}{1-K_{\mathcal{T}}}\|L_T\|_\infty$$

where $\rho = J^*(x)$ is the optimal value of the original MDP, and $\hat{J}(x)$ is the optimal value of the approximate model for the initial state x.

Remark 9 *We note that Assumption 4 with $K_{\mathcal{T}} < 1$ guarantees that the optimal value of the original model is constant, such that $J^*(x) = \rho$, for all $x \in \mathbb{X}$. However, the same may not hold for the approximate model (see Example 2).*

Proof We first note that under Assumption 4 with $K_{\mathcal{T}} < 1$, one can show that (see Demirci et al. (2004) or Theorem 7.3.4 of Yüksel (2025)) the ACOE for the original model admits a solution and that

$$\lim_{\beta \to 1}(1-\beta)J^*_\beta(x) = \rho.$$

Furthermore, we also have that for any finite MDP, and in particular, for the approximate model, we also have that

$$\lim_{\beta \to 1}(1-\beta)\hat{J}_\beta(x) = \hat{J}(x).$$

From Theorem 5 of Kara et al. (2023), it follows that

$$\left|J^*_\beta(x) - \hat{J}_\beta(x)\right| \leq \frac{1}{1-\beta}\left[\|L_c\|_\infty + \frac{K_c}{1-\beta K_{\mathcal{T}}}\|L_T\|_\infty\right].$$

Hence, we can conclude the proof with a triangle inequality:

$$|J^*(x) - \hat{J}(x)| \leq \lim_{\beta \to 1}|(1-\beta)J^*_\beta(x) - (1-\beta)\hat{J}_\beta(x)| \leq \|L_c\|_\infty + \frac{K_c}{1-K_{\mathcal{T}}}\|L_T\|_\infty.$$

■

On Minorization vs. Wasserstein Contraction Conditions. To bound the difference between the value functions of the original MDP and the approximate MDP, we have considered two alternative conditions:

1. The transitions are Wasserstein contractive in the sense that $W_1(\mathcal{T}(\cdot|x,u), \mathcal{T}(\cdot|y,u)) \leq K_{\mathcal{T}} d_{\mathbb{X}}(x,y)$ with $K_{\mathcal{T}} < 1$.
2. The transitions are Lipschitz continuous and they satisfy a minorization condition, i.e., $\mathcal{T}(\cdot|x,u) \geq \nu(\cdot)$ for a non-trivial measure $\nu(\cdot)$.

The first condition implies that the dynamics are contractive on average, which is analogous to contractive behavior in deterministic systems. In particular, under a functional representation of the dynamics via stochastic realizability conditions, we can write the dynamics as

$$X_{t+1} = f(X_t, U_t, W_t)$$

for some measurable f, and for an i.i.d. noise sequence W_t distributed according to a measure $P(\cdot)$ such that the pushforward of $P(\cdot)$ under $f(x,u,\cdot)$ is the kernel $\mathcal{T}(\cdot|x,u)$. In this representation, if we assume that there exists some function $K : \mathbb{W} \times \mathbb{U} \to [0,\infty)$ such that for all $x, y \in \mathbb{X}$, $u \in \mathbb{U}$, $w \in \mathbb{W}$,

$$d_{\mathbb{X}}\left(f(x,u,w), f(y,u,w)\right) \leq K(w,u) d_{\mathbb{X}}(x,y),$$

we then have

$$\begin{aligned} W_1(\mathcal{T}(\cdot|x,u), \mathcal{T}(\cdot|y,u)) &\leq \sup_{\|g\|_{Lip}\leq 1} \left| \int g(f(x,u,w)) - g(f(y,u,w)) \right| P(dw) \\ &\leq \int K(w,u) P(dw) d_{\mathbb{X}}(x,y). \end{aligned}$$

Thus, the Wasserstein contraction assumption for the kernel holds if $\int K(w,u)P(dw) < 1$, that is if the dynamics are contractive on average for different realizations of the noise.

The second condition indicates that the Wasserstein contraction condition can be avoided if the dynamics are mixing, i.e., $\mathcal{T}(\cdot|x,u) \geq \nu(\cdot)$ for a non-trivial measure $\nu(\cdot)$.

We note that the condition $K_{\mathcal{T}} < 1$ guarantees that the average-cost optimal value function for the original model does not depend on the initial state. However, this property may no longer hold for the finite model after discretization. Consider the following simple example:

Example 2 *Let $\mathbb{X} = [-1,1]$, and let the dynamics be uncontrolled and given by*

$$X_{t+1} = \frac{X_t}{2}.$$

The cost is simply equal to the state, that is, $c(x) = x$. It is easy to see that the infinite horizon average cost for this problem is given by $\rho = 0$ independent of the initial condition. Furthermore, we have that

$$(1-\beta)J_\beta(x) = (1-\beta)\sum_{t=0}^{\infty} \beta^t \left(\frac{1}{2}\right)^t x = (1-\beta)\frac{1}{1-\beta/2}x \to 0$$

independent of the initial state x as $\beta \to 1$. On the other hand, if we discretize the state space by mapping $[-1,0) \to -1$ and $[0,1] \to 1$, the infinite horizon average cost becomes $\hat{\rho}(-1) = -1$ and $\hat{\rho}(1) = 1$ which depends on the initial state.

For the minorization condition, we first note that if the original model kernel satisfies this condition, then the finite model kernel also satisfies the same condition:

Lemma 10 *If $\mathcal{T}(\cdot|x,u) \geq \nu(\cdot)$ for a non-trivial measure $\nu(\cdot)$, then the transition kernel for the approximate model also satisfies the same condition such that $\hat{\mathcal{T}}(\cdot|x,u) \geq \nu(\cdot)$ where $\hat{\mathcal{T}}$ is defined in (14). Consequently, for the transition probabilities of the finite model defined in (15), there exists some y_j such that $P(y_j|y_0,u_0) > 0$ for all $y_0 \in \mathbb{Y}$ and $u_0 \in \mathbb{U}$.*

Proof Fix $(x,u) \in \mathbb{X} \times \mathbb{U}$. Take i such that $x \in B_i$. Then

$$\hat{\mathcal{T}}(\cdot|x,u) = \int_{B_i} \mathcal{T}(\cdot|z,u)\mu_{y_i}(dz) \geq \int \nu(\cdot)\mu_{y_i}(dz) = \nu(\cdot).$$

Since $\nu(\cdot)$ is non-trivial, that is since $\nu(\mathbb{X}) > 0$, for any discretization scheme there exists some index j such that $\nu(B_j) > 0$, and thus for any $(y_0, u_0) \in \mathbb{Y} \times \mathbb{U}$,

$$P(y_j|y_0, u_0) = \hat{\mathcal{T}}(B_j|y_0, u_0) \geq \nu(B_j) > 0.$$

■

Even when the original model does not satisfy the minorization condition, it is sometimes possible to construct a discretization scheme which does satisfy the minorization for any rate of discretization:

Example 3 *Let $\mathbb{X} = [-2, 2]$ and consider the dynamics $x_{k+1} = \frac{x_k}{2} + w_k$ where w_k is supported on $\mathbb{Q} \cap [-1, 1]$ where $\mathbb{Q}$ is the set of rationals. For this example, we have that $K_{\mathcal{T}} = \frac{1}{2}$. The kernel $\mathcal{T}$, however, does not satisfy the minorization condition. Indeed, for $x, x' \in \mathbb{X}$*

$$\text{supp}(\mathcal{T}(\cdot|x)) = \frac{1}{2}x + (\mathbb{Q} \cap [-1, 1]).$$

In particular, we have that

$$\text{supp}(\mathcal{T}(\cdot|x)) \cap \text{supp}(\mathcal{T}(\cdot|x')) = \emptyset \text{ when } |x - x'| \notin \mathbb{Q}.$$

Thus, no uniform minorization measure can exist. On the other hand, for any finite partition of $\mathbb{X}$ that has a bin which includes an open neighborhood of the point 0, this bin is visited with positive probability from every other bin. Therefore, the transition kernel of any approximate (finite-state) model based on such a partition satisfies a minorization condition.

Near Optimality of Approximate Policies. The previous results give us an upper bound on the difference between the optimal average-cost value functions. We now focus on the performance of policies designed for the approximate models. In this setting, we note that there might be several policies that achieve the optimal performance for the discretized model under the average cost optimality criterion, and these policies may perform differently when applied to the original problem (see (Kara and Yüksel, 2020, Section 4)).

Therefore, we work with policies that satisfy the ACOE. In particular, we will find performance loss bounds for a policy $\hat{\pi}$ that satisfies:

$$\begin{aligned} \hat{\rho} + \hat{h}(x) &= \hat{c}(x, \hat{\pi}(x)) + \int \hat{h}(x_1)\hat{\mathcal{T}}(dx_1|x, \hat{\pi}(x)) \\ &= \min_{u \in \mathbb{U}} \left\{ \hat{c}(x, u) + \int \hat{h}(x_1)\hat{\mathcal{T}}(dx_1|x, u) \right\}. \end{aligned} \tag{18}$$

Recall from (16), that the performance of an approximate policy is related to the difference $J(x, \hat{\pi}) - \hat{J}(x)$ as well as the corresponding value-function difference, for which we have derived upper bounds. It is not immediately clear, however, how to control this term when the policy $\hat{\pi}$ may be discontinuous and the distance between the transition kernels is controlled in the W_1 distance only.

Theorem 11 *Suppose $\hat{\pi}$ satisfies (18) and is optimal for the approximate model. For a given initial state $x_0 \in \mathbb{X}$, consider the average-cost of $\hat{\pi}$ in the original MDP denoted by $J(x_0, \hat{\pi})$, and the optimal average-cost value of the original MDP denoted by $J^*(x_0)$.*

Under Assumptions 1 and 4, these values are independent of the initial state with $\rho(\hat{\pi}) = J(x_0, \hat{\pi})$ *and* $\rho = J^*(x_0)$*, for all* $x_0 \in \mathbb{X}$*. Fix* δ *such that* $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$ *where* $\kappa := 1 - \nu(\mathbb{X}) > 0$*, and let* D *denote the diameter of the state space. Then, we have*

$$\rho(\hat{\pi}) - \rho \leq \frac{2}{1-\kappa}\left(\|L_c\|_\infty + \left(\frac{K_c D^{1-\delta}\kappa^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta}\right)(\|L_T\|_\infty)^\delta\right).$$

Proof We start by noting that under Assumption 1 and using Lemma 10, we have

$$\mathcal{T}(\cdot|x,u) \geq \nu(\cdot), \qquad \hat{\mathcal{T}}(\cdot|x,u) \geq \nu(\cdot). \tag{19}$$

In particular, if we define the following operators

$$\begin{aligned}
Th(x) &:= c(x,\hat{\pi}(x)) + \int h(x_1)\mathcal{T}(dx_1|x,\hat{\pi}(x)) - \int h(x_1)\nu(dx_1)\\
\hat{T}h(x) &:= \hat{c}(x,\hat{\pi}(x)) + \int h(x_1)\hat{\mathcal{T}}(dx_1|x,\hat{\pi}(x)) - \int h(x_1)\nu(dx_1)\\
T^*h(x) &:= c(x,\pi^*(x)) + \int h(x_1)\mathcal{T}(dx_1|x,\pi^*(x)) - \int h(x_1)\nu(dx_1),
\end{aligned} \tag{20}$$

where π^* is optimal for the average cost problem and solves the ACOE. Using (19), one can then show (see (8)), that these operators are contractions with contraction constant $\kappa := 1 - \nu(\mathbb{X}) < 1$.

Furthermore, these operators admit unique fixed points, say $h(x), \hat{h}(x), h^*(x)$ respectively. As in the proof of Theorem 6, using the kernels $R(\cdot|x,u) = \frac{\mathcal{T}(\cdot|x,u)-\nu(\cdot)}{\kappa}$ and $\hat{R}(\cdot|x,u) = \frac{\hat{\mathcal{T}}(\cdot|x,u)-\nu(\cdot)}{\kappa}$ we write these fixed point equations as:

$$\begin{aligned}
h(x) &= c(x,\hat{\pi}(x)) + \kappa\int h(x_1)R(dx_1|x,\hat{\pi}(x))\\
\hat{h}(x) &= \hat{c}(x,\hat{\pi}(x)) + \kappa\int \hat{h}(x_1)\hat{R}(dx_1|x,\hat{\pi}(x))\\
h^*(x) &= c(x,\pi^*(x)) + \kappa\int h^*(x_1)R(dx_1|x,\pi^*(x)).
\end{aligned} \tag{21}$$

Note that the above equations are in the same form as ACOEs, and hence, we have that $J(x,\hat{\pi}) = \rho(\hat{\pi}) = \int h(x_1)\nu(dx_1)$, $\hat{J}(x) = \hat{\rho} = \int \hat{h}(x_1)\nu(dx_1)$, and $J^*(x) = \rho = \int h^*(x_1)\nu(dx_1)$. We now write

$$J(x,\hat{\pi}) - J^*(x) \leq \left|J(x,\hat{\pi}) - \hat{J}(x)\right| + \left|\hat{J}(x) - J^*(x)\right|.$$

For the first term, we first study the difference $|h(x) - \hat{h}(x)|$:

$$\begin{aligned}
&|h(x) - \hat{h}(x)| = |Th(x) - \hat{T}\hat{h}(x)|\\
&\leq |Th(x) - T\hat{h}(x)| + |T\hat{h}(x) - Th^*(x)| + |Th^*(x) - \hat{T}h^*(x)| + |\hat{T}h^*(x) - \hat{T}\hat{h}(x)|\\
&\leq \kappa\|h - \hat{h}\|_\infty + \kappa\|\hat{h} - h^*\|_\infty + \|L_c\|_\infty + \kappa[h]_{H^\delta}\left(\frac{\|L_T\|_\infty}{\kappa}\right)^\delta + \kappa\|h^* - \hat{h}\|_\infty
\end{aligned} \tag{22}$$

where we used the fact that the used operators are contractions. Furthermore, we used the following upper bound

$$\begin{aligned}
|Th^*(x) - \hat{T}h^*(x)| \leq \Bigg| & c(x,\hat{\pi}(x)) + \kappa \int h^*(x_1) R(dx_1|x,\hat{\pi}(x)) \\
& - \hat{c}(x,\hat{\pi}(x)) - \kappa \int h^*(x_1) \hat{R}(dx_1|x,\hat{\pi}(x)) \Bigg| \\
\leq & \|L_c\|_\infty + \kappa [h^*]_{H^\delta} \left(W_1(R(\cdot|x,\hat{\pi}(x)), \hat{R}(\cdot|x,\hat{\pi}(x))) \right)^\delta \\
\leq & \|L_c\|_\infty + \kappa [h^*]_{H^\delta} \left(\frac{W_1(\mathcal{T}(\cdot|x,\hat{\pi}(x)), \hat{\mathcal{T}}(\cdot|x,\hat{\pi}(x)))}{\kappa} \right)^\delta \\
\leq & \|L_c\|_\infty + \kappa [h^*]_{H^\delta} \left(\frac{\|L_T\|_\infty}{\kappa} \right)^\delta
\end{aligned}$$

where we used Lemma 1. We note that by Lemma 4, $[h^*]_{H^\delta} \leq \frac{K_c D^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta}$. Using Lemma 7, we also have that

$$\|h^* - \hat{h}\|_\infty \leq \frac{1}{1-\kappa} \left(\|L_c\|_\infty + \left(\frac{K_c D^{1-\delta} \kappa^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta} \right) (\|L_T\|_\infty)^\delta \right).$$

Combining the above arguments with (22), we obtain

$$\|h - \hat{h}\|_\infty \leq \frac{1+\kappa}{(1-\kappa)^2} \left(\|L_c\|_\infty + \left(\frac{K_c D^{1-\delta} \kappa^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta} \right) (\|L_T\|_\infty)^\delta \right).$$

For the difference $|\rho(\hat{\pi}) - \hat{\rho}|$, we then have

$$\begin{aligned}
|\rho(\hat{\pi}) - \hat{\rho}| = \left| \int h(x_1)\nu(dx_1) - \int \hat{h}(x_1)\nu(dx_1) \right| & \leq \|h - \hat{h}\|_\infty \nu(\mathbb{X}) \\
& \leq \frac{1+\kappa}{(1-\kappa)} \left(\|L_c\|_\infty + \left(\frac{K_c D^{1-\delta} \kappa^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^\delta} \right) (\|L_T\|_\infty)^\delta \right)
\end{aligned}$$

where we used the identity $\nu(\mathbb{X}) = 1 - \kappa$. Finally, combining this bound with the estimates for $|\rho - \hat{\rho}|$ obtained in Theorem 6, the proof is complete. ■

Further Bounds in terms of State-Space Quantization Error. The performance error of an approximate policy is mainly controlled by the quantities

$$\begin{aligned}
L_c(x,u) &:= |c(x,u) - \hat{c}(x,u)| \\
L_T(x,u) &:= W_1\left(\mathcal{T}(\cdot|x,u), \hat{\mathcal{T}}(\cdot|x,u) \right).
\end{aligned}$$

This suggests that, in order to minimize the loss, the quantization bins should be chosen so that the deviation of the cost function and the transition kernels is small within each bin.

For the cost function, this can be achieved by constructing quantization bins aligned with the level (or sublevel) sets of $c(x,u)$, so that the cost is nearly constant within each bin. However, the transition kernel maps states and actions to the set of probability measures whose pre-image map analysis is far more tedious unless there is further structure on the dynamics. Hence, following the same approach is not effective for the transition kernel. Accordingly, instead of partitioning in terms of the pre-image of the partitions on the space of probability measures, we partition the domain (that is, the state and action sets): The following results show that the error can be made arbitrarily small by choosing a sufficiently high rate of quantization. We define an average loss function $L : \mathbb{X} \to \mathbb{R}$ induced by the quantization: for $x \in \mathbb{X}$ belonging to a quantization bin B_i whose representative state is y_i (i.e., $q(x) = y_i$), a weighted loss function $L(x)$ is defined as

$$L(x) := \int_{B_i} d_{\mathbb{X}}(x, x')\, \mu_{y_i}(dx'). \tag{23}$$

That is, $L(x)$ can be seen as the mean distance from x to the bin B_i under the measure μ_{y_i}. We denote the uniform bound on the quantization error by $L_{\mathbb{X}}$ defined as

$$L_{\mathbb{X}} := \sup_x L(x).$$

We have the following immediate result:

Lemma 12 *Under Assumption 4, we have that for all $x, u \in \mathbb{X} \times \mathbb{U}$*

$$L_c(x,u) = |\hat{c}(x,u) - c(x,u)| \leq K_c L_{\mathbb{X}},$$
$$L_T(x,u) = W_1(\hat{\mathcal{T}}(\cdot|x,u), \mathcal{T}(\cdot|x,u)) \leq K_{\mathcal{T}} L_{\mathbb{X}}.$$

Proof Let $x \in B_i$. For the cost difference, we write

$$|\hat{c}(x,u) - c(x,u)| = \left| \int_{B_i} c(x',u)\mu_{y_i}(dx') - c(x,u) \right| = \left| \int_{B_i} c(x',u) - c(x,u)\mu_{y_i}(dx') \right|$$
$$\leq \int_{B_i} K_c d_{\mathbb{X}}(x,x')\mu_{y_i}(dx') \leq K_c L_{\mathbb{X}}.$$

For the transition difference, for any $\|f\|_{Lip} \leq 1$, we similarly write

$$\left| \int f(x_1)\hat{\mathcal{T}}(dx_1|x,u) - \int f(x_1)\mathcal{T}(dx_1|x,u) \right|$$
$$= \left| \int \int_{B_i} f(x_1)\mathcal{T}(dx_1|x',u)\mu_{y_i}(dx') - \int f(x_1)\mathcal{T}(dx_1|x,u) \right|$$
$$\leq \int_{B_i} K_{\mathcal{T}} d_{\mathbb{X}}(x,x')\mu_{y_i}(dx') \leq K_{\mathcal{T}} L_{\mathbb{X}}.$$

■

Corollary 13 *(of Theorem 11) Suppose $\hat{\pi}$ satisfies (18). Under Assumptions 1 and 4, for any δ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$, we have that*

$$\rho(\hat{\pi}) - \rho \leq \frac{2K_c D^{1-\delta}}{(1-\kappa)\left(1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}\right)} L_{\mathbb{X}}^{\delta}.$$

Remark 14 *The result is stated for finite action spaces. However, we can easily get a further upper bound for continuous action spaces by combining Theorem 5 and Theorem 11 (or Corollary 13). Let $\hat{\pi}$ denote the policy designed for discretized action and state spaces, and let $\rho(\hat{\pi})$ denote the average cost we would receive if we applied the policy $\hat{\pi}$ to the original model. ρ is the optimal value for the original state and action spaces, and ρ' denotes the value of the problem with a discretized action space and a continuous state space. We can then write*

$$\rho(\hat{\pi}) - \rho \leq \left|\rho(\hat{\pi}) - \rho'\right| + \left|\rho' - \rho\right| \leq \frac{2K_c D^{1-\delta}}{(1-\kappa)\left(1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}\right)} L_{\mathbb{X}}^{\delta} + \frac{K_c D^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}} \mathcal{E}(\Lambda)^{\delta}$$

where we have used Corollary 13 for the first term, and Theorem 5(iii) for the second term (see (10) for the definition of $\mathcal{E}(\Lambda)$).

3.2.2 FINITE APPROXIMATIONS VIA WEAK CONTINUITY AND ASYMPTOTIC OPTIMALITY

In Section 3.2.1, we used Assumption 4 to obtain quantitative bounds on approximation errors using the Lipschitz continuity of the cost function and the transition kernel. In this section, we relax the Lipschitz continuity assumption and focus on asymptotic analysis only under continuity of the cost, and weak continuity of the kernel. We emphasize that Lipschitz continuity conditions are generally more restrictive; for example in the belief-MDP reduction for partially observable Markov Decision Processes (POMDPs), there is a significant difference between models which allow for weak Feller continuity vs. those which satisfy Lipschitz regularity (see the tutorial paper Kara and Yüksel (2024b)); an analogous situation applies for problems where lifting to a larger space (such as probability measure valued dynamics) is utilized as in decentralized stochastic control problems, as well as mean-field models. In this subsection, we work with Assumption 2 specialized to finite action spaces:

Assumption 5 *(a) $\mathbb{U}$ is finite.*

(b) $\mathbb{X}$ is compact.

(c) The one-stage cost function c is bounded and continuous.

(d) The stochastic kernel $\mathcal{T}(\,\cdot\,|x,u)$ is weakly continuous in $(x,u) \in \mathbb{X} \times \mathbb{U}$ (that is, weak Feller continuous).

In this section, we will apply a uniform quantization of the compact state space with bin diameter $\frac{1}{n}$ so that

$$L_{\mathbb{X}} = \sup_x L(x) \leq \frac{1}{n}$$

where $L(x)$ is defined in (23).

In Section 4.2.2 of Saldi et al. (2018) total variation continuity was imposed for near optimality of quantized models under the average cost criterion. In the previous subsection, this was relaxed to Wasserstein continuity. The convergence result along the same lines can be shown to work under only Assumptions 1 and 5, albeit without a modulus of continuity.

Theorem 15 *Let π_n denote an optimal policy for the quantized model which solves the ACOE for the approximate model with discretization width $1/n$. For a given initial state $x_0 \in \mathbb{X}$, let $J(x_0, \pi_n)$ denote the average cost of the policy π_n evaluated under the original model and let $J^*(x_0)$ denote the optimal average cost.*

Under Assumption 1, these values are independent of the initial state; we write $\rho(\pi_n) = J(x_0, \pi_n)$ and $\rho = J^*(x_0)$ for all $x_0 \in \mathbb{X}$. Then, under Assumptions 1 and 5, we have

$$\lim_{n\to\infty} \rho(\pi_n) = \rho.$$

Proof The proof relies on the contraction operators and the corresponding fixed point equations used in the proof of Theorem 11, i.e. (21). In particular, we consider the fixed point equations:

$$h_n(x) = c(x, \pi_n(x)) + \kappa \int h_n(x_1) R(dx_1|x, \pi_n(x))$$
$$h^*(x) = c(x, \pi^*(x)) + \kappa \int h^*(x_1) R(dx_1|x, \pi^*(x))$$

where $R(\cdot|x,u) := \frac{\mathcal{T}(\cdot|x,u) - \nu(\cdot)}{\kappa}$ with $\kappa = 1 - \nu(\cdot) < 1$. Note that π_n is an optimal policy for the approximate cost function c_n and the approximate kernel $\mathcal{T}_n$ under the average cost criterion and equivalently under a discounted cost criterion with discount factor κ. Furthermore, the fixed point h_n corresponds to the discounted cost value of the policy π_n under the original model with a discount factor κ, and h^* corresponds to the optimal discounted value function for the original dynamics under the discount factor κ.

We can then use Theorem 4.4 of Kara and Yüksel (2020) to show that $h_n \to h^*$ point-wise under Assumption 5, if

- $R_n(\cdot|x_n, u) \to R(\cdot|x, u)$ weakly for any $x_n \to x$ and for all u
- $c_n(x_n, u) \to c(x, u)$ for any $x_n \to x$ and for all u.

For the first item, it is equivalent to show that $\mathcal{T}_n(\cdot|x_n, u) \to \mathcal{T}(\cdot|x, u)$ weakly. For the cost function we have that:

$$c_n(x_n, u) = \int_{B_{n,i}} c(z, u) \mu_n(dz)$$

where $B_{n,i}$ denotes the quantization bin x_n belongs to and μ_n is the weight measure μ concentrated on the set $B_{i,n}$. Thus, we need to show that for any fixed $\epsilon > 0$, we can find a large enough $N < \infty$ such that for $n > N$, we have that $|c(z, u) - c(x, u)| < \epsilon$ for all $z \in B_{i,n}$.

For fixed $\epsilon > 0$, we can find $\delta > 0$ such that $|c(x, u) - c(z, u)| < \epsilon$ for all $d_{\mathbb{X}}(x, z) < \delta$ since $c(x, u)$ is continuous by assumption. Thus, we now want to find a sufficiently large $N < \infty$ such that for such a δ, $d_{\mathbb{X}}(x, z) < \delta$ for all $z \in B_{i,n}$ for $n \geq N$. Recall that $B_{n,i}$ represents the

quantization bin x_n belongs to, and by construction we have that $d_{\mathbb{X}}(z, x_n) \leq \frac{1}{n}$ for all $z \in B_{i,n}$ which can be made smaller than $\delta/2$ for all $n \geq N_1$ for a sufficiently large N_1. Furthermore, $x_n \to x$ by assumption, and thus we can make $d_{\mathbb{X}}(x_n, x) < \delta/2$ for all $n \geq N_2$ for some other sufficiently large N_2. Picking the greater of N_1 and N_2 implies that

$$d_{\mathbb{X}}(x, z) \leq d_{\mathbb{X}}(x, x_n) + d_{\mathbb{X}}(x_n, z) < \delta$$

for all $n \geq \max(N_1, N_2)$ and for all $z \in B_{i,n}$, which proves the claim that $c_n(x_n, u) \to c(x, u)$ for all $x_n \to x$.

Using identical arguments and noting that for any continuous and bounded f, we have that

$$\begin{aligned}
&\left| \int f(x_1) \mathcal{T}_n(dx_1|x_n, u) - \int f(x_1) \mathcal{T}(dx_1|x, u) \right| \\
&= \left| \int_{B_{n,i}} \int f(x_1) \mathcal{T}_n(dx_1|z, u) \mu_n(dz) - \int f(x_1) \mathcal{T}(dx_1|x, u) \right|
\end{aligned}$$

we can also conclude that $\mathcal{T}_n(\cdot|x_n, u) \to \mathcal{T}(\cdot|x, u)$ weakly for all $x_n \to x$. By Theorem 4.4 of Kara and Yüksel (2020), this shows that $h_n \to h^*$.

We conclude the proof by using the Dominated Convergence Theorem and noting that $\rho(\pi_n) = \int h_n(x_1)\nu(dx_1)$ and $\rho = \int h^*(x_1)\nu(dx_1)$. ■

We now summarize the main results of this section as follows:

Theorem 16 *Suppose $\hat{\pi}$ satisfies (18) and is optimal for the approximate model. For a given initial state $x_0 \in \mathbb{X}$, let $J(x_0, \hat{\pi})$ denote the average cost of $\hat{\pi}$ in the original MDP, $J^*(x_0)$ the optimal average cost of the original MDP, and $\hat{J}(x_0)$ the optimal average cost of the approximate MDP.*

Under Assumption 1, these values are independent of the initial state; we write $\rho(\hat{\pi}) := J(x_0, \hat{\pi})$, $\hat{\rho} := \hat{J}(x_0)$, and $\rho := J^*(x_0)$ for all $x_0 \in \mathbb{X}$.

Fix δ such that $K_{\mathcal{T}} < \left(\frac{1}{\kappa}\right)^{\frac{1-\delta}{\delta}}$ where $\kappa := 1 - \nu(\mathbb{X}) > 0$, and let D denote the diameter of the state space.

i. Under Assumption 4 with $K_{\mathcal{T}} < 1$,

$$|\rho - \hat{\rho}| \leq \frac{K_c}{1 - K_{\mathcal{T}}} L_{\mathbb{X}}.$$

ii. Under Assumptions 1 and 4,

$$|\rho - \hat{\rho}| \leq \frac{K_c D^{1-\delta}}{1 - K_{\mathcal{T}}^{\delta} \kappa^{1-\delta}} \left(L_{\mathbb{X}}\right)^{\delta}$$

and furthermore,

$$\rho(\hat{\pi}) - \rho \leq \frac{2K_c D^{1-\delta}}{(1-\kappa)\left(1 - K_{\mathcal{T}}^{\delta} \kappa^{1-\delta}\right)} \left(L_{\mathbb{X}}\right)^{\delta}.$$

iii. If $L_{\mathbb{X}} = \frac{1}{n}$ (which is possible since $\mathbb{X}$ is compact), then denoting the learned policy by π_n, under Assumptions 1 and 5,

$$\lim_{n\to\infty} \rho(\pi_n) = \rho.$$

Now that we have presented the contraction framework needed for our analysis, we will move on to establishing a Q-learning algorithm and its convergence to near-optimality.

4. Quantized Q-Learning for Continuous Spaces under the Infinite Horizon Average Cost Criterion

In this section, we study Q-learning for a continuous MDP under a discretization map, and investigate whether the algorithm converges and, if so, whether its limit corresponds to the approximate MDP models constructed in the previous sections.

In particular we will present synchronous and asynchronous Q-learning algorithms for systems with continuous spaces and show the convergence of these algorithms to the value functions of appropriately defined finite MDP models consistent with those constructed in Section 3.2.

We denote by $Q^* : \mathbb{Y} \times \mathbb{U} \to \mathbb{R}$ the Q-values (factors, or functions) for the finite model introduced in Section 3.2 for some weight measure $\mu \in \mathcal{P}(\mathbb{X})$. For any discretized state $y_i \in \mathbb{Y}$ and any control action $u \in \mathbb{U}$, the Q value of the pair (y_i, u) satisfies the following equality:

$$\hat{\rho} + Q^*(y_i, u) = C(y_i, u) + \sum_{y_j \in \mathbb{Y}} P(y_j | y_i, u) \min_{v \in \mathbb{U}} Q^*(y_j, v) \tag{24}$$

where P and C are defined in (15) and $\hat{\rho}$ is the value of the average cost problem for the finite model.

Theorem 11 provides bounds on the performance loss of the policies designed for discretized models (and Theorem 15 generalizes this to the case with only weakly continuous models, though with only asymptotic convergence). Hence, if we can find iterations that converge to the Q values in (24), we will be able to obtain performance results through Theorems 11 and 15.

In the following, we present a synchronous and an asynchronous algorithm. It is important to note that the quantized process $q(X_t)$ is not an MDP, and in fact should be viewed as a POMDP, a perspective used by Kara and Yüksel (2023) (see also Kara et al., 2023).

For the asynchronous setup, we work with a single trajectory of the original process, and the data are given sequentially so that we cannot rely on the Markov properties. Accordingly, we generalize the proof method given by Kara and Yüksel (2023) for the average cost criterion and impose ergodicity properties under an exploration policy.

4.1 Synchronous Quantized Q-Learning for Continuous Space Average Cost MDPs

We first present a synchronous Q-learning algorithm. As noted above, the proof is more direct in this case. In the following, we present the synchronous Q-learning algorithm, which we will use in this section. Note that the discretization part of the algorithm follows the same steps as introduced in Section 3.2.

Recall the quantization of the state space with a collection of disjoint sets $\{B_i\}_{i=0}^{M-1}$ such that $\bigcup_i B_i = \mathbb{X}$, and $B_i \bigcap B_j = \emptyset$ for any $i \neq j$ with a representative state, $y_i \in B_i$, for each disjoint set. Denote the new finite state space by $\mathbb{Y} := \{y_0, y_1, \ldots, y_{M-1}\}$.

Algorithm 1 Synchronous Quantized Q-Learning

Input: Q_0, $\mathbb{Y}$, $\mathbb{U}$, $\{\mu_y\}_{y\in\mathbb{Y}}$, reference (y_0, u_0)
for $t = 0, \dots, L-1$ **do**
  for each $(y_i, u_j) \in \mathbb{Y} \times \mathbb{U}$ **do**
    Sample $x_i \sim \mu_{y_i}$, $X_1^{i,j} \sim \mathcal{T}(\cdot|x_i, u_j)$
    Update

$$Q_{t+1}(y_i, u_j) = (1-\alpha_t)Q_t(y_i, u_j) + \alpha_t\Big(c(x_i, u_j) + \min_v Q_t(q(X_1^{i,j}), v)\Big) \tag{25}$$

  end for
  Normalize $\hat{Q}_{t+1}(y_i, u_j) = Q_{t+1}(y_i, u_j) - Q_{t+1}(y_0, u_0)$
end for
return $\hat{Q}_L$

The algorithm updates the Q-values for all $(y, u) \in \mathbb{Y} \times \mathbb{U}$ pairs synchronously. To perform the updates for a discrete state y_i (or quantization bin B_i) and an action u_j, a *continuous* state is sampled from the bin B_i according to a pre-determined measure $x_i \sim \mu_{y_i}(\cdot)$. Along with this sampled state, the corresponding future state X_1^{ij} is also taken from the dataset. The update in (25) then uses the future continuous state X_1^{ij} and the cost $c(x_i, u_j)$ corresponding to the sampled state x_i under the action u_j. Finally, the Q-values are normalized by subtracting $Q_t(y_0, u_0)$ for a pre-selected reference pair (y_0, u_0) from all Q-values.

The next result shows that these iterations converge to the Q values given in (24) for an appropriate weight measure.

Theorem 17 *We assume that the approximate transition probabilities for the finite model satisfy for all y, y', u, u'*

$$\frac{1}{2}\sum_j \big|P(y_j|y, u) - P(y_j|y', u')\big| \leq \beta < 1. \tag{26}$$

If $\alpha_t = \frac{1}{t+1}$ and the iterations are given by (25), then Q_t converges to Q^* (see (24)) under the span semi-norm, and $\hat{Q}_t$ converges to $\hat{Q}^*$ under the uniform norm where $\hat{Q}(y, u) = Q(y, u) - Q(y_0, u_0)$ for a predetermined pair (y_0, u_0). We also have that

$$\hat{\rho} + \hat{V}^*(y_0) = C(y_0, u_0) + \sum_{y_1\in\mathbb{Y}} \hat{V}^*(y_1)P(y_1|y_0, u_0)$$

where $\hat{V}^*(y) = \min_u \hat{Q}^*(y, u)$ and where $\hat{\rho}$ is the value of the approximate model introduced in Section 3.2 under the average cost criterion. Accordingly, the results of Theorem 16 follow.

A proof is given in Appendix B. We note that under Assumption 1, (26) always holds. In particular, if $\mathcal{T}(\cdot|x, u) \geq \nu(\cdot)$, then Lemma 10 implies that $P(y_j|y, u) > 0$ for some y_j and for all y, u. This then implies the condition (26). On the other hand, there may be settings where the original problem does not satisfy the minorization condition in Assumption 1, however, one may find discretization schemes which do satisfy condition (26). Indeed, Example 3 is a case where the

original model does not satisfy Assumption 1; however, one can always find a discretization which does satisfy (26). Accordingly, for the convergence result, we impose the weaker condition, that is (26). Nonetheless, to show the near optimality of the learned policies, we still need Assumption 1.

4.2 Asynchronous Quantized Q-Learning for Continuous Space Average Cost MDPs

The Q-learning algorithm we presented earlier is constructed under the assumption that we can generate samples from every quantization bin synchronously. We now present an algorithm that learns the Q-values and an optimal policy of the finite model constructed in Section 3.2, from a single trajectory. In this setting, the main challenge is that only a single trajectory of the original continuous state MDP is available. The question is whether Q-learning applied with the quantized process converges.

The decision maker applies an arbitrary admissible policy π and collects realizations of quantized states, actions, and stage-wise cost under this policy: $Y_0, U_0, c(X_0, U_0), Y_1, U_1, c(X_1, U_1) \dots$ where Y_t denotes the representative (quantized) state corresponding to X_t.

Assumption 6 *For the finite model transition probabilities $P(y_j|y,u)$ (see (15)), where the weighting measure is given by the invariant measure of the state process under the exploration policy, there exists a bin B_j, with representative state y_j, such that*

$$P(y_j|y,u) > 0,$$

for every $y \in \mathbb{Y}$ and $u \in \mathbb{U}$. Furthermore, the index j of this bin is known.

We propose a shifted Q-learning algorithm by subtracting the value $\delta V_t(y_j)$ for some sufficiently small $\delta > 0$, where $V_t(y') = \min_u Q_t(y', u)$. This yields, for all $(y,u) \in \mathbb{Y} \times \mathbb{U}$

$$Q_{t+1}(y,u) = (1 - \alpha_t(y,u))\, Q_t(y,u) + \alpha_t(y,u) \left(c(X_t, U_t) + \min_{v \in \mathbb{U}} Q_t(Y_{t+1}, v) - \delta V_t(y_j) \right). \tag{27}$$

Algorithm 2 Asynchronous Quantized Q-Learning

Input: Q_0, quantizer $q : \mathbb{X} \to \mathbb{Y}$, exploration policy π, horizon L, parameter $\delta > 0$
Initialize $N(y,u) = 0$ for all $(y,u) \in \mathbb{Y} \times \mathbb{U}$
Set Q_0
for $t = 0, \dots, L-1$ **do**
  Observe $(X_t, U_t, c(X_t, U_t), X_{t+1})$
  $y \leftarrow q(X_t), \quad y' \leftarrow q(X_{t+1})$
  $N(y, U_t) \leftarrow N(y, U_t) + 1$
  $\alpha_t(y, U_t) \leftarrow \frac{1}{1+N(y,U_t)}$
  $Q_{t+1}(y, U_t) \leftarrow (1 - \alpha_t(y, U_t)) Q_t(y, U_t) + \alpha_t(y, U_t)(c(X_t, U_t) + \min_v Q_t(y', v) - \delta V_t(y_j))$
  $U_{t+1} \sim \pi(\cdot|y')$
end for
return Q_L

We impose the following assumptions for convergence

Assumption 7

(1.) We set $\alpha_t(y,u) = 0$ *unless* $(Y_t, U_t) = (y,u)$*. Otherwise, let*

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

2. *Under the (memoryless or stationary) exploration policy* $\pi(\cdot|\cdot)$*,* X_t *is uniquely ergodic and thus has a unique invariant measure* μ_π *such that* $\mu_\pi(B_i) > 0$ *for every quantization bin* B_i*.*

3. *The exploration policy* π *has full support over actions at every quantized state:* $\pi(u|y) > 0$ *for all* $(y,u) \in \mathbb{Y} \times \mathbb{U}$*.*

We note that a sufficient condition for the second item in Assumption 7 is that the state process $\{X_t\}_{t\geq 0}$ is positive Harris recurrent under the exploration policy. In particular, together with the minorization condition in Assumption 6, the process becomes positive Harris recurrent. Item (3) is a full support exploration condition. It is satisfied, for example, by any ϵ-greedy policy with $\epsilon > 0$ or by a uniformly randomized policy over the finite action set $\mathbb{U}$. Together, Items (2) and (3) ensure that every quantized state-action pair $(y,u) \in \mathbb{Y} \times \mathbb{U}$ is visited infinitely often almost surely, which is required both for the step-size conditions of Lemma 18 and for the empirical-average convergence in the proof of Theorem 19.

For the discounted cost criterion, an analogous result was proven by Kara and Yüksel (2023) (see also related results by Szepesvàri and Smart (2004) and Singh et al. (1994)). We first recall a key result by Bertsekas and Tsitsiklis (1996b, Prop. 4.5) and by Singh et al. (2000, Lemma 1).

Lemma 18 *(Bertsekas and Tsitsiklis, 1996b, Prop. 4.5)(Singh et al., 2000, Lemma 1) Consider a stochastic process* $(\alpha_t, \Delta_t, F_t)$*,* $t \geq 0$*, where* $\alpha_t, \Delta_t, F_t : \mathbb{S} \to \mathbb{R}$ *for some finite set* $\mathbb{S}$ *and satisfy the equations*

$$\Delta_{t+1}(s) = (1 - \alpha_t(s))\Delta_t(x) + \alpha_t(s)F_t(s)$$

Let P_t *be a sequence of increasing* σ*-fields such that* α_0 *and* Δ_0 *are* P_0 *measurable and* $\alpha_t, \Delta_t, F_{t-1}$ *are* P_t *measurable. Assume that the following hold:*

- $\sum_t \alpha_t(s) = \infty$*,* $\sum_t \alpha_t^2(s) < \infty$ *almost surely,*
- $|E[F_t(\cdot)|P_t]|_\infty \leq \beta \|\Delta_t\|_\infty + c_t$ *where* $\beta < 1$ *and* c_t *converges to zero almost surely.*
- $Var(F_t(s)|P_t) \leq K(1 + \|\Delta_t\|_\infty)^2$ *for some constant* $K < \infty$*.*

Then Δ_t *converges to zero almost surely.*

Convergence of asynchronous methods is challenging because the quantized state variable does not satisfy the Markov property and the error term is not a martingale noise. This issue does not arise in the synchronous method due to independent sampling across iterations. To deal with this, we utilize the ergodicity assumption on the state process and decompose the non-Markov noise term so that it can be written as the difference between the stationary mean and the empirical average. This difference is then folded into the decaying error term, i.e., the c_t term in Lemma 18.

Theorem 19 *Under Assumption 6 and Assumption 7, the algorithm given in (27) converges almost surely, for sufficiently small δ (specifically, $\delta < \min_{y,u\in\mathbb{Y}\times\mathbb{U}} P(y_j|y,u)$ where y_j is as in Assumption 6), to Q^* which satisfies*

$$\hat{\rho} + Q^*(y,u) = C(y,u) + \sum_{z\in\mathbb{Y}} P(z|y,u) \min_{v\in\mathbb{U}} Q^*(z,v). \tag{28}$$

P and C are defined in (15) with the weighting measure being the stationary distribution of the state process X_t under the exploration policy.

Furthermore, for the learned value function and the policy, near optimality and convergence results of Theorem 16 are applicable.

Proof The proof can be found in Appendix D. ■

5. Extensions, Reflections, and Refinements

In this section, we present some reflections and refinements. To avoid creating a distraction from the main text, these have been summarized in this final section. The first one involves arriving at near optimality via discounted cost minimization with discount parameters close to 1, the second involves the relaxation of the minorization condition at the expense of further regularity in the kernel, and the third discussion is on empirical quantized model learning and equivalence with the quantized Q-learning convergence results.

5.1 Approximate Optimality via Discounted Cost Criterion Approximation: Beyond Minorization

Consider the following two conditions: (i) There exists a solution to the average cost optimality equation, and (ii) this solution is obtained via the vanishing discount method. Under these conditions, it follows (see Cregg et al., 2024, Theorems 1 and 2) (see also Yüksel, 2025, Theorem 7.3.6) that a near-optimal policy for the discounted cost problem is also near-optimal for the average cost problem.

Accordingly, a further method would be to approximate the Q-learning algorithm by its classical discounted version. This approach, in particular, does not make explicit use of the ergodicity or minorization condition (1) (except for its use in establishing the existence of a solution to the average cost optimality equation in our paper; this condition is not necessary in general). For example, for the belief-MDP reduction of partially observable models, such a condition does not hold; yet one can establish conditions under which a solution to the average cost optimality equation exists and is obtained via the vanishing discount technique (see, e.g., Borkar (2000); Stettner (2019); Runggaldier and Stettner (1994); Platzman (1980); Hsu et al. (2006); Demirci et al. (2004)). However, the question of approximation rates as the discount parameter approaches unity remains open for such problems. We leave this important direction for future research, but only note that one can arrive at near average cost optimality via such a general method under these relatively mild conditions as well.

5.2 Near Optimality of Finite Approximations when ACOE holds but Minorization Does Not

For the difference of the value functions between the original and the approximate model, we were able to relax Assumption 1, and use Assumption 8 with $K_{\mathcal{T}} < 1$ instead. However, we could not use the same method to show near optimality of the approximate model policies. We now show that under additional regularity conditions, but not requiring minorization, we can do this for the asymptotic analysis; that is, we can show that the policy designed for the discretized model when applied to the original model is near optimal for the original model, with the performance loss decaying to zero as the quantization gets finer. Such conditions are particularly relevant for controlled (continuous-time) diffusions which are time-discretized. We make the following assumption (as in Theorem 4.1 of Yüksel (2024)):

Assumption 8 *Suppose that*

(i) We have $\mathbb{X} \subset \mathbb{R}^d$ for some finite d and $\mathbb{X}$ is compact, and $\mathbb{U}$ is finite.

(ii) For every stationary policy $\pi \in \Pi_S$ there exists a unique invariant probability measure.

(iii) The kernel $\mathcal{T}(dy|x,u)$ is such that, the family of conditional probability measures $\mathcal{T}(dy|x,u), x \in \mathbb{X}, u \in \mathbb{U}\}$ admit densities $f_{x,u}(y)$ with respect to a reference measure ψ, and all such densities are bounded and equicontinuous (over $x \in \mathbb{X}, u \in \mathbb{U}$).

(iv) $\mathcal{T}$ is weak Feller continuous.

Theorem 20 *Suppose that Assumption 4 with $K_{\mathcal{T}} < 1$ holds, and Assumption 8 holds. Then,*

$$\lim_{n\to\infty} J(x, \pi_n) = J^*(x).$$

Proof The proof can be found in Appendix E. ■

5.3 Equivalence with Empirical Model Learning

The implication of Theorem 19 is that the Q-learning algorithm (27) converges to a limit where the limit solves the optimality equation (28) for an approximate model defined by the stationary distribution of the state process X_t under the exploration policy. In particular, this model is precisely the empirical limit of a Markovian approximation in the following sense: Let the exploration policy π given in the (quantized) Q-learning algorithm give rise to the invariant probability measure μ^*_π. The limiting Q-function $Q^*(y,u)$ corresponds to the optimal Q-function of an approximate MDP defined over the quantized state space $\mathbb{Y}$. The effective cost $C^*(y,u)$ is the average cost over the bin B_y weighted by the invariant distribution μ^*_π conditioned on bin B_y:

$$C^*(y,u) = \mathbb{E}_{x\sim\mu^*_\pi | x\in B_y}[c(x,u)] = \int_{B_y} \frac{\mu^*_\pi(dx)}{\mu^*_\pi(B_y)} c(x,u). \tag{29}$$

Observe that the above is, (see e.g. Kara and Yüksel, 2024a, Theorem 2.1), equal to the almost sure limit of the empirical expression on the right hand side below:

$$C^*(y,u) = \lim_{N\to\infty} \frac{\sum_{k=0}^{N-1} c(X_k,U_k) 1_{\{X_k\in B_y, U_k = u\}}}{\sum_{k=0}^{N-1} 1_{\{X_k\in B_y, U_k=u\}}}. \tag{30}$$

Similarly, the effective transition probability $P^*(y'|y,u)$ represents the probability of transitioning from bin B_y to bin $B_{y'}$ under action u, averaged over the invariant distribution:

$$P^*(y'|y,u) = \mathbb{P}_{x\sim\mu^*_\pi|x\in B_y}[q(X_{t+1}) = y'|X_t = x, U_t = u] = \int_{B_y} \frac{\mu^*_\pi(dx)}{\mu^*_\pi(B_y)} \mathcal{T}(B_{y'}|x,u). \quad (31)$$

Likewise, by (Kara and Yüksel, 2024a, Theorem 2.1), the above is the almost sure empirical limit of the right hand side below:

$$P^*(y'|y,u) = \lim_{N\to\infty} \frac{\sum_{k=0}^{N-1} 1_{\{X_{k+1}\in B_{y'}\}} 1_{\{X_k\in B_y, U_k=u\}}}{\sum_{k=0}^{N-1} 1_{\{X_k\in B_y, U_k=u\}}} \quad (32)$$

An interpretation of the above result then is that one can first obtain the approximate model given with (29-31) by forcing the data into a Markovian model for both the empirical cost estimate (30) and empirical transition kernel estimate (32), and then solve the MDP as if this empirically constructed model is the actual one, instead of running Q-learning whose limit is then optimal precisely for this learned/empirically constructed model. Accordingly, for this learned value function and the policy, near optimality and convergence results of Propositions 6 and 8, and Theorems 11 and 15 are applicable.

6. Simulation and Case Study

We consider a continuous space control problem. We let the state space be $\mathbb{X} = [0,1]$, and the action space to be $\mathbb{U} = [-1,1]$. The stage wise cost function is given by

$$c(x,u) = 0.7(1-x) + 0.2(u+1)$$

For the dynamics we assume: if $u > 0$, for a given state x

$$x_1 \sim \begin{cases} \text{Unif}[\text{x}, \min(\text{x}+\text{u}, 1)] \text{ w.p. } 0.9 \\ \text{Unif}[0,1] \text{ w.p. } 0.1. \end{cases}$$

If $u \leq 0$

$$x_1 \sim \begin{cases} \text{Unif}[\max(\text{x}+\text{u}, 0), \text{x}] \text{ w.p. } 0.9 \\ \text{Unif}[0,1] \text{ w.p. } 0.1 \end{cases}$$

The action controls the direction and magnitude of the stochastic state transitions: positive actions try to move the state upward and negative actions tend to move it downward, with larger actions leading to more likely movements. The cost function favors states closer to the upper end of the state space and penalizes large control inputs, and thus creates a tradeoff between keeping a high quality state and limiting control effort.

For exploration, we use a randomized control policy such that $u_t \sim \text{Unif}[0,1]$ for all t. We will analyze the problem numerically after the discretization of the action space. We verify Assumption 4 and Assumption 1. Assumption 1 is satisfied with $\nu(\cdot) = 0.1 \times \text{Unif}[0,1]$. For Assumption 4, the Lipschitz constant of $c(x,u)$ is 0.7 since the cost is linear in the state variable. For the transition

kernel, the first order Wasserstein distance is equal to the L_1 distance of the CDF functions. Hence, one can check that the Lipschitz constant of the kernel is bounded by 0.9.

Figure 1 shows the convergence behavior of the relative value function for both the synchronous and asynchronous algorithms when the state and action spaces are discretized with 5 discretization bins. Recall that the relative value function that satisfies the ACOE is not unique and any shifted version of the function satisfies the ACOE. Hence, for better comparison we normalize them by subtracting the minimum value the functions when we plot them. The limit values of the asynchronous and synchronous algorithms differ slightly because the weight measure μ (see (13)) is determined differently in each case. For the asynchronous algorithm, μ is the invariant measure of the state process under the exploration policy, whereas for the synchronous algorithm, μ is a user-specified sampling distribution over each bin; for the plotted values, we used the uniform measure.

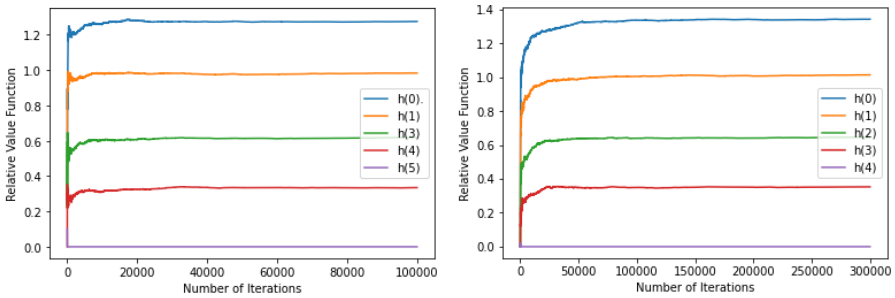

Figure 1: Relative value function convergence for synchronous(left) and asynchronous(right) algorithms

Figure 2 shows the convergence of the value constant, i.e. $\delta V_t(y_j)$, for different initial values $x_0 = 0.3, 0.5, 0.8$ for the asynchronous algorithm. We use the discretized action values $\hat{\mathbb{U}} = \{-1, 0, 1\}$, and for the state space we choose the bins to be the intervals $[0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1]$. As it can be seen from the figure, and as expected, the limit value constant does not depend on the initial state.

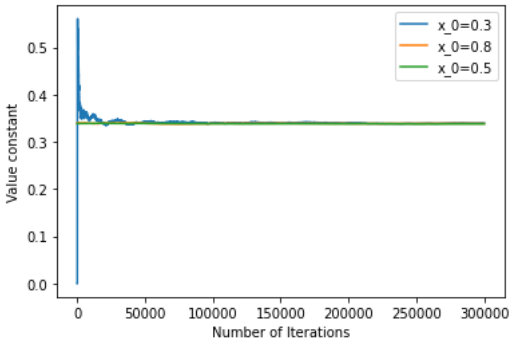

Figure 2: Algorithm convergence under different initial conditions

In Figure 3,the upper curve shows the total average cost achieved by the learned policy with $\hat{\mathbb{U}} = \{-1, 0, 1\}$ for different state-space quantization rates, while the lower curve shows the performance

of the learned policies with $\hat{\mathbb{U}} = \{-1, -0.5, 0, 0.5, 1\}$ under varying state-space quantizations. For the state space, we use uniform quantization, i.e. if the size of the discrete state space is $M = 3$, the quantization bins are $[0, \frac{1}{3}], (\frac{1}{3}, \frac{2}{3}], (\frac{2}{3}, 1]$. It is clear from the figure that as the quantization rate increases the regret decreases. Note that the asynchronous and the synchronous algorithms learn the same policy, therefore we do not plot them separately. In the same figure, we also plot the discretization error ($L_{\mathbb{X}}$), and the accumulated average cost. As our results also suggest, the cost increases linearly with the increasing discretization error.

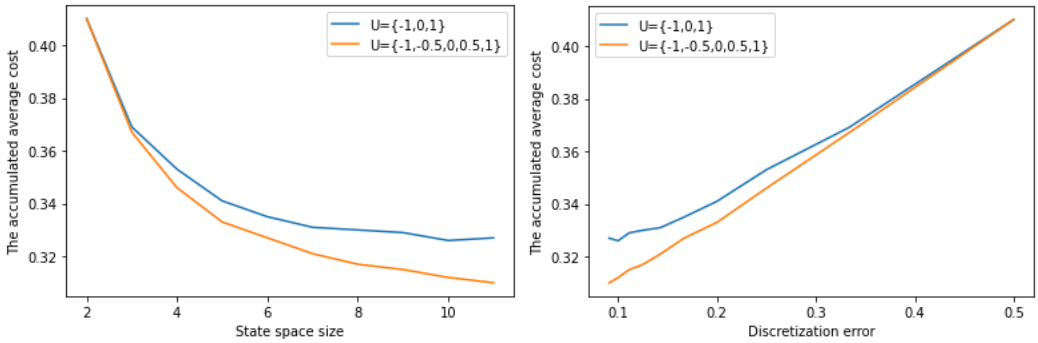

Figure 3: Learned policy performance under different quantization rates

7. Concluding Remarks

For infinite-horizon average-cost criterion problems, we presented approximation and reinforcement learning results for Markov Decision Processes with standard Borel spaces. We first provided a discretization based approximation method for fully observed Markov Decision Processes (MDPs) with continuous spaces under average cost criterion, and we derived error bounds for the approximations when the dynamics are only weakly continuous under certain ergodicity assumptions. In particular, we relaxed the total variation condition given in prior work to weak continuity as well as Wasserstein continuity conditions. We then presented synchronous and asynchronous Q-learning algorithms for continuous spaces via quantization, and established their convergence. We showed that the convergence is to the optimal Q values of the finite approximate models constructed via quantization.

A direction for future research is the following. Since we are considering an average cost problem, one can obtain an online Q-learning algorithm where the past is used to optimally adapt the exploration policy, so that the optimal cost is obtained for each sample path under mild ergodicity conditions. This can be done, e.g., by increasing the exploration lengths and adapting the resulting policy using diminishing exploration.

Acknowledgments

The authors are grateful to an anonymous referee and the Editor for their care, attention, and insightful and constructive comments.

Appendix A. Proof of Lemma 4

Proof We define iteratively

$$\begin{aligned} h_{k+1}(x) &= \inf_{u\in\mathbb{U}}\left\{c(x,u)+\int h_k(x_1)\mathcal{T}(dx_1|x,u)-\int h_k(x_1)\nu(dx_1)\right\} \\ &= \inf_{u\in\mathbb{U}}\left\{c(x,u)+\kappa\int h_k(x_1)R(dx_1|x,u)\right\} \end{aligned}$$

where $R(\cdot|x,u) := \frac{\mathcal{T}(\cdot|x,u)-\nu(\cdot)}{\kappa}$, with $h_0 \equiv 0$. Note that $h_k \to h$ uniformly.

We write for some $x, y \in \mathbb{X}$:

$$\begin{aligned} |h_{k+1}(x)-h_{k+1}(y)| &\leq \sup_u\left\{|c(x,u)-c(y,u)|+\kappa\left|\int h_k(x_1)R(dx_1|x,u)-\int h_k(x_1)R(dx_1|y,u)\right|\right\} \\ &\leq K_c d_{\mathbb{X}}(x,y)^{1-\delta}d_{\mathbb{X}}(x,y)^{\delta}+\kappa[h_k]_{H^\delta}\left[\frac{K_{\mathcal{T}}}{\kappa}d_{\mathbb{X}}(x,y)\right]^{\delta} \\ &\leq \left(K_c D^{1-\delta}+\kappa[h_k]_{H^\delta}\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}\right)d_{\mathbb{X}}(x,y)^{\delta}. \end{aligned}$$

Noting $d_{\mathbb{X}}(x,y) \leq D$, we obtain

$$[h_{k+1}]_{H^\delta} \leq K_c D^{1-\delta}\sum_{t=0}^{k}\left(\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}\right)^t \leq \frac{K_c D^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}}.$$

Together, with the fact that $h_k \to h$ uniformly and that the Hölder constant is lower semicontinuous under uniform convergence, we conclude that

$$[h]_{H^\delta} \leq \frac{K_c D^{1-\delta}}{1-\kappa\left(\frac{K_{\mathcal{T}}}{\kappa}\right)^{\delta}}.$$

■

Appendix B. Proof of Theorem 17

Proof The main part of the proof is the convergence of the algorithm. If convergence is shown, then (i) follows from Proposition 8 and Lemma 12, (ii) follows from Proposition 6 and Theorem 11 together with Lemma 12, and (iii) follows from Theorem 15. Now, let

$$F(Q)(y,u) := C(y,u)+\sum_{u,j}\min_{u_1}Q(y_1,u_1)P(y_1|y,u).$$

The process Q_t satisfies the following form, with $S_t = Q_t - Q^*$:

$$S_{t+1}(y,u) = (1-\alpha_t)S_t(y,u)+\alpha_t\Big((F(Q_t)(y,u)-F(Q^*)(y,u))+\hat{\rho}+w_t\Big),$$

where $\alpha_t = \frac{1}{t}$. Furthermore $w_t := \Big(c(x,u)+\min_v Q_t(Y_1,v)-F(Q_t)(y,u)\Big)$, where x is generated from the bin y belongs to according to the measure μ_y. Note that w_t is a zero-mean random variable. We will consider the following two parallel dynamics, as in (Jaakkola et al., 1994, Theorem 1):

$$\begin{aligned}
S^a_{t+1}(y,u) &= (1-\alpha_t(y,u))S^a_t(y,u) + \alpha_t(y,u)w_t, \\
S^b_{t+1}(y,u) &= (1-\alpha_t(y,u))S^b_t(y,u) + \alpha_t(y,u)\Big(F(Q_t)(y,u) - F(Q^*_t)(y,u) + \hat{\rho}\Big),
\end{aligned} \tag{33}$$

We have $S_t(y,u) = S^a_{t+1}(y,u) + S^b_{t+1}(y,u)$. We will show further below that $S^a_{t+1}(y,u) \to 0$ almost surely. For now we assume this holds and focus on

$$\|S^a_t + S^b_t\|_{sp} = \max_{y,u}(S^b_t(y,u) + S^a_t(y,u)) - \min_{y,u}(S^b_t(y,u) + S^a_t(y,u)).$$

Under Assumption 1 $\|(F(Q_t)(\cdot,\cdot) - F(Q^*)(\cdot,\cdot))\|_{sp} \le \beta\|S_t\|_{sp} \le \beta\|S^a_t\|_{sp} + \beta\|S^b_t\|_{sp}$, since $\|P(\cdot|y,u) - P(\cdot|y',u')\|_{TV} \le \beta < 1$. With $\|S^a_t\|_\infty \to 0$ almost surely, for every $\epsilon > 0$, there exists N such that for $t \ge N$, $\|S^a_{t+1}\|_\infty \le \frac{\epsilon}{2}$ and so that $\|S^a_{t+1}\|_{sp} \le \epsilon$ (where we suppress the sample path dependence). In the following, we assume that $t \ge N$. For some M large enough, let $\hat{\beta} := \beta\frac{(M+1)}{M} < 1$. Furthermore, for $\|S^b_t\|_{sp} > M\epsilon$, we note that for any $(y',u') \in \mathbb{Y}\times\mathbb{U}$

$$\beta\|S^b_t(y,u) - S^b_t(y',u') + \epsilon\| \le \hat{\beta}\|S^b_t\|_{sp}.$$

For $(y',u') \in \mathbb{Y}\times\mathbb{U}$, $S^b_{t+1}(y',u') = (1-\alpha_t)S^b_t(y',u') + \alpha_t\Big(F(Q_t)(y',u') - F(Q^*)(y',u') + \hat{\rho}\Big)$:

$$\begin{aligned}
|S^b_{t+1}(y,u) - S^b_{t+1}(y',u')| &\le (1-\alpha_t)|S^b_t(y,u) - S^b_t(y',u')| \\
&\quad + \left|\alpha_t\Big(F(Q_t)(y,u) - F(Q^*)(y,u) - \Big(F(Q_t)(y',u') - F(Q^*)(y',u')\Big)\right| \\
&\le (1-\alpha_t)\|S^b_t\|_{sp} + \alpha_t(\beta\|S^a_t\|_{sp} + \beta\|S^b_t\|_{sp}) \qquad (34) \\
&\le (1-\alpha_t)\|S^b_t\|_{sp} + \alpha_t\hat{\beta}\|S^b_t\|_{sp} = \Big[1-\alpha_t(1-\hat{\beta})\Big]\|S^b_t\|_{sp} \qquad (35)
\end{aligned}$$

In particular, opening the last bound iteratively, and noting that $\alpha_t = \frac{1}{t+1}$, we can write that for some $l \ge 1$ we can write $\|S^b_{t+l}\|_{sp} \le \|S^b_t\|_{sp}\prod_{k=t}^{t+l}(1-\frac{1-\hat{\beta}}{k})$. This product can be made arbitrarily small for any t by choosing l large enough. Therefore, there exists some $l < \infty$, such that $\|S^b_{t+l}\|_{sp} \le M\epsilon$. Furthermore, once $\|S^b_t\|_{sp} \le M\epsilon$, we can show via (34) and $\beta(M+1)/M < 1$ that it will remain there thereafter. Thus, for any $\epsilon > 0$, for large enough t, we have that $\|S^b_t\|_{sp} \le M\epsilon$. Since $\epsilon > 0$ is arbitrary, the convergence result follows.

We now discuss S^a_t. Taking the square of S^a_t, we obtain:

$$E[(S^a_{t+1}(y,u))^2|\mathcal{F}_t] \le (S^a_t(y,u))^2 - 2\alpha_t(S^a_t(y,u))^2 + \alpha_t^2(S^a_t(y,u))^2 + \alpha_t^2 w_t^2 \tag{36}$$

First, we have that for any $T > 0$,

$$E[\sum_{t=0}^{T-1}(2\alpha_t - \alpha_t^2)(S^a_t(y,u))^2 \quad \le \quad (S^a_0(y,u))^2 + E[\sum_{t=0}^{T-1}\alpha_t^2 w_t^2] \tag{37}$$

We now show that the right hand term is bounded. Consider:

$$Q_{t+1}(y,u) = (1-\alpha_t)Q_t(y,u) + \alpha_t(c(x,u) + \min_v Q_t(q(x_1),v))$$

which implies that

$$|Q_{t+1}(y,u)| \leq (1-\alpha_t)\|Q_t\|_\infty + \alpha_t(c(x,u) + \|Q_t\|_\infty) \leq \|Q_t\|_\infty + \alpha_t\|c\|_\infty,$$

And thus, since this holds for all (y,u) pairs, $\|Q_{t+1}\|_\infty \leq \|Q_t\|_\infty + \alpha_t\|c\|_\infty$. By bounding the partials sums of harmonic series $\sum_{k=1}^t \frac{1}{k}$, the above implies that $\|Q_{t+1}\|_\infty \leq \log(t)\|c\|_\infty + M_1$, for some finite M_1. However, $\alpha_t^2 w_t^2 \leq (1/t)^2\Big(2\|c\|_\infty^2 + 4(\log(t)\|c\|_\infty + M_1)^2\Big)$ is a summable sequence, and the right hand side of (37) is bounded. Furthermore, re-writing (36), in the expression

$$E[(S_{t+1}^a(y,u))^2|\mathcal{F}_t] \leq (S_t^a(y,u))^2 - (2\alpha_t - \alpha_t^2)(S_t^a(y,u))^2 + \alpha_t^2 w_t^2,$$

the term $\alpha_t^2 w_t^2$ is finite almost surely. This implies, by (Neveu, 1975, p. 33, Exercise II-4) (see also (Yüksel, 2025, Exercises 4.4.13 and 4.4.14) for a more explicit discussion), that S_t^a converges to some random variable almost surely. The finiteness of the right hand term in (37) then implies that the limit must be zero: Suppose not; as α_t is not summable, there exists an infinite sequence of times so that each summation of α_t between the times is bounded from below by a positive constant. Through this, if $(S_t^a)^2$ were not to converge to zero (given that it does converge to something), it would remain above a positive constant after a sufficiently large time, and then it would follow that $\sum_t (2\alpha_t - \alpha_t^2)S_t^{a2}$ would not remain bounded. Therefore, if this were to happen with non-zero measure, the expectation would be unbounded, which in turn would, as $T \to \infty$, violate (37).

Finally, if the sequence converges under the span semi-norm, for some constant $\hat{\rho}$.

$$\hat{\rho} + Q^*(y,u) = F(Q^*)(y,u) = C(y,u) + \sum_{y'} P(y'|y,u)\min_v Q^*(t',v)$$

Note that the minimum of u, for each y, is the solution to the Average Cost Optimality Equation. Hence, the stationary policy $\{f^*\}$ is optimal. Furthermore, $\hat{Q}^*$ is only a constant shifted version of Q^*, $\hat{Q}^*$ also satisfies the ACOE. By definition, we have $\hat{Q}^*(y_0,u_0) = 0$. Thus, we have that

$$\hat{\rho} + \hat{V}^*(y_0) = C(y_0,u_0) + \sum_{y_1 \in \mathbb{Y}} \hat{V}^*(y_1)P(y_1|y_0,u_0).$$

■

Appendix C. On Lemma 18

We provide a short proof for Lemma 18 as applied to our theorem, essentially building on Bertsekas and Tsitsiklis (1996b). Write

$$\Delta_{t+1}(x) = (1-\alpha_t(x))\Delta_t(x) + \alpha_t(x)F_t(x)$$

as

$$\Delta_{t+1}(x) = (1-\alpha_t(x))\Delta_t(x) + \alpha_t(x)\Big(E[F_t(x)|P_t] + F_t(x) - E[F_t(x)|P_t]\Big)$$

We will take B_t so that $\|\Delta_t\|_\infty \leq B_t$ with $\hat{w}_t := \frac{F_t(x)-E[F_t(x)|P_t]}{B_t}$ so that $E[\hat{w}_t^2] \leq K$ by assumption. To this end, we take $B_0 = 1 + \|\Delta_0\|_\infty$ and for $t \in \mathbb{Z}_+$,

$$B_{t+1} = \max(B_t, 1 + \|\Delta_{t+1}\|_\infty)$$

Let

$$R_t(x) = \frac{\Delta_t(x)}{B_t}.$$

Write

$$\Delta_{t+1}(x) \leq (1-\alpha_t(x))B_t R_t(x) + \alpha_t(x)\Big(E[F_t(x)|P_t] + B_t\hat{w}_t\Big)$$

and

$$\Delta_{t+1}(x) \leq B_t\Big((1-\alpha_t(x))R_t(x) + \alpha_t(x)\Big(\beta R_t(x) + \hat{w}_t\Big)\Big) \tag{38}$$

Therefore,

$$R_{t+1}(x)B_{t+1} \leq B_t\Big((1-\alpha_t(x))R_t(x) + \alpha_t(x)\Big(\beta R_t(x) + \hat{w}_t\Big)\Big)$$

and since $\frac{B_t}{B_{t+1}} \leq 1$,

$$R_{t+1}(x) \leq \Big((1-\alpha_t(x))R_t(x) + \alpha_t(x)\Big(\beta R_t(x) + \hat{w}_t\Big)\Big)$$

We know that $(1-\alpha_t(x))R_t(x) + \alpha_t(x)\Big(\beta R_t(x) + \hat{w}_t\Big) \to 0$ almost surely (Tsitsiklis (1994) or (Yüksel, 2025, Theorem 9.1.1)) as $\hat{w}_t$ has a uniformly bounded variance. Accordingly, for large enough t,

$$R_{t+1}(x) \leq \epsilon,$$

and by (38)

$$B_{t+1} = \max(B_t, 1 + B_t\epsilon_t)$$

This implies then that B_t is bounded (almost surely; though not necessarily by a constant over all realizations), and so is Δ_t. Once we have that Δ_t is bounded, we can write

$$S^b_{t+1}(x) = (1-\alpha_t(x))S^b_t(x) + \alpha_t(x)\Big(E[F_t(x)|P_t]\Big)$$
$$S^a_{t+1}(x) = (1-\alpha_t(x))S^a_t(x) + \alpha_t(x)\Big(F_t(x) - E[F_t(x)|P_t]\Big)$$

Now, we have that in S^a_t above, the term $\Big(F_t(x) - E[F_t(x)|P_t]\Big)$ has a conditionally bounded covariance, even though there is no uniform bound. Nonetheless, (Bertsekas and Tsitsiklis, 1996b, Corollary 4.1) implies that $S^a_t \to 0$ in this case as well. With $S^a_t \to 0$, $S_b(t) \to 0$ also via Tsitsiklis (1994) or (Yüksel, 2025, Theorem 9.1.1).

Appendix D. Proof of Theorem 19

Proof For small enough δ, we define a positive measure ν' such that

$$\nu'(y_j) = \delta,$$

and $\nu'(y) = 0$ otherwise where y_j is as in Assumption 6. ν' then satisfies

$$P(\cdot|y,u) \geq \nu'(\cdot)$$

since we selected $\delta < \min_{y,u} P(y_j|y,u)$. We now define the following new transition measure (though, not a conditional probability measure)

$$\hat{P}(\cdot|y,u) = P(\cdot|y,u) - \nu'(\cdot).$$

Then, as noted earlier, the following operator is a contraction with $1-\delta$ on bounded measurable functions $\mathcal{B}(\mathbb{Y}\times\mathbb{U})$ (see (8) and (Hernández-Lerma, 2012, p.61))

$$(\hat{\mathbb{T}}f)(y,u) = C(y,u) + \sum_{y'} \min_u f(y',u)\hat{P}(y'|y,u).$$

We define the fixed point, say $Q^*(y,u)$, of the mapping $\hat{\mathbb{T}}$, such that

$$\begin{aligned} Q^*(y,u) = (\hat{\mathbb{T}}Q^*)(y,u) &= C(y,u) + \sum_{y'} \min_u Q^*(y',u)\hat{P}(y'|y,u) \\ &= C(y,u) + \sum_{y'} \min_u Q^*(y',u)P(y'|y,u) - \sum_{y'} \min_u Q^*(y',u)\nu'(y') \\ &= C(y,u) + \sum_{y'} \min_u Q^*(y',u)P(y'|y,u) - \delta \min_u Q^*(y_j,u) \end{aligned}$$

which satisfies an ACOE with $\delta \min_u Q^*(y_j,u) = \hat{\rho}$.

We claim that the iterations converge to Q^*. Now, let $\Delta(y,u) := Q_t(y,u) - Q^*(y,u)$, we then write

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

where

$$\begin{aligned} F_t(y,u) = &(c(X_t,U_t) + V_t(Y_{t+1}) - \delta V_t(y_j)) \\ &- \left(C(y,u) + \sum_{y'} V^*(y')P(y'|y,u) - \delta V^*(y_j) \right) \end{aligned}$$

where $Y_{t+1} = q(X_{t+1})$. We now write $F_t = \hat{F}_t + r_t$ by adding and subtracting $V^*(Y_{t+1})$ with

$$\begin{aligned} \hat{F}_t(y,u) &= V_t(Y_{t+1}) - \delta V_t(y_j) - (V^*(Y_{t+1}) - \delta V^*(y_j)) \\ r_t(y,u) &= c(X_t,U_t) - C(y,u) + V^*(Y_{t+1}) - \sum_{y'} V^*(y')P(y'|y,u). \end{aligned} \tag{39}$$

We define two processes $\delta_t(y,u)$ and $v_t(y,u)$:

$$\delta_{t+1}(y,u) = (1-\alpha_t(y,u))\delta_t(y,u) + \alpha_t(y,u)\hat{F}_t(y,u)$$
$$v_{t+1}(y,u) = (1-\alpha_t(y,u))v_t(y,u) + \alpha_t(y,u)r_t(y,u).$$

We first show that $v_t(y,u) \to 0$ for all (y,u). Due to how we have chosen the learning rates, one can show that

$$v_{t+1}(y,u) = \frac{\sum_{k=0}^{t-1} r_k(y,u)\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}{\sum_{k=0}^{t-1}\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}$$

where $Y_k = q(X_k)$. By Items (2) and (3) of Assumption 7, the stationary probability of the event $\{(Y_t,U_t)=(y,u)\}$ equals $\mu_\pi(B_y)\cdot\pi(u|y) > 0$ for every $(y,u) \in \mathbb{Y}\times\mathbb{U}$. Hence every pair (y,u) is visited infinitely often almost surely, so the denominator above diverges to infinity as $t\to\infty$ and the ratio is well-defined for all sufficiently large t. This property also ensures that the step size conditions $\sum_t \alpha_t(y,u) = \infty$ and $\sum_t \alpha_t^2(y,u) < \infty$ of Lemma 18 hold almost surely for every (y,u). Note that the joint process $\{X_t, Y_t, U_t\}$ is Markovian and has a unique stationary measure under Assumption 7 given by

$$Pr(X_t \in A, Y_t = y, U_t = u) = \int_A \pi(u|y)\mathbb{1}_{\{q(X_t)=y\}}\mu_\pi(dx)$$

for any $A \in \mathcal{B}(\mathbb{X})$, and for any $y,u \in \mathbb{Y}\times\mathbb{U}$ where μ_π is the stationary distribution of the hidden state process under the exploration policy π.

Thus, we first write assuming $y \in B_i$

$$\frac{\sum_{k=0}^{t-1} c(X_k,U_k)\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}{\sum_{k=0}^{t-1}\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}} = \frac{\frac{1}{t}\sum_{k=0}^{t-1} c(X_k,U_k)\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}{\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}$$
$$\to \frac{\int_{B_i} c(x,u)\pi(u|y_i)\mu_\pi(dx)}{\pi(u|y_i)\mu_\pi(B_i)} = \frac{\int_{B_i} c(x,u)\mu_\pi(dx)}{\mu_\pi(B_i)} = C(y,u).$$

Using identical arguments, we can also show that

$$\frac{\sum_{k=0}^{t-1} V^*(Y_{t+1})\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}}{\sum_{k=0}^{t-1}\mathbb{1}_{\{(Y_k,U_k)=(y,u)\}}} \to \sum_{y'} V^*(y')P(y'|y,u).$$

Hence, we have that $v_t(y,u)\to 0$ almost surely for all $(y,u)\in\mathbb{Y}\times\mathbb{U}$. To show that δ_t converges to zero, we will use Lemma 18. Under Assumption 6

$$E[\hat{F}_t(y_i,u_i)|h_t] \le \beta\|V_t - V^*\|_\infty \le \beta\|\Delta_t\|_\infty \le \beta\|\delta_t\| + \beta\|v_t\|_\infty$$

where $\beta := (1-\delta) < 1$, and $\|v_t\|_\infty$ converges to zero almost surely. We finally need to verify that

$$Var(\hat{F}_t|h_t) \le K(1+\|\delta_t\|_\infty)^2$$

where $h_t = \{y_t,u_t,\dots,y_0,u_0\}$. We write

$$Var(\hat{F}_t|h_t) = E\Bigg[\Big(V_t(Y_{t+1}) - \delta V_t(y_j) - V^*(Y_{t+1}) + \delta V^*(y_j)$$

$$
\begin{aligned}
&- \int V_t(y')P(y'|h_t) + \delta V_t(y_j) + \int V^*(y')P(y'|h_t) - \delta V^*(y_j)\Big)^2\Big] \\
&= E\Big[\Big(V_t(Y_{t+1}) - V^*(Y_{t+1}) - \int V_t(y')P(y'|h_t) + \int V^*(y')P(y'|h_t)\Big)^2\Big] \\
&\leq \|V_t - V^*\|_\infty^2 \leq \|\delta_t + v_t\|_\infty^2.
\end{aligned}
$$

Note that $\|v_t\|_\infty$ remains bounded uniformly over t (and over all sample paths as V^*, c are bounded in (39)), since the cost function $c(x,u)$ and $V^*(y)$ are uniformly bounded and thus $r_t(y,u)$ is bounded. Hence, we can find some $\hat{K}$ such that

$$Var(\hat{F}_t|h_t) \leq \hat{K}(1 + \|\delta_t\|_\infty)^2.$$

Thus, we can conclude that $\|\delta_t\|_\infty \to 0$ almost surely.

■

Appendix E. Proof of Theorem 20

Proof To make the analysis explicit let $\mathcal{T}_n$ be the discretized finite state/action model given by (15). Notably, let π_n be optimal for $\mathcal{T}_n$. We then seek to bound:

$$J(\mathcal{T}, \pi_n) - J(\mathcal{T}_n, \pi_n) + J(\mathcal{T}_n, \pi_n) - J(\mathcal{T}, \pi)$$

The second term above goes to zero by Proposition 8. We consider then the first term

$$J(\mathcal{T}, \pi_n) - J(\mathcal{T}_n, \pi_n)$$

Instead of working with the ACOE, we will find it convenient to work with the induced invariant probability measures. Let the invariant measures be: for $\mathcal{T}$ under π_n be ν^n (with $\nu^n \ll \psi$); and for $\mathcal{T}_n$ under π_n be $\hat{\nu}^n$, so that

$$\nu^n(dx') = \int \nu^n(dx) \int \mathcal{T}(dx'|x,u)\pi_n(du|x) \tag{40}$$

$$\hat{\nu}^n(dx') = \int \hat{\nu}^n(dx) \int \mathcal{T}_n(dx'|x,u)\pi_n(du|x) \tag{41}$$

where (41) is constructed via the invariance equation

$$\hat{\nu}^n(B^n_{y'}) = \sum_y \hat{\nu}^n(B^n_y) \int \hat{\nu}^n(dx|y) \int (\mathcal{T}(B^n_{y'}|s,u)\mu^n_y(ds))\pi_n(du|x). \tag{42}$$

Here, μ^n_y is a pre-defined weighting measure and $\hat{\nu}^n(dx|y)$ is an artificial normalization as it does not impact the integration on the right hand side noting that $\pi_n(du|x)$ is a constant policy for all $x \in B^n_y$. Note that by the construction above as $\mathcal{T}(\cdot|s,u) \ll \psi(\cdot)$, we have that we can extend $\hat{\nu}^n(B^n_{y'})$ to the entire $\mathbb{X}$ so that $\hat{\nu}^n(dx) := \sum_y \hat{\nu}^n(B^n_y) \int \hat{\nu}^n(dx|y) \ll \psi(dx)$.

The question is, does

$$\int \nu^n(dx)c(x,\pi_n(x)) - \int \hat{\nu}^n(dx)c(x,\pi_n(x)) \to 0$$

as $n \to \infty$? Consider (40) and take $n \to \infty$. We have that ν^n to some $\bar{\nu}$ along a subsequence weakly by compactness, where this convergence is in total variation by equi-continuity of densities with respect to ψ. Furthermore, along a further subsequence, $\pi_n \to \pi^*$ for some π^* in Young topology at input ψ which then implies that this is also true at input $\bar{\nu}$ by (Yüksel, 2024, Lemma 3.6). Since convergence of $\nu^n(dx) \to \bar{\nu}(dx)$ is also in total variation (Yüksel, 2024, Theorem 4.1), this then implies that $\nu^n(dx)\pi_n(du|x)$ converges weakly to some $\bar{\nu}(dx)\pi^*(du|x)$.

For the second term (42), we note that $\hat{\nu}^n(dx)$ is defined first by its discrete support on the bins B_y^n via the weighting measures μ_y, but this can be extended to the entire $\mathbb{X}$ so that $\hat{\mu}^n \ll \psi$ as noted above. Therefore, by the equi-continuity condition, this sequence of measures will have a converging subsequence which does so in total variation (Yüksel, 2024, Theorem 4.1) to some measure $\tilde{\nu} \ll \psi$ in total variation. This then implies that the measure $\hat{\nu}^n(dx)\pi_n(du|x)$ converges weakly, along a subsequence, to some $\tilde{\nu}(dx)\pi^*(du|x)$,

Since $\mathcal{T}$ is weakly continuous, it follows that for (42), by (Serfozo, 1982, Theorem 3.5) or (Langen, 1981, Theorem 3.5), the limit leads to invariance and therefore the limits $\bar{\nu}$ and $\tilde{\nu}$ have to be invariant under the same policy $\pi^*(du|x)$. However, since the invariant measure is to be unique given the policy π^* by hypothesis, it must be that these limit measures are identical. Therefore, the induced costs

$$\int \nu^n(dx)\pi_n(du|x)c(x,u) - \int \hat{\nu}^n(dx)\pi_n(du|x)c(x,u) \to 0.$$

It should be noted that π_n converges to a limit π^* under the Young topology (and not pointwise in x); this is why the analysis above is needed. One could also note that the analysis above can be generalized to the case where $\mathbb{X} = \mathbb{R}^d$, however with the required tightness conditions on the set of invariant measures as in (Yüksel, 2024, Theorem 4.1) and the associated ACOE conditions.

■

Appendix F. Alternative Ergodicity Conditions

Proposition 21 *(Hernández-Lerma et al., 1991, Theorem 3.2) Consider the following.*

a. *There exists a state $x^* \in \mathbb{X}$ and a number $\beta > 0$ such that $\mathcal{T}(\{x^*\}|x,\pi) \geq \beta$, for all $x \in \mathbb{X}, \pi \in \Pi_s$.*

b. *There exists a positive integer t and a non-trivial measure ν on $\mathbb{X}$ such that $\mathcal{T}^t(\cdot|x,\pi) \geq \nu(\cdot)$ for all $x \in \mathbb{X}, \pi \in \Pi_s$.*

c. *For each $\pi \in \Pi_s$, the transition kernel $\mathcal{T}(dy|x,\pi)$ has a density $p(y|x,\pi)$ with respect to a sigma-finite measure m on $\mathbb{X}$, and there exist $\epsilon > 0$ and $C \in \mathcal{B}(\mathbb{X})$ such that $m(C) > 0$ and $p(y|x,\pi) \geq \epsilon$ for all $y \in C, x \in \mathbb{X}, \pi \in \Pi_s$.*

d. *For each $\pi \in \Pi_s$, $\mathcal{T}(dy|x,\pi)$ has a density $p(y|x,\pi)$ with respect to a sigma-finite measure m on $\mathbb{X}$, and $p(y|x,\pi) \geq p_0(y)$ for all $x, y \in \mathbb{X}$, $\pi \in \Pi_s$, where p_0 is a non-negative measurable function with $\int p_0(y)m(dy) > 0$.*

e. There exists a positive integer t and a measure ν on $\mathbb{X}$ such that $\nu(\mathbb{X}) < 2$ and $\mathcal{T}^t(\cdot|x,\pi) \leq \nu(\cdot)$ for all $x \in \mathbb{X}, \pi \in \Pi_s$.

f. There exists a positive integer t and a positive number $\beta < 1$ such that $\|\mathcal{T}^t(\cdot|x,\pi) - \mathcal{T}^t(\cdot|x',\pi)\|_{TV} \leq 2\beta$ for all $x, x' \in \mathbb{X}, \pi \in \Pi_s$.

g. There exists a positive integer t and a positive number β for which the following holds: For each $\pi \in \Pi_s$, there is a probability measure ν_π on $\mathbb{X}$ such that $\mathcal{T}^t(\cdot|x,\pi) \geq \beta\nu_\pi(\cdot)$ for all $x \in \mathbb{X}$.

h. There exist positive numbers c and β, with $\beta < 1$, for which the following holds: For each $\pi \in \Pi_s$, there is a probability measure p_π on $\mathbb{X}$ such that $\|\mathcal{T}^t(\cdot|x,\pi) - p_\pi(\cdot)\|_{TV} \leq c\beta^t$ for all $x \in \mathbb{X}, t \in \mathbb{N}$.

i. The state process is uniformly ergodic such that $\lim_{t\to\infty} \|\mathcal{T}^t(\cdot|x,\pi) - p_\pi(\cdot)\|_{TV} = 0$ uniformly in $x \in \mathbb{X}$ and $\pi \in \Pi_s$.

The conditions above are related as follows:

$$a \to b$$
$$e \to f$$
$$c \to d \to b \to f \leftrightarrow g \leftrightarrow h \leftrightarrow i.$$

References

J. Abounadi, D. Bertsekas, and V. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

D. J. Aldous. *Weak convergence and the general theory of processes*. Editeur inconnu, 1981.

A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control and Optimization*, 31:282–344, 1993.

K. Asadi, D. Misra, and M. Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 264–273. PMLR, 2018.

J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. All adapted topologies are equal. *Probability Theory and Related Fields*, 178(3):1125–1172, 2020a.

J. Backhoff-Veraguas, D. Bartl, M. Beiglböck, and M. Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3):601–632, 2020b.

D. Bartl and J. Wiesel. Sensitivity of multiperiod optimization problems with respect to the adapted wasserstein distance. *SIAM Journal on Financial Mathematics*, 14(2):704–720, 2023.

D. Bartl, M. Beiglböck, and G. Pammer. The wasserstein space of stochastic processes. *Journal of the European Mathematical Society*, 2024.

E. Bayraktar, Y. Y. Dolinsky, and J. Guo. Continuity of utility maximization under weak convergence. *Mathematics and Financial Economics*, pages 1–33, 2020.

M. Beiglböck, B. Jourdain, W. Margheriti, and G. Pammer. Approximation of martingale couplings on the line in the adapted weak topology. *Probability Theory and Related Fields*, 183(1):359–413, 2022.

D. Bertsekas. Convergence of discretization procedures in dynamic programming. *IEEE Trans. Autom. Control*, 20(3):415–419, Jun. 1975.

D. Bertsekas and J. Tsitsiklis. Neuro-dynamic programming. Athena Scientific, 1996a.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996b.

P. Billingsley. *Probability and Measure*. Wiley, 3rd edition, 1995.

V. Borkar and S. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

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

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.

M. Bravo and R. Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

W. Chae, K. Hong, Y. Zhang, A. Tewari, and D. Lee. Learning infinite-horizon average-reward linear mixture MDPs of bounded span. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 2737–2745. PMLR, 2025.

Z. Chen. Non-asymptotic guarantees for average-reward q-learning with adaptive stepsizes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=kGlrPZuHPq`.

C. Chow and J. N. Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE transactions on automatic control*, 36(8):898–914, 1991.

O. Costa and F. Dufour. Average control of Markov decision processes with Feller transition probabilities and general action spaces. *Journal of Mathematical Analysis and Applications*, 396(1): 58–69, 2012.

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.

Y. Demirci, A. 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*, 62:2859–2883, 2004.

F. Dufour and T. Prieto-Rumeau. Approximation of Markov decision processes with general state space. *J. Math. Anal. Appl.*, 388:1254–1267, 2012.

E. Feinberg, P. Kasyanov, and N. Zadioanchuk. Average cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.*, 37(4):591–607, Nov. 2012.

C. Gaskett and A. Z. D. Wettergreen. Q-learning in continuous state and action spaces. In *Australasian joint conference on artificial intelligence*, pages 417–428. Springer, 1999.

E. Gordienko and O. Hernández-Lerma. Average cost Markov control processes with weighted norms: Existence of canonical policies. *Appl. Math.*, 23(2):199–218, 1995.

A. Gosavi. Reinforcement learning for long-run average cost. *European journal of operational research*, 155(3):654–674, 2004.

M. F. Hellwig. Sequential decisions under uncertainty and the maximum theorem. *Journal of Mathematical Economics*, 25(4):443–464, 1996.

O. Hernández-Lerma. *Adaptive Markov control processes*, volume 79. Springer Science & Business Media, 2012.

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

O. Hernández-Lerma and J. B. Lasserre. *Further topics on discrete-time Markov control processes*. Springer, 1999.

O. Hernández-Lerma, R. M. de Oca, and R. Cavazos-Cadena. Recurrence conditions for markov decision processes with borel state space: a survey. *Annals of Operations Research*, 28(1):29–46, 1991.

C. J. Himmelberg, T. Parthasarathy, and F. S. V. Vleck. Optimal plans for dynamic programming problems. *Mathematics of Operations Research*, 1(4):390–394, 1976.

K. Hong and A. Tewari. A computationally efficient algorithm for infinite-horizon average-reward linear MDPs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 23751–23773. PMLR, 2025.

K. Hong, W. Chae, Y. Zhang, D. Lee, and A. Tewari. Reinforcement learning for infinite-horizon average-reward linear MDPs via approximation by discounted-reward MDPs. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 2989–2997. PMLR, 2025.

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.

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

N. Jiang. Notes on state abstractions. https://nanjiang.cs.illinois.edu/files/cs542f22/note4.pdf, 2018. Unpublished lecture notes, CS 542, University of Illinois Urbana–Champaign.

Y. Jin, R. Gummadi, Z. Zhou, and J. Blanchet. Feasible q-learning for average reward reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR, 2024.

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

A. 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.

A. Kara and S. Yüksel. Q-learning for stochastic control under general information structures and non-markovian environments. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=1Yp6xpTV55. Featured Certification.

A. Kara and S. Yüksel. Partially observed optimal stochastic control: Regularity, optimality, approximations, and learning. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 6709–6721, 2024b.

A. 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.

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.

V. R. Konda and J. N. Tsitsiklis. Onactor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003. doi: 10.1137/S0363012901385691. URL https://doi.org/10.1137/S0363012901385691.

K. Kuratowski and C. Ryll-Nardzewski. A general theorem on selectors. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys*, 13(1):397–403, 1965.

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

D. Maran, A. M. Metelli, and M. Restelli. Tight performance guarantees of imitator policies with continuous actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9073–9080, 2023.

F. C. Melo, S. P. Meyn, and I. M. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

S. Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

J. Neveu. Discrete-parameter martingales. revised edition, 1975.

D. Ormoneit and Ś. Sen. Kernel-based reinforcement learning. *Machine learning*, 49(2):161–178, 2002.

G. Pammer. A note on the adapted weak topology in discrete time. *Electronic Communications in Probability*, 29:1–13, 2024.

M. Pirotta, M. Restelli, and L. Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100:255–283, 2015.

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

S. Pradhan and S. Yüksel. Continuity of cost in Borkar control topology and implications on discrete space and time approximations for controlled diffusions under several criteria. *Electronic Journal of Probability*, 29:1–32, 2024.

E. Rachelson and M. G. Lagoudakis. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics*, 2010. URL `https://api.semanticscholar.org/CorpusID:14029770`.

W. J. Runggaldier and L. Stettner. *Approximations of discrete time partially observed control problems*. Giardini Pisa, 1994.

N. Saldi. Finite-state approximations to discounted and average cost constrained markov decision processes. *IEEE Transactions on Automatic Control*, 64(7):2681–2696, 2019. doi: 10.1109/TAC.2018.2890756.

N. Saldi and S. Yüksel. Kernel mean embedding topology: Weak and strong forms for stochastic kernels and implications for model learning. *arXiv preprint arXiv:2502.13486*, 2025.

N. Saldi, T. Linder, and S. Yüksel. Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control. *IEEE Trans. Automatic Control*, 60:553 –558, 2015a.

N. Saldi, S. Yüksel, and T. Linder. Finite-state approximation of Markov decision processes with unbounded costs and Borel spaces. In *IEEE Conf. Decision Control*, Osaka. Japan, December 2015b.

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.

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.

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

M. Schäl. A selection theorem for optimization problems. *Archiv der Mathematik*, 25(1):219–224, 1974.

M. Schäl. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitsth*, 32:179–296, 1975.

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

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

S. Singh, T. Jaakkola, M. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38:287–308, 2000.

S. P. Singh, T. Jaakkola, and M. I. Jordan. Reinforcement learning with soft state aggregation. *Advances in neural information processing systems*, pages 361–368, 1995.

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

W. A. Suttle, A. Bedi, B. Patel, B. M. Sadler, A. Koppel, and D. Manocha. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level Monte Carlo actor-critic. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33240–33267. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/suttle23a.html`.

C. Szepesvári. Algorithms for reinforcement learning. volume 4, pages 1–103, 2010.

C. Szepesvàri and W. Smart. Interpolation-based q-learning. 2004.

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

J. N. Tsitsiklis and B. V. Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

S. Vakili and J. Olkhovskaya. Kernel-based function approximation for average reward reinforcement learning: An optimist no-regret algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2410.23498.

O. Vega-Amaya. The average cost optimality equation: a fixed point approach. *Bol. Soc. Mat. Mexicana*, 9(3):185–195, 2003a.

O. Vega-Amaya. The average cost optimality equation: a fixed point approach. *Bol. Soc. Mat. Mexicana*, 9(1):185–195, 2003b.

C. Villani. *Optimal transport: old and new*. Springer, 2009.

Y. Wan, A. Naik, and R. S. Sutton. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

Y. Wang, A. Velasquez, G. K. Atia, A. Prater-Bennette, and S. Zou. Model-free robust average-reward reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36431–36469. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wang23am.html.

C.-Y. Wei, M. Jafarnia-Jahromi, R. Jain, and T. Jaksch. Learning infinite-horizon average-reward MDPs with linear function approximation. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 3007–3015. PMLR, 2021. arXiv:2007.11849.

Y. Wu, D. Zhou, and Q. Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward MDPs with linear function approximation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pages 3883–3913. PMLR, 2022.

Z. Yang, Y. Chen, M. Hong, and Z. Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in neural information processing systems*, 32, 2019.

S. Yüksel. On Borkar and Young relaxed control topologies and continuous dependence of invariant measures on control policy. *SIAM Journal on Control and Optimization*, 62(4):2367–2386, 2024.

S. Yüksel. *Optimization and Control of Stochastic Systems*. Queen's University, Lecture Notes, available online, 2025.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940, 2021a.

S. Zhang, Z. Zhang, and S. T. Maguluri. Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021b.

Y. Zhang and K. W. Ross. On-policy deep reinforcement learning for the average-reward criterion. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12535–12545. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhang21q.html.