Near Optimal Approximations and Finite Memory Policies for POMPDs with Continuous Spaces*

Dedicated to Professor Peter E. Caines, on the occasion of his 80th birthday.

KARA Ali Devran · BAYRAKTAR Erhan · YÜKSEL Serdar

Abstract   The authors study an approximation method for partially observed Markov decision processes (POMDPs) with continuous spaces. Belief MDP reduction, which has been the standard approach to study POMDPs requires rigorous approximation methods for practical applications, due to the state space being lifted to the space of probability measures. Generalizing recent work, in this paper the authors present rigorous approximation methods via discretizing the observation space and constructing a fully observed finite MDP model using a finite length history of the discrete observations and control actions. The authors show that the resulting policy is near-optimal under some regularity assumptions on the channel, and under certain controlled filter stability requirements for the hidden state process. The authors also provide a Q learning algorithm that uses a finite memory of discretized information variables, and prove its convergence to the optimality equation of the finite fully observed MDP constructed using the approximation method.

Keywords   Filter stability, POMDP, reinforcement learning, stochastic control.

1   Introduction

Partially Observed Markov Decision Problems (POMDPs) offer a practically rich and relevant, and mathematically challenging model. Analysis of optimality and computation of solutions become complex even for finite models. Hence, approximations are required for efficient (near) optimality and control analysis. We present a brief review in the following.

KARA Ali Devran (Corresponding author)

Department of Mathematics, Florida State University, FL 32306-2400, USA. Email: akara@fsu.edu.

BAYRAKTAR Erhan

Department of Mathematics, University of Michigan, MI 48109, USA. Email: erhan@umich.edu.

YÜKSEL Serdar

Department of Mathematics and Statistics, Queen's University, ON K7L 3N6, Canada.

Email: yuksel@queensu.ca.

 Springer

Most of the studies in the literature are algorithmic and computational contributions such as [1, 2]. These studies develop computational algorithms, utilizing structural convexity/concavity properties of the value function under the discounted cost criterion. [3] provides an insightful algorithm which may be regarded as a quantization of the belief space; however, no rigorous convergence results are provided. [4, 5] also present quantization based algorithms for the belief state, where the state, measurement, and the action sets are finite.

Rigorous approximation techniques for POMDPs have mostly focused on finite state or finite observation models. Some works that study POMDPs with continuous state or observation spaces include [6–9].

[6, 7] introduce a rigorous approximation analysis after establishing weak continuity conditions on the transition kernel defining the (belief-MDP) via the non-linear filter (see [10, 11]), and it is shown that the finite model approximations obtained through quantization are asymptotically optimal as the number of quantization bins increases. The finite model is constructed by choosing a finite number of belief points which are sufficiently close to any point in the original belief space, however, this may not be an easy task in general, as the belief space consists of probability measures.

Another rigorous set of studies with continuous observation and state spaces is [8, 9] where the authors provide an explicit quantization method for the set of probability measures containing the belief states, where the state space is parametrically representable under strong density regularity conditions. The quantization is done through the approximations as measured by Kullback-Leibler divergence (relative entropy) between probability density functions.

We refer the reader to the survey papers by [12–14] and [15] for further structural results as well as algorithmic and computational methods for approximating POMDPs. Notably, for POMDPs[15] presents structural results on optimal policies under monotonicity conditions of the value function in the belief variable.

[16] studies near optimality of finite window policies for average cost problems where the state, action and observation spaces are finite; under the condition that the liminf and limsup of the average cost are equal and independent of the initial state, the paper establishes the near-optimality of (non-stationary) finite memory policies. In another related direction, [17] studies finite memory approximation techniques for POMDPs with finite state, action and measurements. The POMDP is reduced to a belief MDP and the worst and best case predictors prior to the N most recent information variables are considered to build an approximate belief MDP. The original value function is bounded using these approximate belief MDPs that use only finite memory, where the finiteness of the state space is critically used.

Our paper generalizes several recent studies [18–21] and the review paper [22]. In these studies, finite memory approximations as well as quantized belief approximations are studied, and near optimality of the approximations and their associated Q learning algorithms are provided under certain controlled filter stability assumptions, both for discounted and average cost criteria. However, the methods provided in these papers are tailored towards finite observation spaces, and the efficiency of the algorithms can be affected significantly for large observation spaces. The current paper addresses this gap.

🖄 Springer

We dedicate this paper to Professor Peter E. Caines, who has shaped stochastic control theory fundamentally in a variety of directions. In the context of this paper on control with partial information, the contributions of Professor Caines in a variety of contexts include: Filtering theory and associated stability analysis[23], partially observable jump parameter systems[24], control with Poisson measurements[25], mean-field games with noisy measurements[26], observer design for automata[27], and filtering in Riemannian manifolds[28]. A particular connection with our current paper is that observability entails filter stability (see [29–31]) which then leads to near optimality of sliding window control policies, not unlike the case in classical stochastic linear systems theory[23] where observability implies near optimality of finite memory output feedback. Our paper makes this analysis precise for a large class of systems.

Contributions In this paper we develop rigorous approximations and learning results for POMDPs with Borel state, action, and measurement spaces. Furthermore, under regularity conditions, we obtain rigorous (finite dimensional) approximations, and under explicit filter stability conditions, we obtain near optimality bounds of finite window based policies which then leads to finite model approximations.

Notably, we quantize either the state space or the measurement space, noting that it is well known by now that action spaces can be quantized with arbitrarily low loss in performance under weak continuity conditions (see [32] and [33, Theorem 3.16]).

Quantization of the hidden state space results in a belief state process which takes values from a simplex. Hence, the problem can be solved via this simplex valued approximate belief state and leads to a finite-dimensional filter which can be computer with relative ease. We present sufficient conditions to guarantee the near optimality of this discretization method in Section 3.

Furthermore, in Section 4, we study the performance of finite window based policies and relate the performance loss due to finite window restriction to controlled filter stability.

In Section 5, towards a finite model approximation, we quantize the measurements and obtain performance loss bounds due to the approximations of the measurements to finite ones. This quantization leads to a garbling of information and therefore the performance loss can be obtained with such a perspective, as we show in the paper.

The analysis above culminates in a joint analysis of the performance loss due to measurement quantization and finite window restriction in Section 6.

By quantizing the measurement space we are able to utilize refined filter stability conditions. In particular, discretization on the observation variables results in an approximate non-linear filter where the conditioning is based on a less informative information process. We present explicit conditions under which the approximate filter is stable as well. This in turn implies near-optimality of the discretized finite-window information variables.

Finally, Section 7 applies the analysis above to reinforcement learning via finite memory Q-learning, which is shown to converge to near optimality under the conditions noted above.

Compared to the results in the literature, this paper is the first one to our knowledge which considers the case with continuous space measurements and establishes explicit conditions on the near optimality of finite window policies, and also with a learning algorithm under convergence

guarantees to near optimality.

Before we start out analysis, building on [34], we first review some technical tools we will need along the paper.

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

Two important notions of convergence for sequences of probability measures are weak convergence and convergence under total variation. For some $N \in \mathbb{N}$, a sequence $\{\mu_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathbb{X})$ is said to converge to $\mu \in \mathcal{P}(\mathbb{X})$ *weakly* if $\int_{\mathbb{X}} c(x)\mu_n(dx) \to \int_{\mathbb{X}} c(x)\mu(dx)$ 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

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

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

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

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

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

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

and the second equality follows from the dual formulation of the Wasserstein distance (see [35, Remark 6.5]). Note that the weak convergence and the Wasserstein convergence are equivalent if the underlying space is compact.

We can now define the following regularity properties for the transition kernels:

- $\mathcal{T}(\cdot|x, u)$ is said to be weakly continuous in (x, u), 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 as $t \to \infty$, $\frac{g(t)}{t} \uparrow \infty$, 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 1.1 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

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

where μ 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 so 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 (see [36, 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 (x, u) pair such that, there exists some $\alpha < \infty$ with

$$|f(x_n, u_n, w) - f(x, u, w)| \le \alpha (|x_n - x| + |u_n - u|),$$

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

$$W_1 \left(\mathcal{T}(\cdot | x_n, u_n), \mathcal{T}(\cdot | x, u) \right) = \sup_{Lip(g) \le 1} \left| \int g(x_1) \mathcal{T}(dx_1 | x_n, u_n) - \int g(x_1) \mathcal{T}(dx_1 | x, u) \right|$$
$$= \sup_{Lip(g) \le 1} \left| \int g(f(x_n, u_n, w)) \mu(dw) - \int g(f(x, u, w)) \mu(dw) \right|$$
$$\le \int |f(x_n, u_n, w) - f(x, u, w)| \, \mu(dw)$$
$$\le \alpha (|x_n - x| + |u_n - u|).$$

2 Partially Observed Markov Decision Processes and Standard Solution Methods

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

We will assume that the channel admits a density function $g(x,y)$ with respect to a reference measure $\lambda(dy) \in \mathcal{P}(\mathbb{Y})$ such that

$$O(dy|x) = g(x,y)\lambda(dy).$$

$\mathbb{U} \subset \mathbb{R}^k$ denotes the action space for some $k \in \mathbb{N}$.

An *admissible policy* γ is a sequence of control functions $\{\gamma_t, t \in \mathbb{Z}_+\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables

$$H_t = \{Y_{[0,t]}, U_{[0,t-1]}\}, \quad t \in \mathbb{N}, \quad I_0 = \{Y_0\},$$

where $U_t = \gamma_t(I_t), t \in \mathbb{Z}_+$, are the $\mathbb{U}$-valued control actions and $Y_{[0,t]} = \{Y_s, 0 \leq s \leq t\}$, $U_{[0,t-1]} = \{U_s, 0 \leq s \leq t-1\}$. When we consider a particular realization of the information variables, we use the notation h_t. We define Γ to be the set of all such admissible policies. The update rules of the system are determined by the following relationships:

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

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

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

$B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}), t \in \mathbb{N}$, where $\mathcal{T}$ is the transition kernel of the model which is a stochastic kernel from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{X}$. We let the objective of the agent (decision maker) be the minimization of the infinite horizon discounted cost,

$$J_\beta(\mu, O, \gamma) = \sum_{t=0}^{\infty} \beta^t E_\mu^{O,\gamma}\left[c(X_t, U_t)\right], \tag{1}$$

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

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

The solution of the problem in its current formulation requires one to find a control over the information variables

$$h_t = \{y_0, \cdots, y_t, u_0, \cdots, u_{t-1}\}.$$

Hence, the length of the information grown over time. In particular, the space I_t lives in, $\mathbb{Y}^t \times \mathbb{U}^{t-1}$ grows exponentially. This leads one to use compressed versions of the original information variable I_t. A generic solution method for POMDPs then involves a compression scheme

$$h_t \to z_t, \tag{2}$$

where the new state space $\mathcal{Z}$ does not grow over time and z_t defines controlled Markov chain lets one to use standard methods to solve the optimal control problem. Ideally, the compression map (2) is without loss of optimality. However, one might also work with lossy compression maps with controllable loss bounds, where the compressed state z_t computationally appealing, which will be the main focus of this paper.

In what follows, we will introduce different compression schemes, and present approximations with provable error bounds.

We start by introducing the most commonly used approach for the compression of the original information state, where the decision maker keeps track of the posterior distribution of the state X_t given the available history h_t. In the following section, we formalize this approach.

2.1   Reduction to Fully Observed Models Using Belief States

It is by now a standard result that, for optimality analysis, every POMDP can be reduced to a completely observable Markov decision process (see [37, 38]), whose states are the posterior state distributions or *beliefs* of the observer or the filter process; that is, the state at time t is

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

We call this equivalent process the filter process . The filter process lives in $\mathcal{Z} := \mathcal{P}(\mathbb{X})$ such that $\{z_t\}_t \subset \mathcal{P}(\mathbb{X})$. Therefore, the state space of the new model can be viewed as $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ and the action space remains as $\mathbb{U}$. Here, $\mathcal{Z}$ is equipped with the Borel σ-algebra generated by the topology of weak convergence[39]. Then, the transition probability η of the filter process can be constructed as follows. If we define the measurable function

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

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

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

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

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

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

It is well known that an optimal control policy of the original POMDP can use the belief Z_t as a sufficient statistic for optimal policies (see [37, 38]), provided they exist. More precisely, the filter process is equivalent to the original POMDP in the sense that for any optimal policy for the filter process, one can construct a policy for the original POMDP which is optimal. Existence then follows under measurable selection conditions, e.g., satisfied by the aforementioned weak Feller continuity of the belief MDP (see [40]). In particular, this applies when the belief-MDP is weak Feller (see [10, 11]), the action spaces are compact and the cost function is continuous and bounded.

Even though, the belief MDP approach provides a strong tool for the analysis of POMDPs, solution of POMDPs may not be computationally feasible in general. This stems from mainly two reasons:

- Computation of the belief state $h_t \to z_t = P(X_t \cdot |h_t)$ is not tractable in general, except for a few special cases, e.g., Kalman filter for linear and Gaussian additive noise dynamics, or for finite models.

- The belief space $\mathcal{Z} = \mathcal{P}(\mathbb{X})$ is always uncountable even when $\mathbb{X}$, $\mathbb{Y}$ and $\mathbb{U}$ are finite. Thus, even one is able to track and compute the belief state, the solution of the control may not be practical.

Therefore, approximation of the belief-MDP is usually needed.

2.2 Discretization of Action Sets

For a weak Feller belief MDP (see [10, 11]), [33, Theorem 3.16] (see also [41]) has established near optimality of finite action policies. If $\mathbb{U}$ is compact, a finite collection of action sets can be constructed, with arbitrary approximation error. Accordingly, we will assume that the action spaces are finite in the following.

3 Approximation via Quantization of the (Hidden) State Space

The first approach we present builds on the discretization of the hidden state space $\mathbb{X}$. This methods will rely on the model knowledge.

We first focus on the discretization of the observation variables. For the discretization, we choose disjoint subsets $\{A_i\}_{i=0}^{M} \subset \mathbb{X}$ such that $\cup_i A_i = \mathbb{X}$. We define a finite set

$$\mathbb{X}_M := \{x_0, \cdots, x_M\}.$$

Furthermore, the discretization map $\psi_{\mathbb{X}} : \mathbb{X} \to \widehat{\mathbb{X}}_M$ is defined such that

$$\psi_{\mathbb{X}}(x) = x_i, \quad \text{if } x \in A_i. \tag{6}$$

In other words, the discretization checks what bin or set x belongs to and maps it to the representative element of that bin.

Using this discretization scheme, we introduce appropriately normalized transition and channel kernels $\widehat{\mathcal{T}}$ and $\widehat{O}$, following and generalizing (see [6] and [33, Chapter 4]): We first choose a weighting measure $\pi(\cdot) \in \mathcal{P}(\mathbb{X})$. For some $x \in A_i$, we define

$$\widehat{\mathcal{T}}(\cdot|x,u) = \int_{A_i} \mathcal{T}(\cdot|z,u)\pi_i(dz), \quad \widehat{O}(\cdot|x) = \int_{A_i} O(\cdot|z)\pi_i(dz), \tag{7}$$

where $\pi_i(\cdot) = \frac{\pi(\cdot)}{\pi(A_i)}$. Note that the approximate channel density can be written similarly as

$$\widehat{g}(x,y) = \int_{A_i} g(z,y)\pi_z(dz).$$

We also define the following auxiliary cost function:

$$\widehat{c}(x,u) = \int c(z,u)\pi_i(dz).$$

These approximate kernel and functions can be seen to defined either on the finite set $\widehat{\mathbb{X}}$ or on the original space $\mathbb{X}$ by extending them as constant over the quantization bins.

We define the infinite horizon discounted cost for this approximation by

$$\widehat{J}_\beta(\mu, \widehat{O}, \gamma) = \sum_{t=0}^{\infty} \beta^t \widehat{E}_\mu^{\widehat{O},\gamma} \left[\widehat{c}(X_t, U_t) \right], \tag{8}$$

for some discount factor $\beta \in (0,1)$, over the set of admissible policies $\gamma \in \Gamma$, where $\widehat{E}_\mu^{\widehat{O},\gamma}$ denotes the expectation with initial state probability measure μ and the transition kernel for the discretized model $\widehat{\mathcal{T}}$ and the channel $\widehat{O}$ under policy γ. We define the optimal cost for the discretized model by

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

The above construction implies the following immediate result.

Assumption 3.1   We assume that

a. $|c(x,u) - c(x',u)| \le K_c \|x - x'\|$, for some $K_c < \infty$ and $\|c\|_\infty < \infty$.

b. $W_1\left(\mathcal{T}(\cdot|x,u), \mathcal{T}(\cdot|z,u)\right) \le K_T \|x - z\|$, for all $u \in \mathbb{U}$.

c. $\|O(\cdot|x) - O(\cdot|z)\|_{TV} \le K_O \|x - z\|$, (or $|g(x,y) - g(z,y)| \le K_O \|x - z\|$),

for all $x, z \in \mathbb{X}$ for some $K_f < \infty$.

Lemma 3.2 *Under Assumption* 3.1, *we have that*

$$W_1\left(\mathcal{T}(\cdot|x,u),\widehat{\mathcal{T}}(\cdot|x,u)\right) \leq K_T L_{\mathbb{X}},$$

$$\|O(\cdot|x) - \widehat{O}(\cdot|x)\|_{TV} \leq K_O L_{\mathbb{X}}, \quad (|g(x,y) - \widehat{g}(x,y)| \leq K_O L_{\mathbb{X}}),$$

where

$$L_{\mathbb{X}} = \max_{i \in \{1,\cdots,M\}} \sup_{x,z' \in A_i} \|x - z\|. \tag{9}$$

Proof The proof can be found in Appendix A.1. ∎

The following result will be critical to prove the main result of this section:

Lemma 3.3 *Under Assumption* 3.1, *we have that*

$$\left| E\left[c(X_t, \gamma(Y_{[0,t]}))\right] - \widehat{E}\left[\widehat{c}(X_t, \gamma(Y_{[0,t]}))\right] \right| \leq K_O \|c\|_\infty L_{\mathbb{X}} \sum_{n=0}^{t} \sum_{m=0}^{n} K_T^m + K_c L_{\mathbb{X}} \sum_{n=0}^{t} K_T^n,$$

for any policy $\gamma \in \Gamma$ *where* E *and* $\widehat{E}$ *represent the expectation operators respectively under the models* $(\mathcal{T}, O)$ *and* $(\widehat{\mathcal{T}}, \widehat{O})$.

Proof The proof can be found in Appendix A.2. ∎

Proposition 3.4 *Under Assumption* 3.1, *if we further assume that* $\beta K_T < 1$, *we can the write*

$$J_\beta(\mu_0, O, \gamma) - \widehat{J}_\beta(\mu_0, \widehat{O}, \gamma) \leq \frac{K_O \|c\|_\infty L_{\mathbb{X}}}{(1-\beta)^2(1-\beta K_T)} + \frac{K_c L_{\mathbb{X}}}{(1-\beta)(1-\beta K_T)},$$

for any admissible policy $\gamma \in \Gamma$.

Proof Note that we have

$$J_\beta(\mu_0, O, \gamma) = \sum_{t=0}^{\infty} \beta^t E_{\mu_0}^\gamma \left[c(X_t, \gamma(Y_{[0,t]}))\right], \quad \widehat{J}_\beta(\mu_0, \widehat{O}, \gamma) = \sum_{t=0}^{\infty} \beta^t \widehat{E}_{\mu_0}^\gamma \left[\widehat{c}(X_t, \gamma(Y_{[0,t]}))\right].$$

Hence the result is a direct application of Lemma 3.3. ∎

3.1 Near Optimal Policy Construction Based on Finite State Space

Recall that the control of the partially observed system can be reduced to the control of a belief MDP. For the belief state based control, the decision maker calculates and applies the optimal policy $\gamma^*(\cdot) : \mathcal{P}(\mathbb{X}) \to \mathbb{U}$, by tracking the belief state

$$\pi_t = \Pr(X_t \in \cdot | Y_{[0,t]}, U_{[0,t]}).$$

As a result of discretization of the hidden state space, the agent can construct the conditional probabilities on the finite state space $\widehat{\mathbb{X}}$, using the observations. At time t, the belief state on $\widehat{\mathbb{X}}$ can be written as

$$\widehat{\pi}_t(i) = \Pr(\widehat{X}_t = i | y_{[0,t]}, u_{[0,t-1]})$$

$$= \frac{\sum_{\widehat{x}_{t-1}} \widehat{g}(i, y_t) \widehat{\mathcal{T}}(i | \widehat{x}_{t-1}, u_{t-1}) \widehat{\pi}_{t-1}(\widehat{x}_{t-1})}{\sum_{\widehat{x}_{t-1}} \sum_{\widehat{x}_t} \widehat{g}(\widehat{x}_t, y_t) \widehat{\mathcal{T}}(\widehat{x}_t | \widehat{x}_{t-1}, u_{t-1}) \widehat{\pi}_{t-1}(\widehat{x}_{t-1})}$$

$$=: G(\widehat{\pi}_{t-1}, y_t, u_{t-1})(i), \tag{10}$$

and thus the decision maker can track the belief state via these iterations. Furthermore, the control policy can be solved for the belief state on the simplex $\mathcal{P}(\widehat{\mathbb{X}}) \subset \mathbb{R}^{|\widehat{\mathbb{X}}|}$ using the belief MDP that corresponds to the transition and channel model $\widehat{\mathcal{T}}$, and $\widehat{O}$ (or the channel density $\widehat{g}(\widehat{x}, y)$). The resulting policy $\widehat{\gamma} : \mathcal{P}(\widehat{\mathbb{X}}) \to \mathbb{U}$ can also be realized as a mapping from $h_t = \{y_{[0,t]}, u_{[0,t-1]}\}$ to $\mathbb{U}$ since the belief at any time t is also a function of h_t, i.e.,

$$\widehat{\gamma}(\widehat{\pi}_t) = \widehat{\gamma}(\Pr(\widehat{X}_t \in \cdot | y_{[0,t]}, u_{[0,t-1]})).$$

One can then directly apply Proposition 3.4 to find an upper-bound for the performance loss of the approximate control policy:

Theorem 3.5 *Under Assumption* 3.1, *if we further assume that* $\beta K_T < 1$, *we can the write*

$$J_\beta(\mu_0, O, \widehat{\gamma}) - J_\beta^*(\mu_0, O) \le \frac{2K_O}{(1-\beta)^2(1-\beta K_T)} L_{\mathbb{X}} + \frac{2K_c}{(1-\beta)(1-\beta K_T)} L_{\mathbb{X}},$$

where $\widehat{\gamma}$ *is the optimal policy for the control problem based on the discretized hidden state space* $\widehat{\mathbb{X}}$. *Note that the first term above represent the realized accumulated cost if one applies the policy based on the the finite space for the true system, whereas the second term is the optimal cost that can be achieved for the correct model under admissible policies.*

Proof   We write

$$J_\beta(\mu_0, O, \widehat{\gamma}) - J_\beta^*(\mu_0, O) \le \left| J_\beta(\mu_0, O, \widehat{\gamma}) - \widehat{J}_\beta^*(\mu_0, \widehat{O}) \right| + \left| \widehat{J}_\beta^*(\mu_0, \widehat{O}) - J_\beta^*(\mu_0) \right|.$$

The first term is bounded by Proposition 3.4 as both costs are induced by the same policy $\widehat{\gamma}$. For the second term, we can also use Proposition 3.4 with the following extra step: If $\widehat{J}_\beta^*(\mu_0, \widehat{O}) > J_\beta^*(\mu_0, O)$,

$$\left| \widehat{J}_\beta^*(\mu_0, \widehat{O}) - J_\beta^*(\mu_0, O) \right| \le \widehat{J}_\beta(\mu_0, \widehat{O}, \gamma^*) - J_\beta^*(\mu_0, O),$$

if $\widehat{J}_\beta^*(\mu_0, \widehat{O}) \le J_\beta^*(\mu_0, O)$, we can then write

$$\left| \widehat{J}_\beta^*(\mu_0, \widehat{O}) - J_\beta^*(\mu_0, O) \right| \le J_\beta(\mu_0, O, \widehat{\gamma}) - \widehat{J}_\beta^*(\mu_0, \widehat{O}),$$

where we used the fact that $J_\beta^*(\mu_0, O) = J_\beta(\mu_0, O, \gamma^*)$ and $\widehat{J}_\beta^*(\mu_0, \widehat{O}) = \widehat{J}_\beta(\mu_0, \widehat{O}, \widehat{\gamma})$, and the fact that $\widehat{\gamma}$ and γ^* are optimal for the corresponding models, so using any other policy increases the cost. Hence, in either case, we can use Proposition 3.4 which concludes the proof.   ∎

3.2 Asymptotic Analysis

In this section, we analyze the performance of the hidden state discretization method asymptotically. For this part, rather than choosing an arbitrary quantization scheme, we will work with a particularly constructed quantization scheme. We assume here that $\mathbb{X}$ is compact, and for some $n \in \mathbb{N}$, we select a finite set (whose existence is guaranteed under the compactness assumption) $\widehat{\mathbb{X}}_n = \{x_{n,1}, \cdots, x_{n,k_n}\}$ such that

$$\min_{1, \cdots, k_n} \|x - x_{n,i}\| < 1/n, \quad \text{ for all } x \in \mathbb{X}.$$

That is, $\widehat{\mathbb{X}}_n$ is a $1/n$ net in $\mathbb{X}$. We choose the construction sets A_i's, in a nearest neighbor way such that

$$\max_i \sup_{x,x' \in A_i} \|x - x'\| < 1/n.$$

We will construct the approximate transition and channel kernels in the same way as (3.1). But, we will denote them by $\widehat{\mathcal{T}}_n$ and $\widehat{O}_n$ to emphasize their dependence on $n \in \mathbb{N}$.

Furthermore, instead of using Assumption 3.1 with Lipschitz continuity we will assume the following:

Assumption 3.6 We assume that

i. $c(x, u)$ is continuous in (x, u).

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

iii. The channel $O(\cdot|x)$ is total variation continuous in x.

Remark 3.7 Under the density assumption on the channel $O(\cdot|x)$, we will sometimes equivalently use the assumption that $g(x, y)$ is continuous in x.

The following then is a direct implication of Assumption 3.6, see also [42, Theorem 16].

Lemma 3.8 *Let $\mathbb{U}$ be finite. Recall that $\mathbb{U}$ is assumed finite, with arbitrarily small loss. Under Assumption* 3.6*, we have that*

$$\widehat{\mathcal{T}}_n(\cdot|x_n, u) \to \mathcal{T}(\cdot|x, u),$$

weakly for every $x_n \to x$, and for every $u \in \mathbb{U}$. Furthermore,

$$\|\widehat{O}_n(\cdot|x_n) - O(\cdot|x)\|_{TV} \to 0,$$

for every $x_n \to x$, and for every $u \in \mathbb{U}$.

Theorem 3.9 *Under Assumption* 3.6*, we can then write*

$$J_\beta(\mu_0, O, \widehat{\gamma}_n) \to J_\beta^*(\mu_0, O),$$

where $\widehat{\gamma}_n$ is the optimal policy for the control problem based on the discretized hidden state space $\widehat{\mathbb{X}}_n$ and models $\widehat{\mathcal{T}}_n$ and $\widehat{O}_n$.

Proof The proof follows similar steps as in the proof of Theorem 3.5, also see [43, Theorem 4.4] where only transition discrepancy is considered. ∎

3.3   A Discussion on the Hidden Space Discretization

One can construct a belief MDP model based on the discretization of the hidden state space, and using the transition and channel models introduced in (7). This approximation results in a fully observed control problem where the state space is the simplex on $\widehat{\mathbb{X}}$ where $|\widehat{\mathbb{X}}| = M$.

As a result of this approximation, the belief state on the finite space can be computed iteratively using (10). Although, the resulting state space is continuous even after the discretization of the hidden space, one can apply further quantization directly on the resulting simplex to find near optimal controllers defined on a finite space; (see [44, Section 5]). Notably, observe that [45, Theorem 2.6] relates the Wasserstein error with the quantization diameter.

We also note that the hidden state space discretization indirectly approximates the belief space $\mathcal{P}(\mathbb{X})$ as well. In particular, for the finite space $\widehat{\mathbb{X}} = \{x_1, \cdots, x_M\}$, all the probability measures that assign the same measure to the quantization bins, B_i's, form an equivalence class after the discretization of the hidden space. That is if $P, Q \in \mathcal{P}(\mathbb{X})$ are such that $P(B_i) = Q(B_i)$ for all $i \in \{1, \cdots, M\}$ then we group P and Q together (and all the other measures that have the same property). Hence, the methods in [6, 33] and the Wasserstein and weak continuity of the belief kernel (see [11, 18, 19]) as an alternative approach to establish near optimality.

However, we should recall that this approximation relies on the knowledge of the model. In particular, the belief state on the finite set $\widehat{\mathbb{X}}$ is computed with the model knowledge. Hence, combining this approximation method with learning methods may not be possible if one only has access to the noisy observation, $y_{[0,t]}$, and control action, $u_{[0,t]}$, variables for learning period, the estimation of the kernels $\mathcal{T}(\cdot|x,u)$ and $O(\cdot|x)$ is possible if one has access to the hidden state process as well.

Hence, in what follows, we will focus on alternative compression schemes that uses finite memory information variables.

4   Finite Memory Based Compression

In this section, we will focus on compression schemes using the finite memory information variables. We note that unlike the methods in Section 3, we keep the hidden state $\mathbb{X}$ as it is, and do not discretize it for the techniques used in this section.

4.1   An Alternative Finite Window Belief-MDP Reduction

The construction in this section will build on [21]. The belief MDP reduction reveals that for the optimal performance, belief state is a sufficient knowledge about the system, hence, one needs to keep track of all the past observation and actions variables. In this section, we will focus on the effect of the finite memory use on the performance decrease as opposed to the full memory of the past. To this end, we construct an alternative fully observed MDP reduction using the predictor from N stages earlier and the most recent N information variables (that is, measurements and actions). This new construction allows us to highlight the most recent information variables and *compress* the information coming from the past history via the predictor as a probability measure valued variable.

Consider the following state variable at time t:

$$\widehat{z}_t = (\pi_{t-N}^-, h_t^N), \tag{11}$$

where, for $N \geq 1$

$$\pi_{t-N}^- = P^\mu(X_{t-N} \in \cdot | y_{t-N-1}, \cdots, y_0, u_{t-N-1}, \cdots, u_0),$$
$$h_t^N = \{y_t, \cdots, y_{t-N}, u_{t-1}, \cdots, u_{t-N}\},$$

and $h_t^N = y_t$ for $N = 0$ with μ being the prior probability measure on X_0. Above, we use the notation

$$P^\mu(X_{t-N} \in \cdot | y_{t-N-1}, \cdots, y_0, u_{t-N-1}, \cdots, u_0) := \Pr(X_{t-N} \in \cdot | y_{t-N-1}, \cdots, y_0, u_{t-N-1}, \cdots, u_0),$$

which is the posterior distribution of X_{t-N} conditioned on $\{y_{t-N-1}, \cdots, y_0, u_{t-N-1}, \cdots, u_0\}$ with a prior distribution $\mu \sim X_0$.

The state space with this representation is $\widehat{\mathcal{Z}} = \mathcal{P}(\mathbb{X}) \times \mathbb{Y}^{N+1} \times \mathbb{U}^N$ where we equip $\widehat{\mathcal{Z}}$ with the product topology where we consider the weak convergence topology on the $\mathcal{P}(\mathbb{X})$ coordinate and the usual (coordinate) topologies on $\mathbb{Y}^{N+1} \times \mathbb{U}^N$ coordinates.

We can now define the stage-wise cost function and the transition probabilities. Consider the new cost function $\widehat{c} : \widehat{\mathcal{Z}} \times \mathbb{U} \to \mathbb{R}$, using the cost function $\widetilde{c}$ of the belief MDP (defined in (5)) such that

$$\widehat{c}(\widehat{z}_t, u_t) = \widehat{c}(\pi_{t-N}^-, h_t^N, u_t) = \int_{\mathbb{X}} c(x_t, u_t) P^{\pi_{t-N}^-}(dx_t | y_t, \cdots, y_{t-N}, u_{t-1}, \cdots, u_{t-N}). \tag{12}$$

In particular, noting that

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

we have

$$P^{\pi_{t-N}^-}(dx_t | y_t, \cdots, y_{t-N}, u_{t-1}, \cdots, u_{t-N}) = P^\mu(dx_t | y_t, \cdots, y_0, u_{t-1}, \cdots, u_0),$$

which implies

$$\widehat{c}(\widehat{z}_t, u_t) = \widehat{c}(\pi_{t-N}^-, h_t^N, u_t) = \int_{\mathbb{X}} c(x_t, u_t) P^{\pi_{t-N}^-}(dx_t | y_t, \cdots, y_{t-N}, u_{t-1}, \cdots, u_{t-N})$$
$$= \int_{\mathbb{X}} c(x_t, u_t) P^\mu(dx_t | y_t, \cdots, y_0, u_{t-1}, \cdots, u_0).$$

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

$$A = B \times \{\widehat{y}_{t+1}, \widehat{y}_t, \widehat{u}_t\}, \quad B \in \mathcal{B}(\mathcal{P}(\mathbb{X})), \quad \widehat{y}_{t+1}, \widehat{y}_t, \widehat{u}_t \in \mathbb{Y}^2 \times \mathbb{U}$$

conditioned on a path of state and action variables $\widehat{z}_t, \cdots, \widehat{z}_0, u_t, \cdots, u_0$, we write

$$\Pr(\widehat{Z}_{t+1} \in A | \widehat{z}_t, \cdots, \widehat{z}_0, u_t, \cdots, u_0)$$

$$= \Pr(\Pi_t^- \in B, Y_{t+1} = \widehat{y}_{t+1}, Y_t = \widehat{y}_t, U_t = \widehat{u}_t | \pi_{[0,t-1]}^-, y_{[0,t]}, u_{[0,t]})$$

$$= \mathbb{1}_{\{y_t = \widehat{y}_t, u_t = \widehat{u}_t, G(\pi_{t-1}^-, y_{t-1}, u_{t-1}) \in B\}} P^{\pi_{t-1}^-}(\widehat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1})$$

$$= \Pr(\Pi_t^- \in B, Y_{t+1} = \widehat{y}_{t+1}, Y_t = \widehat{y}_t, U_t = \widehat{u}_t | \pi_{t-1}^-, y_t, y_{t-1}, u_t, u_{t-1})$$

$$= \Pr(\widehat{Z}_{t+1} \in A | \widehat{z}_t, u_t) =: \int_A \widehat{\eta}(d\widehat{z}_{t+1} | \widehat{z}_t, u_t), \tag{13}$$

where the map G is defined as

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

Hence, we have a proper fully observed MDP, with the cost function $\widehat{c}$, transition kernel $\widehat{\eta}$ and the state space $\widehat{\mathcal{Z}}$. Furthermore, we denote by $\widehat{\Gamma}$, the set of admissible policies for the new state model defined in this section.

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

$$J_\beta^*(\widehat{z}) = \inf_{u \in \mathbb{U}} \left(\widehat{c}(\widehat{z}, u) + \beta \int J_\beta^*(\widehat{z}_1) \widehat{\eta}(d\widehat{z}_1 | \widehat{z}, u) \right). \tag{14}$$

This construction results in a fully observed control problem with a state space $\widehat{\mathcal{Z}}$ that does not grow over time, and is without loss of optimality. However, the first coordinate of the state variable $\widehat{z}_t = (\pi_{t-N}^-, h_t^N)$, which is the predictor at time $t - N$, requires the model knowledge and can be practically hard to track much like the belief state. Therefore, we introduce an approximation method in the following section that maps this predictor to a fixed point and keeps the finite window information variables untouched.

4.2   Approximation of the Finite Window Belief-MDP

The alternative belief MDP model constructed in the previous section lives in the state space

$$\widehat{\mathcal{Z}} = \{\pi, y_{[0,N]}, u_{[0,N-1]} : \pi \in \mathcal{P}(\mathbb{X}), y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\},$$

where the first coordinate summarizes the past information, and the second and the last coordinates carry the information from the most recent N time steps.

Consider the following set $\widehat{\mathcal{Z}}_{\pi^*}^N$ for a fixed $\pi^* \in \mathcal{P}(\mathbb{X})$,

$$\widehat{\mathcal{Z}}_{\pi^*}^N = \{\pi^*, y_{[0,N]}, u_{[0,N-1]} : y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N\},$$

such that the state at time t is $\widehat{z}_t^N = (\pi^*, h_t^N)$. Compared to the state $\widehat{z}_t = (\pi_{t-N}^-, h_t^N)$ defined in (11), this approximate model uses π^* as the predictor, no matter what the real predictor at time $t - N$ is.

The cost function is defined as

$$\widehat{c}(\widehat{z}_t^N, u_t) = \widehat{c}(\pi^*, h_t^N, u_t) = \int_\mathbb{X} c(x_t, u_t) P^{\pi^*}(dx_t | y_t, \cdots, y_{t-N}, u_{t-1}, \cdots, u_{t-N}). \tag{15}$$

We define the controlled transition model for some $\widehat{z}_{t+1}^N = (\pi^*, h_{t+1}^N)$ and $\widehat{z}_t^N = (\pi^*, h_t^N)$,

$$\widehat{\eta}^N(\widehat{z}_{t+1}^N|\widehat{z}_t^N, u_t) = \widehat{\eta}^N(\pi^*, h_{t+1}^N|\pi^*, h_t^N, u_t) := \widehat{\eta}(\mathcal{P}(\mathbb{X}), h_{t+1}^N|\pi^*, h_t^N, u_t), \qquad (16)$$

where $\widehat{\eta}$ is the transition kernel of the alternative finite window belief MDP reduction, see (13). In words, starting from the state $\widehat{z}_t^N = (\pi^*, h_t^N)$ under the action u_t, for any possible future state under the dynamics of the true world (i.e., $\widehat{\eta}$), we apply a crude quantization for the first coordinate and map any $\pi_{t-N+1}^- \in \mathcal{P}(\mathbb{X})$ to π^*.

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

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

$$= \mathbb{1}_{\{y_t=\widehat{y}_t, u_t=\widehat{u}_t\}} P^{\pi^*}(\widehat{y}_{t+1}|y_t, y_{t-1}, u_t, u_{t-1}). \qquad (17)$$

We denote the optimal value function for the approximate model by J_β^N. We also denote by ϕ^N an optimal policy for the resulting approximate model. Note that both J_β^N and ϕ^N are defined on the set $\widehat{\mathcal{Z}}_{\pi^*}^N$. However, we can simply extend them to the set $\widehat{\mathcal{Z}}$ by defining for any $\widehat{z} = (\pi, y_1, y_0, u_0) \in \widehat{\mathcal{Z}}$ (assuming $N = 1$),

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

Another interpretation of the extension is the following: Since the predictor π^* is fixed ϕ^N can be thought as a map from h_t^N to $\mathbb{U}$, hence it maps the most recent N information variables to the control actions. At any given time, one can simply choose the control action by looking at the most recent N information variables according to ϕ^N and ignore the past. See Figure 1 for a comparison of finite memory belief reduction
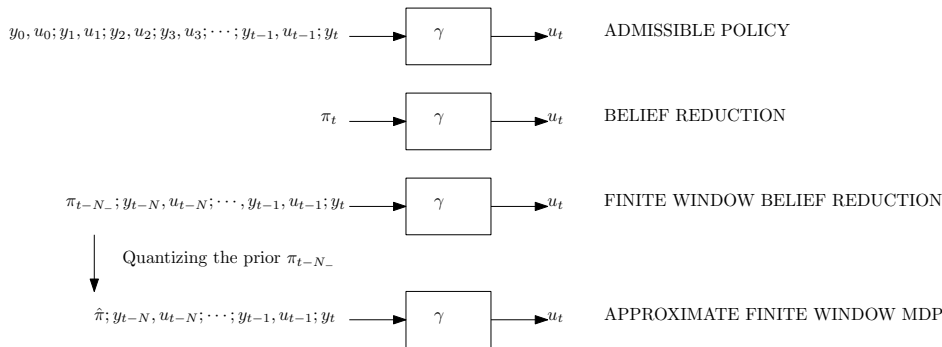

Figure 1 Construction of the Finite-Window Approximate MDP from the Finite-Window Belief-MDP[21]. The quantization of the finite window MDP model leads to the collapse of the first coordinate to a fixed measure

We define the following constant:

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

$$- P^{\pi^*}(X_{t+N} \in \cdot|Y_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV} \big], \qquad (18)$$

which is the expected bound on the total variation distance between the posterior distributions of X_{t+N} conditioned on the same observation and control action variables $Y_{[t,t+N]}, U_{[t,t+N-1]}$ when the prior distributions of X_t are given by π_t^- and π^*. Above, $\widehat{\Gamma}$ denote the set of admissible policies for the alternative finite window based state model defined in Subsection 4.1.

Proposition 4.1 (see [21, Theorem 3.3])   *Let* $X_0 \sim \mu$ *and consider* $H_0^N = \{Y_0, \cdots, Y_N, U_0, \cdots, U_{N-1}\}$ *with a policy* $\widehat{\gamma}$ *acting on the first* N *steps. For* $\widehat{Z}_0 = (\mu, H_0^N)$,

$$E_\mu^{\widehat{\gamma}} \left[\left| J_\beta(\widehat{Z}_0, \widetilde{\phi}^N) - J_\beta^*(\widehat{Z}_0) \right| \right] \le \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t,$$

where $J_\beta(\widehat{Z}_0, \widetilde{\phi}^N)$ *is the cost under the finite window policy* $\widetilde{\phi}^N$, *and* $J_\beta^*(\widehat{Z}_0)$ *is the optimal value function for the initial point* $\widehat{Z}_0$ *(see* (14)*). Furthermore, the expectation is with respect to the random realizations of* H_0^N.

The finite memory approximation method we have presented in this section results in a compression scheme with a loss of optimality since the information prior to time $t - N$ is ignored. This loss, on the other hand, is controllable with the term L_t (18). Furthermore, the state space of this approximation is $\mathbb{Y}^N \times \mathbb{U}^{N-1}$ where N is fixed and thus the state space does not grow over time. To track this finite memory state, the controller does not have to perform an extra compression step and can simply use the finite memory variables as they are.

However, for continuous observation spaces, the finite memory approximation will still not be leading to a finite model. To this end, in the next section, we will focus on the discretization of the observation space. We first focus on the observation discretization without the finite memory requirement, and then we combine the observation discretization with the finite memory method.

5   Quantization of Observations and Loss in Performance

The belief MDP reduction reveals that for the optimal performance, belief state is a sufficient knowledge about the system That is, at any given time t, the controller can use

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

Hence, one needs to keep track of all the past observation and control action variables. Even when the memory is not a constraint, computation of the belief state is not trivial as the observation space is assumed to be continuous.

We first focus on the discretization of the observation variables. For the discretization, we choose disjoint subsets $\{B_i\}_{i=0}^M \subset \mathbb{Y}$ such that $\cup_i B_i = \mathbb{Y}$. We define a finite set

$$\mathbb{Y}_M := \{y_0, \cdots, y_M\}.$$

Furthermore, the discretization map $\psi_{\mathbb{Y}} : \mathbb{Y} \to \widehat{\mathbb{Y}}_M$ is defined such that

$$\psi_{\mathbb{Y}}(y) = y_i, \quad \text{if } y \in B_i. \tag{19}$$

In other words, the discretization checks what bin or set y belongs to and maps it to the representative element of that bin. Selection of the disjoint sets depends on the application and the control problem.

Using the discretization procedure, one can define a further POMDP model, which uses a channel, say $\widehat{O}(\cdot|x)$, defined on the finite space $\widehat{\mathbb{Y}}_M$. Note that this channel is different than the one introduced in (7), but we use the same notation. For any $y_i \in \widehat{\mathbb{Y}}_M$

$$\widehat{O}(y_i|x) = O(B_i|x). \tag{20}$$

This channel is clearly degraded with respect to the original channel, since y_i is conditionally independent of x given the original observation y.

Therefore, for prior distribution on x_0 given by μ, one can define the optimal cost under the discrete observations by

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

where the policies, $\widehat{\gamma} \in \widehat{\Gamma}$ are measurable with respect to the σ-algebra generated by the new information variables where the observations take values from the finite set $\widehat{\mathbb{Y}}_M$.

5.1 Effect of Observation Discretization

In this section, we study the effect of observation discretization on the optimal value. Recall that the original channel is modeled as a stochastic kernel and denoted by $O(dy|x)$. Furthermore, we assume that this channel admits a density function with respect to a reference measure $\lambda(dy)$ such that

$$O(dy|x) = g(x, y)\lambda(dy).$$

The observation space discretization is introduced in Section 5, and the resulting channel is denoted by $\widehat{O}$, which is defined on the finite space $\widehat{\mathbb{Y}}_M$ such that for any $y_i \in \widehat{\mathbb{Y}}_M$

$$\widehat{O}(y_i|x) = O(B_i|x),$$

where B_i's are the quantization bins (or the aggregate sets). We will further assume that the observation space $\mathbb{Y} \subset \mathbb{R}^n$ is compact. Our goal in this section is to find bounds for the term

$$\inf_{\widehat{\gamma} \in \widehat{\Gamma}} J_\beta(\mu, \widehat{O}, \widehat{\gamma}) - \inf_{\gamma \in \Gamma} J_\beta(\mu, O, \gamma),$$

where $\widehat{\Gamma}$ is the set of policies that use the discretized observations, and Γ is the policies that use the original observations, and μ is the prior distribution of x_0. Furthermore, the first term is greater as $\widehat{O}$ is a degraded channel in comparison to O.

We define the following notation, denoting the uniform error bound due to the discretization of the observations

$$L_{\mathbb{Y}} = \max_{i \in \{1, \cdots, M\}} \sup_{y, y' \in B_i} |y - y'|. \tag{21}$$

We now state the assumptions and the result formally.

Assumption 5.1   • $O(dy|x) = g(x,y)\lambda(dy)$, and $g(y,x)$ is continuous in y for every $x \in \mathbb{X}$.

• Stage-wise cost function $c(x,u)$ is bounded such that $\sup_{x,u} c(x,u) := \|c\|_\infty < \infty$.

Assumption 5.2   • $\mathbb{Y} \subset \mathbb{R}^n$ is compact.

• $O(dy|x) = g(x,y)\lambda(dy)$, and $g(y,x)$ is Lipschitz in y, such that $|g(x,y) - g(x,y')| \leq \alpha_\mathbb{Y}|y - y'|$ for every $y, y' \in \mathbb{Y}$ and $x \in \mathbb{X}$ for some $\alpha_\mathbb{Y} < \infty$.

• Stage-wise cost function $c(x,u)$ is bounded such that $\sup_{x,u} c(x,u) := \|c\|_\infty < \infty$.

Theorem 5.3   *Under Assumption* 5.1*,*

$$\lim_{L_\mathbb{Y} \to 0} \left(J_\beta^*(\mu, \widehat{O}) - J_\beta^*(\mu, O) \right) = 0,$$

where $J_\beta^*(\mu, \widehat{O}) := \inf_{\widehat{\gamma} \in \widehat{\Gamma}} J_\beta(\mu, \widehat{O}, \widehat{\gamma})$ *and* $J_\beta^*(\mu, O) := \inf_{\gamma \in \Gamma} J_\beta(\mu, O, \gamma)$.

Theorem 5.4   *Under Assumption* 5.2*,*

$$J_\beta^*(\mu, \widehat{O}) - J_\beta^*(\mu, O) \leq \frac{\beta}{(1-\beta)^2}\|c\|_\infty \alpha_\mathbb{Y} L_\mathbb{Y}.$$

Proof   We will only prove Theorem 5.4, the proof for Theorem 5.3 follows from almost identical steps.

We start by introducing a new channel $O'(dy|x)$ acting on the original observation space $\mathbb{Y}$. The new channel is conditionally independent of everything else given the output of the discrete channel $\widehat{O}$, and it gives a uniformly distributed (with respect to the measure $\lambda(dy)$) random value from B_i if the output of the finite channel $\widehat{O}$ is y_i. Hence, the new channel O' admits a density function, say g', with respect to λ, such that for some $y' \in B_i$

$$g'(x, y') = \frac{O(B_i|x)}{\lambda(B_i)} = \frac{1}{\lambda(B_i)} \int_{B_i} g(x,y)\lambda(dy).$$

Note also that, since we assume that $\mathbb{Y}$ is compact, the subsets B_i's are bounded and the uniform distribution is well defined on these sets. Using the assumption that $g(x,y)$ is Lipschitz in y, we can conclude that

$$|g'(x,y) - g(x,y)| \leq \alpha_\mathbb{Y} \max_i \sup_{y,y' \in B_i} |y - y'| \leq \alpha_\mathbb{Y} L_\mathbb{Y}. \tag{22}$$

This new channel is introduced only for mathematical convenience, in order to define the degraded POMDP on the same observation space as the original POMDP model. Furthermore, since the only information one can infer about the hidden state x, is through the discrete observations y_i, the channel O' is a garbling of the channel $\widehat{O}$. Hence, it is a direct consequence of Blackwell's informativeness theorem[46] that

$$J_\beta^*(\mu, O') \geq J_\beta(\mu, \widehat{O}).$$

In fact, one can show that the costs are indeed equal, however, we will only use the inequality as it is sufficient for our purpose. Furthermore, since O' sends values from the original observation space $\mathbb{Y}$, the POMDP model with O' also uses the policies from the set Γ. Therefore, denoting the optimal policy designed for the original channel O by γ^*, if we apply this policy for the channel O', we get a larger cost, i.e., $J_\beta(\mu, O', \gamma^*) \geq J_\beta^*(\mu, O')$. Hence, for the rest of the proof, we will assume that both models use the same policy.

By [47, Theorem 6.2] we establish the following relation:

$$|J_\beta(\mu, O', \gamma) - J_\beta(\mu, O, \gamma)| \leq \sum_k \beta^k \|c\|_\infty k \sup_{x \in \mathbb{X}} \|O'(.|x) - O(.|x)\|_{TV}.$$

Furthermore, using (22), we can write that

$$\sup_{x \in \mathbb{X}} \|O'(.|x) - O(.|x)\|_{TV} \leq \alpha_{\mathbb{Y}} L_{\mathbb{Y}},$$

which concludes the proof. ∎

6 Near Optimality of Finite Memory Feedback with Approximated Discrete Observations

Although, the observations are discretized, the resulting belief state still takes values from the space of probability measures and the sufficient information for the optimal control of the POMDP with the finite observation space, requires the controller to keep track of all the history. Therefore, as the next step of the approximation, we only focus on a finite memory of the information variables. We restrict the controller to use the new information variables

$$\widehat{H}_t^N = \{\widehat{Y}_{[t-N,t]}, U_{[t-N,t-1]}\},$$

where $\widehat{Y}_t$ takes values from the finite set $\widehat{\mathbb{Y}}_M$. If we denote the set of policies that use $\widehat{H}_t^N$, by $\widehat{\Gamma}^N$, the question we ask is the following: Can we find bounds on

$$\inf_{\gamma^N \in \widehat{\Gamma}^N} J_\beta(\mu, \widehat{O}, \gamma^N) - J_\beta^*(\mu, O), \tag{23}$$

that is what do we loose if we use a finite memory of discretized observations in comparison the optimal cost of the original POMDP with continuous observations. This problem also motivates the natural, though interesting, question: If one has a computational constraint, is it better to use a finer discretization of the observation space, or use a longer memory by keeping the dicretizaton rate the same.

To analyze the problem, we first discretize the observations. We then analyze the effect of finite memory with the discretized observations. On the effect of the finite memory on the performance, in Subsection 4.2, we have presented bounds on the loss if one only uses a fixed length of most recent information variables. We remind that the presented results are be valid for any channel model including the degraded channel model resulting from the observation discretization.

6.1   Combined Approximation Algorithm and the Near Optimality of the Approximate Policies Under Filter Stability

In this section, we introduce the approximation methodology, which involves a combination of observation space discretization and use of finite memory. We then provide an upper bound on the performance loss when the approximation policy is applied in the original POMDP model, in terms of the approximation parameters.

The approximation algorithm is as follows:

1) Pick disjoint subsets $\{B_i\}_{i=1}^M \subset \mathbb{Y}$ such that $\cup_i B_i = \mathbb{Y}$. Furthermore, pick the representative observation set $\widehat{\mathbb{Y}}_M = \{y_1, \cdots, y_M\}$, and the discretization map $\psi_{\mathbb{Y}} : \mathbb{Y} \to \widehat{\mathbb{Y}}_M$ is defined such that

$$\psi_{\mathbb{Y}}(y) = y_i, \quad \text{if } y \in B_i.$$

2) Define a new channel $\widehat{O}(\cdot|x)$ on the finite set $\widehat{\mathbb{Y}}_M$ such that for any $y_i \in \widehat{\mathbb{Y}}_M$,

$$\widehat{O}(y_i|x) = O(B_i|x).$$

3) Pick a window length N, and define a fully observed MDP, with state space $\widehat{\mathbb{Y}}_M^N \times \mathbb{U}^{N-1}$, the cost function is defined by (15), and the transition model defined by (16) or (17). Note that, to define the cost and the transitions, the new channel $\widehat{O}$ is used.

4) Solve the optimality equation for the finite fully observed MDP, to construct a policy $\widehat{\gamma}^N$.

Theorem 6.1 *Suppose Assumption* 5.1 *holds and the belief kernel η is weakly continuous* ([10, 11]). *Let $X_0 \sim \mu$ and consider $H_0^N = \{Y_0, \cdots, Y_N, U_0, \cdots, U_{N-1}\}$ with a policy $\widehat{\gamma}$ acting on the first N steps. For $\widehat{Z}_0 = (\mu, H_0^N)$ (recall (11)), if the policy found after the steps above, $\widehat{\gamma}^N$ acts after the first N steps*

$$\lim_{L_{\mathbb{Y}} \to 0} E_\mu^{\widehat{\gamma}} \left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\widehat{Z}_0) \right] \leq \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^\infty \beta^t L_t,$$

where L_t is the filter stability term for the original channel O and the expectation is with respect to the random realizations of H_0^N.

Theorem 6.2 *Suppose Assumption* 5.2 *holds. Let $X_0 \sim \mu$ and consider $H_0^N = \{Y_0, \cdots, Y_N, U_0, \cdots, U_{N-1}\}$ with a policy $\widehat{\gamma}$ acting on the first N steps. For $\widehat{Z}_0 = (\mu, H_0^N)$ (recall (11)), if the policy found after the steps above, $\widehat{\gamma}^N$ acts after the first N steps*

$$E_\mu^{\widehat{\gamma}} \left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\widehat{Z}_0) \right] \leq \frac{\beta}{(1-\beta)^2} \|c\|_\infty \alpha_{\mathbb{Y}} L_{\mathbb{Y}} + \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^\infty \beta^t L_t,$$

where

$$L_{\mathbb{Y}} := \max_i \sup_{y,y' \in B_i} |y - y'|,$$

$$L_t := \sup_{\widehat{\gamma} \in \widehat{\Gamma}} E_\mu^{\widehat{\gamma}, \widehat{O}} \left[\|P^{\pi_t^-}(X_{t+N} \in \cdot|\widehat{Y}_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\pi^*}(X_{t+N} \in \cdot|\widehat{Y}_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV} \right],$$

and $\alpha_{\mathbb{Y}}$ is the Lipschitz constant of the density function g of the channel O.

Proof [For Theorems 6.1 and 6.2] We first consider the marginal distribution of X_2:

$$\mu_2(\cdot) = \int \mathcal{T}(\cdot|x_1, \widehat{\gamma}(y_1, y_0))O(dy_1|x_1)\mathcal{T}(dx_1|x_0, \widehat{\gamma}(y_0))O(dy_0|x_0)\mu(dx_0).$$

Note that we have

$$E_\mu^{\widehat{\gamma}}\left[J_\beta^*(\widehat{Z}_0)\right] = J_\beta^*(\mu_2, O).$$

We write

$$E_\mu^{\widehat{\gamma}}\left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\widehat{Z}_0)\right] \leq E_\mu^{\widehat{\gamma}}\left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\mu_2, \widehat{O})\right] + E_\mu^{\widehat{\gamma}}\left[J_\beta^*(\mu_2, \widehat{O}) - J_\beta^*(\widehat{Z}_0)\right]$$

$$= E_\mu^{\widehat{\gamma}}\left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\mu_2, \widehat{O})\right] + E_\mu^{\widehat{\gamma}}\left[J_\beta^*(\mu_2, \widehat{O}) - J_\beta^*(\mu_2, O)\right].$$

Note that $\widehat{\gamma}^N$ is the finite window policy for the channel $\widehat{O}$. The first difference is bounded by Proposition 4.1. Therefore, the first bound is related to the controlled filter stability under the discrete channel $\widehat{O}$. For the second term, we use Theorem 5.4.

For the proof of Theorem 6.1, the second term goes to 0, by Theorem 5.3. For the first term, we again use Proposition 4.1 with the fact that the filter stability term under channel $\widehat{O}$ converges to the filter stability term under the channel O if the belief kernel η is weak Feller. ∎

6.1.1 A Discussion on Theorem 6.2

One can observe that the first term in the upper bound presented in Theorem 6.2, is only related to the observation discretization and is not affected by the finite memory use. The result indicates that higher Lipschitz constants will lead to higher performance improvements as the dicretization rate grows. Clearly, if the discretization rate is already high, increasing it further will not make the same impact. For example, for a bounded observation space $\mathbb{Y}$, if the size of the chosen finite set is M, $L_\mathbb{Y}$ can be approximated by $\frac{1}{M}$ under uniform quantization. Thus, increasing the size to $M + 1$ would result in

$$K\alpha_\mathbb{Y}\frac{1}{M} - K\alpha_\mathbb{Y}\frac{1}{M+1} = K\alpha_\mathbb{Y}\frac{1}{M(M+1)}$$

decrease on the first term of the upper bound where $K = \frac{\beta\|c\|_\infty}{(1-\beta)^2}$.

The second term is related to the controlled filter stability under the discrete channel. This term is affected by both the window length N, and the rate of the discretization. Increasing the window length will clearly decrease the second term. The better discretization rate will change the second term, since how informative the observations are will also impact the filter stability. One can perform simulations to study the exact impact of the discretization on the controlled filter stability. In the following, we will provide some upper bounds for the second term, using Dobrushin coefficient (Definition 6.3) of the original channel O.

6.1.2 Further Upper-Bounds for Filter Stability via Dobrushin Coefficient

In this section, we discuss the term L_t term defined in (18).

Definition 6.3 ([48, Equation 1.16])   For a kernel operator $K : S_1 \to \mathcal{P}(S_2)$ (that is a regular conditional probability from S_1 to S_2) for standard Borel spaces S_1, S_2, we define the Dobrushin coefficient as:

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

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

Note that for the channel O Dobrushin coefficient is readily defined, for the transition kernel T, let

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

We now state the following corollary to Theorem 6.2 and [49, Theorem 3.3].

Corollary 6.4   *Suppose Assumption 5.2 holds and let $\alpha := (1 - \widetilde{\delta}(\mathcal{T}))(2 - \delta(O)) < 1$. Let $X_0 \sim \mu$ and consider $H_0^N = \{Y_0, \cdots, Y_N, U_0, \cdots, U_{N-1}\}$ with a policy $\widehat{\gamma}$ acting on the first N steps. For the policy found with the algorithm presented in Subsection 6.1, $\widehat{\gamma}^N$, acting after the first N steps we can write that,*

$$E_\mu^{\widehat{\gamma}} \left[J_\beta(\widehat{Z}_0, \widehat{\gamma}^N) - J_\beta^*(\widehat{Z}_0) \right] \leq \frac{\|c\|_\infty}{(1-\beta)^2} \left(\beta \alpha_\mathbb{Y} L_\mathbb{Y} + 4\alpha^N \right).$$

Proof   [49, Theorem 3.3] implies that

$$L_t := \sup_{\widehat{\gamma} \in \widehat{\Gamma}} E_\mu^{\widehat{\gamma}, \widehat{O}} \left[\|P^{\pi_t^-}(X_{t+N} \in \cdot | \widehat{Y}_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\pi^*}(X_{t+N} \in \cdot | \widehat{Y}_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV} \right]$$

$$\leq 2 \left((1 - \widetilde{\delta}(\mathcal{T}))(2 - \delta(\widehat{O})) \right)^N.$$

Hence, we only need to show that

$$\delta(\widehat{O}) \geq \delta(O).$$

However, this is a direct consequence of the definition of Dobrushin coefficient, since for any discrete channel $\widehat{O}$, the quantization bins $\{B_i\}$ forms a partition for $\mathbb{Y}$. ∎

Corollary 6.4 separates the effects of discretization and finite memory use, in two different terms. However, we caution the reader that this is only done for an upper bound. In fact, we do not bound the filter stability term of the discrete channel $\widehat{O}$, using the filter stability term of the original channel O. Instead, we bound an upper bound of the filter stability term for $\widehat{O}$, using an upper bound for the filter stability term of the channel O. This can be done since [49, Theorem 3.3] favors "non-informative" channels.

6.1.3   Utilizing Finite Measurements via a Hilbert Projective Metric Based Bound for Filter Stability

Via a somewhat different, and more direct, derivation, [20, Subsection 4.2 and Theorem 17] presented the following alternative condition involving sample path-wise uniform filter stability term

$$\overline{L}_{TV}^N := \sup_{pi \in \mathcal{P}(\mathbb{X})} \sup_{y_{[0,N]}, u_{[0,N-1]}} \left\| P^\pi(\cdot|y_{[0,N]}, u_{[0,N-1]}) - P^{\pi^*}(\cdot|y_{[0,N]}, u_{[0,N-1]}) \right\|_{TV}, \tag{24}$$

where π^* has been fixed in Subsection 4.2. to show the following *uniform* error bound:

$$\sup_z \left| J_\beta(z, \gamma_N) - J_\beta^*(z) \right| \leq \frac{2(1 + (\alpha_{\mathcal{Z}} - 1)\beta)}{(1-\beta)^3(1 - \alpha_{\mathcal{Z}}\beta)} \|c\|_\infty \overline{L}_{TV}^N, \tag{25}$$

for all $\beta \in (0,1)$ under a contraction condition, for some constant $\alpha_{\mathcal{Z}}$ defined in [20].

Accordingly, a further method, and one which leads to complementary conditions given the above, for filter stability is via the Hilbert projective metric (see [50, Lemma 3.8]). Via the Birkhoff-Hopf theorem, a controlled version of a contraction via the Hilbert metric[50] can be utilized[19]:

Definition 6.5 (Mixing kernel) The non-negative kernel K defined on $\mathbb{X}$ is mixing, if there exist a constant $0 < \varepsilon \leq 1$, and a non-negative measure λ on $\mathbb{X}$, such that

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

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

With

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

we note the following:

Assumption 6.6 1) $O(y|x) \geq \varepsilon > 0$ for every $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.

2) The transition kernel $\mathcal{T}(.|., u)$ is a mixing kernel (see Definition 6.5) for every $u \in \mathbb{U}$.

Lemma 6.7 (see [51]) *Under Assumption 6.6, there exists a constant $r < 1$ such that*

$$h(F(\mu, y, u), F(\nu, y, u)) \leq rh(\mu, \nu), \tag{26}$$

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

Theorem 6.8 (see [19]) *Under Assumption 6.6, there exist a constant $r < 1$ and K such that*

$$\overline{L}_{TV}^N \leq r^{N-1}K. \tag{27}$$

Here, $K = \frac{2}{\log 3} \sup h(\pi_1, \pi_1^*)$ and $r = \sup_{u \in \mathbb{U}} \frac{1 - \varepsilon_u^2 \varepsilon}{1 + \varepsilon_u^2 \varepsilon}$ where $\pi_1 = F(\pi_0, y, u)$ and $\pi_1^* = F(\pi^*, y, u)$ and the supremum is taken over all π_0, π^*, y, u.

Corollary 6.9 (see [19]) *Under Assumption 6.6, there exist a constant $r < 1$ and K such that*

$$E_{\pi_0^-}^{\widehat{\gamma}} \left[\left| \widetilde{J}_\beta^N \left(\widehat{Z}_0, \widetilde{\phi}^N \right) - J_\beta^* \left(\widehat{Z}_0 \right) \right| \right] \leq \frac{2\|c\|_\infty}{(1-\beta)^2} r^{N-1} K. \tag{28}$$

Here, $K = \frac{2}{\log 3} \sup_{\pi_0^-, y, u} h(F(\pi_0^-.y, u), F(\pi^*, y, u))$ and $r = \sup_{u \in \mathbb{U}} \frac{1 - \varepsilon_u^2 \varepsilon}{1 + \varepsilon_u^2 \varepsilon}$.

The implication of the quantized approximation is that we can ensure that upon quantization each bin can satisfy the condition that $\widehat{O}(y|x) > \varepsilon > 0$. That is, the original measurement kernel may not satisfy the contraction condition but the finite measurement model may, leading to the applicability of the bound above. Note that the conditions are complementary when compared with those building on the Dobrushin coefficient based analysis above.

6.2   Comparison of Discretization Rate and Finite Window Length

Consider, for example, the original channel O, is fully "non-informative" and provides a random observation y uniformly distributed on $\mathbb{Y}$, then the discretization of the observation space will not lead to any improvement as the Lipschitz constant, $\alpha_{\mathbb{Y}} = 0$. However, using a longer window, will improve the performance. In particular, for this case the Dobrushin coefficient of the channel becomes $\delta(O) = 1$, thus $\alpha = (1 - \delta(\widetilde{T}))$.

For the other extreme case, we consider a "fully informative" channel. Since, we have a density assumption on the channel, consider instead an additive channel of the from $y_t = x_t + v_t$ where v_t is a zero mean Gaussian with small variance. Note that, although we assume $\mathbb{Y}$ to be compact, we can extend the analysis by focusing on high probability events on compact subsets. For this case, finer quantization of the observation space will lead to greater improvements on the performance as the Lipschitz constant of the density of the channel will be large. However, as the channel is already informative, using a longer memory will not provide further information on the hidden state x.

For the extreme cases, we observe that for "fully informative" channels it is better to increase the discretization rate rather than increasing the window length. For "fully non-informative" channels, however, one should increase the window length instead of increasing the discretization rate. Even though, we observe this trend for the extreme cases, for general observation models, one needs give a quantitative definition of an "informative" channel. [30, 52] study stochastic observability for measurement channels for non-linear dynamics and establish the filter stability. This notion can be used to further compare the effect of discretization and memory use for "observable" channels. We should also note that increasing the size of the finite observation set $\widehat{\mathbb{Y}}_M$ from M to $M + 1$, will increase the size of the state space of the fully observed MDP constructed by the algorithm in Subsection 6.1 by a factor of $\left(\frac{M+1}{M}\right)^N$, and increasing the window size from N to $N + 1$ will increase the state space size by a factor of $M \times |\mathbb{U}|$.

Remark 6.10   We note that if the goal is only asymptotic convergence of the approximation error, we can relax the geometric convergence conditions of the filter stability (and only consider stochastic observability of the controlled kernels (see [30]) so that $L_t \to 0$ as the window length increases with no rate of convergence), $\mathbb{U}$ is compact (not necessarily finite) and the density of the measurement kernels, denoted with f is only continuous in y, with no Lipschitz regularity.

7   Q-Learning for the Approximate Models

7.1   Q-Learning Using Belief Discretization

We will first introduce a discretization based Q learning over the belief spaces. We first set a finite number of quantization bins, $\{Z_1, \cdots, Z_k\} \subset \mathcal{P}(\mathbb{X})$ such that $\cup_{i=1}^k = \mathcal{P}(\mathbb{X})$. We use the map $\Psi : \mathcal{P}(\mathbb{X}) \to \{Z_1, \cdots, Z_k\}$ for quantization.

A natural way of choosing the quantization for the belief space is as follows: (i) Quantize the hidden space $\mathbb{X}$ to a finite set, (ii) for the resulting simplex use further discretization over the probability mass of the resulting bins (see [44]).

Finally, we use the following Q learning algorithm over the finite set of quantization bins. We denote the i to denote the quantization bin Z_i:

$$Q_{k+1}(i, u) = (1 - \alpha_k(i, u))Q_k(i, u) + \alpha_k(i, u)\left(C_k + \beta \min_v Q_k(\psi(\pi_{t+1}), v)\right).$$

This algorithm can be shown to converge under suitable ergodicity conditions on the belief process (see [53, Subsection 3.4]). Furthermore, the near optimality of the learned policies can be guaranteed using the weak continuity of the belief kernel (see [34, Section 3.3]).

We emphasize, however, that this algorithm assumes that the controller has access to the belief process $\{\pi_t\}$. In other words, the controller can map the the history h_t to the belief state π_t, and this mapping can be challenging in general.

7.2   Q-Learning Using Finite Memory with Discrete Observations

In this section, we show that the proposed approximate model using the finite memory can be learned using a Q learning algorithm.

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

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

For every observation y_t, we use the representative element from the finite $\widehat{\mathbb{Y}}_M$, using the map $\psi_{\mathbb{Y}}$ (see (19)). With an abuse of notation, we use $\psi_{\mathbb{Y}}(h_t^N)$ to denote the finite history information with discrete observations.

For these new approximate states, we follow the usual Q learning algorithm such that for any $I \in \mathbb{Y}^{N+1} \times \mathbb{U}^N$ and $u \in \mathbb{U}$,

$$\begin{aligned} Q_{k+1}(\psi_{\mathbb{Y}}(I), u) &= (1 - \alpha_k(\psi_{\mathbb{Y}}(I), u))Q_k(\psi_{\mathbb{Y}}(I), u) \\ &\quad + \alpha_k(\psi_{\mathbb{Y}}(I), u)\left(C_k + \beta \min_v Q_k(\psi_{\mathbb{Y}}(I_1), v)\right), \end{aligned} \tag{29}$$

where $I_1 = \{Y_{t+1}, y_t, \cdots, y_{t-N+1}, u_t, \cdots, u_{t-N+1}\}$ which are the finite memory information variables we observe following I and the control u. C_k denotes the cost we observe at time time k after we apply u, hence, if the hidden state is x_k, we have $C_k = c(x_k, u)$.

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

Assumption 7.1   1) $\alpha_t(\psi_{\mathbb{Y}}(I), u) = 0$ unless $(\psi_{\mathbb{Y}}(h_t), u_t) = (\psi_{\mathbb{Y}}(I), u)$. Furthermore,

$$\alpha_t(\psi_{\mathbb{Y}}(I), u) = \frac{1}{1 + \sum_{k=0}^t 1_{\{\psi_{\mathbb{Y}}(I_k) = \psi_{\mathbb{Y}}(I), u_k = u\}}}.$$

We note that, this means that $\alpha_k(\psi_{\mathbb{Y}}(I), u) = \frac{1}{k}$ if $\psi_{\mathbb{Y}}(I_k) = \psi_{\mathbb{Y}}(I), u_k = u$, if k is the instant of the kth visit to $(\psi_{\mathbb{Y}}(I), u)$.

2) Under every stationary {memoryless or finite memory exploration} policy, say γ, the true state process, $\{X_t\}_t$, is positive Harris recurrent and in particular admits a unique invariant measure π^*_γ.

3) During the exploration phase, every $(\psi_{\mathbb{Y}}(I), u)$ pair is visited infinitely often.

Theorem 7.2   *Under Assumption* 7.1,

i) *The algorithm given in* (29) *converges almost surely to some* Q^* *which are the optimal* Q *values of an MDP constructed in Subsection* 4.2 *with the discrete channel* $\widehat{O}$.

ii) *For any policy* γ^N *that is constructed using* Q^* *if we assume that the controller has access to at least* $N + 1$ *observations and* N *control action variables, when it starts acting, we have*

$$J_\beta(\mu, O, \gamma^N) - J^*_\beta(\mu, O) \leq \frac{\beta}{(1-\beta)^2}\|c\|_\infty \alpha_{\mathbb{Y}} L_{\mathbb{Y}} + \frac{2\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t.$$

Proof   The result is a direct implication of [21, Theorem 4.1]. In [21, Theorem 4.1], it is shown that the Q learning using finite memory information variables for a POMDP with a finite observation space, converges. The Q iterations constructed here then can be thought to learn a POMDP model with the discrete channel $\widehat{O}$ which has a finite observation space. Hence, (ii) is a result of Theorem 6.2. ∎

The positive Harris recurrence condition can be relaxed; please see [53] and [54, Lemma 6].

8   Concluding Remarks and a Discussion

We studied two approximation algorithms for POMDPs with continuous state and observation spaces. The first method discretizes the hidden state space and genrates a belief MDP defined over a simplex. The near optimality of this method relies on the regularity of the transition and the channel kernel of the POMDP. Furthermore, this method is tailored more towards settings where the model is known. The second method is done by dicretizing the observation space and using finite memory of discrete information variables. The near optimality of this method relies on the controlled filter stability under the discretized observations. Furthermore, we provided a Q learning algorithm for the constructed approximate finite models.

Conflict of Interest

The authors declare no conflict of interest.

References

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

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

[3] Vlassis N and Spaan M T J, Perseus: Randomized point-based value iteration for pomdps, *Journal of Artificial Intelligence Research*, 2005, **24**: 195–220.

[4] Pineau J, Gordon G, and Thrun S, Anytime point-based approximations for large pomdps, *Journal of Artificial Intelligence Research*, 2006, **27**: 335–380.

[5] Smith T and Simmons R, Point-based pomdp algorithms: Improved analysis and implementation, 2012, arXiv: 1207.1412.

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

[7] Saldi N, Yüksel S, and Linder T, Finite model approximations for partially observed Markov decision processes with discounted cost, *IEEE Transactions on Automatic Control*, 2020, arXiv: 1710.07009.

[8] Zhou E, Fu M C, and Marcus S I, A density projection approach to dimension reduction for continuous-state POMDPs, *Decision and Control, 2008, CDC 2008, 47th IEEE Conference on*, 2008, 5576–5581.

[9] Zhou E, Fu M C, and Marcus S I, Solving continuous-state POMDPs via density projection, *IEEE Transactions on Automatic Control*, 2010, **55**(5): 1101–1116.

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

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

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

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

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

[15] Krishnamurthy V, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*, Cambridge University Press, Cambridge, 2016.

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

[17] White III C C and Scherer W T, Finite-memory suboptimal design for partially observed Markov decision processes, *Operations Research*, 1994, **42**(3): 439–455.

[18] Demirci Y E, Kara A D, and Yüksel S, Average cost optimality of partially observed mdps: Contraction of non-linear filters and existence of optimal solutions, *SIAM Journal on Control and Optimization*, 2004, **62**: 2859–2883.

[19] Demirci Y E, Kara A D, and Yüksel S, Refined bounds on near optimality finite window policies in pomdps and their reinforcement learning, 2024, arXiv: 2409.04351.

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

[21] Kara A D and Yüksel S, Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability, *Mathematics of Operations Research*, 2023,

        48(4): 2066–2093.

[22]   Kara A D and Yuksel S, Partially observed optimal stochastic control: Regularity, optimality, approximations, and learning, *IEEE* 51*st IEEE Conference on Decision and Control* (*CDC*), 2024, arXiv: 2412.06735.

[23]   Caines P E, *Linear Stochastic Systems*, John Wiley & Sons, New York, 1988.

[24]   Caines P E and Zhang J-F, On the adaptive control of jump parameter systems via nonlinear filtering, *SIAM Journal on Control and Optimization*, 1995, **33**(6): 1758–1777.

[25]   Ades M, Caines P E, and Malhamé R P, Stochastic optimal control under poisson-distributed observations, *IEEE Transactions on Automatic Control*, 2000, **45**(1): 3–13.

[26]   Sen N and Caines P E, Mean field games with partial observation, *SIAM Journal on Control and Optimization*, 2019, **57**(3): 2064–2091.

[27]   Caines P E, Greiner R, and Wang S, Classical and logic-based dynamic observers for finite automata, *IMA Journal of Mathematical Control and Information*, 1991, **8**(1): 45–80.

[28]   Ng S K and Caines P E, Nonlinear filtering in riemannian manifolds, *IMA Journal of Mathematical Control and Information*, 1985, **2**: 25–36.

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

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

[31]   van Handel R, Nonlinear filtering and systems theory, *Proceedings of the* 19*th International Symposium on Mathematical Theory of Networks and Systems* (*MTNS semi-plenary paper*), 2010, https://api.semanticscholar.org/CorpusID:16553338.

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

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

[34]   Kara A D, Saldi N, and Yüksel S, Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity, *Journal of Machine Learning Research*, 2023, **199**: 1–34.

[35]   Villani C, *Optimal Transport: Old and New*, Springer, New York, 2009.

[36]   Billingsley P, *Probability and Measure*, 3rd Edition, Wiley, New York, 1995.

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

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

[39]   Billingsley P, *Convergence of Probability Measures*, 2nd Edition, Wiley, New York, 1999.

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

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

321–337.

[42] Kara A D and Yüksel S, Robustness to approximations and model learning in MDPs and POMDPs, *Modern Trends in Controlled Stochastic Processes: Theory and Applications, V. III, Springer*, 2021, **41**: 166–191.

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

[44] Saldi N, Yüksel S, and Linder T, Asymptotic optimality of finite model approximations for partially observed Markov decision processes with discounted cost, *IEEE Transactions on Automatic Control*, 2019, **65**(1): 130–142.

[45] Kreitmeier W, Optimal vector quantization in terms of Wasserstein distance, *Journal of Multivariate Analysis*, 2011, **102**(8): 1225–1239.

[46] Blackwell D, Equivalent comparisons of experiments, *The Annals of Mathematical Statistics*, 1953, **24**: 265–272.

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

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

[49] McDonald C and Yüksel S, Exponential filter stability via Dobrushin's coefficient, *Electronic Communications in Probability*, 2019, DOI: 10.48550/arXiv.1910.08463.

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

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

[52] Van Handel R, Discrete time nonlinear filters with informative observations are stable, *Electron. Commun. Probab.*, 2008, **13**: 562–575.

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

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

Appendix

A.1 Proof of Lemma 3.2

Consider some $x \in A_i$, we then write

$$
\begin{aligned}
W_1\left(\mathcal{T}(\cdot|x,u), \widehat{\mathcal{T}}(\cdot|x,u)\right) &= \sup_{\|f\|_{Lip}\leq 1} \left| \int f(x_1)\mathcal{T}(dx_1|x,u) - \int f(x_1)\widehat{\mathcal{T}}(dx_1|x,u) \right| \\
&= \sup_{\|f\|_{Lip}\leq 1} \left| \int f(x_1)\mathcal{T}(dx_1|x,u) - \int f(x_1)\mathcal{T}(dx_1|z,u)\pi_i(dz) \right| \\
&\leq \sup_{\|f\|_{Lip}\leq 1} \int_{z\in A_i} \left| \int_{x_1\in\mathbb{X}} f(x_1)\mathcal{T}(dx_1|x,u) - \int f(x_1)\mathcal{T}(dx_1|z,u) \right| \pi_i(dz)
\end{aligned}
$$

$$\leq \int_{z \in A_i} W_1 \left(\mathcal{T}(\cdot|x, u), \mathcal{T}(\cdot|z, u) \right) \pi_i(dz)$$

$$\leq K_T \sup_{z \in A_i} \|x - z\| \leq K_T L_{\mathbb{X}}.$$

The proof for the bound on the approximate channel follows from identical steps. ∎

A.2   Proof of Lemma 3.3

We start by defining the following functions for rotational convenience.

$$F_k(x_{t-k}, y_{[0,t-k]}) = E\left[c(X_t, \gamma(Y_{[0,t]})) | x_{t-k}, y_{[0,t-k]} \right],$$
$$\widehat{F}_k(x_{t-k}, y_{[0,t-k]}) = \widehat{E}\left[\widehat{c}(X_t, \gamma(Y_{[0,t]})) | x_{t-k}, y_{[0,t-k]} \right],$$

in particular

$$F_0(x_t, y_{[0,t]}) = c(x_t, y_{[0,t]}), \quad \widehat{F}_0(x_t, y_{[0,t]}) = \widehat{c}(x_t, y_{[0,t]}).$$

Furthermore, we are interested in the difference $|E[F_t(x_0, y_0)] - \widehat{E}[\widehat{F}_t(x_0, y_0)]|$.

We first note the following due to iterated expectations

$$F_k(x_{t-k}, y_{[0,t-k]})$$
$$= E\left[c(X_t, \gamma(Y_{[0,t]})) | x_{t-k}, y_{[0,t-k]} \right]$$
$$= E\left[E\left[c(X_t, \gamma(Y_{[0,t]})) | X_{t-k+1}, Y_{[0,t-k+1]} \right] | x_{t-k}, y_{[0,t-k]} \right]$$
$$= E\left[F_{k-1}(X_{t-k+1}, Y_{[0,t-k+1]}) | x_{t-k}, y_{[0,t-k]} \right]$$

$$= \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \mathcal{T}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})), \qquad \text{(A.1)}$$

which can also be written for the functions $\widehat{F}_k$. Using this iterative form, we find an upper-bound for the Lipschitz constant of $F_k(x_{t-k}, y_{[0,t-k]})$ with respect to x_{t-k} uniformly over all $y_{[0,t-k]}$:

$$\left| F_k(x_{t-k}, y_{[0,t-k]}) - F_k(x'_{t-k}, y_{[0,t-k]}) \right|$$
$$= \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \mathcal{T}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right.$$
$$\left. - \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \mathcal{T}(dx_{t-k+1}|x'_{t-k}, \gamma(y_{[0,t-k]})) \right|$$
$$\leq \left\| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \right\|_{Lip} K_T |x_{t-k} - x'_{t-k}|$$
$$\leq (K_O \|c\|_\infty + \|F_{k-1}\|_{Lip}) K_T \|x_{t-k} - x'_{t-k}\|,$$

where we used Assumption 3.1 (b) for the first inequality. For the last step, we used the following:

$$\left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \int F_{k-1}(x'_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x'_{t-k+1}) \right|$$

$$\leq \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \right.$$

$$\left. - \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x'_{t-k+1}) \right|$$

$$+ \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x'_{t-k+1}) \right.$$

$$\left. - \int F_{k-1}(x'_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x'_{t-k+1}) \right|$$

$$\leq \|F_{k-1}\|_\infty \|O(\cdot|x_{t-k+1}) - O(\cdot|x'_{t-k+1})\|_{TV} + \|F_{k-1}\|_{Lip} \|x_{t-k+1} - x'_{t-k+1}\|$$

$$\leq (K_O \|c\|_\infty + \|F_{k-1}\|_{Lip}) \|x_{t-k+1} - x'_{t-k+1}\|.$$

One can then show that

$$\|F_k\|_{Lip} \leq K_O \|c\|_\infty \left(K_T + \cdots + K_T^k \right) + \|F_0\|_{Lip} K_T^k$$

$$= K_O \|c\|_\infty \left(K_T + \cdots + K_T^k \right) + K_c K_T^k. \tag{A.2}$$

The iterative form in (A.1) implies that

$$\left| F_k(x_{t-k}, y_{[0,t-k]}) - \widehat{F}_k(x_{t-k}, y_{[0,t-k]}) \right|$$

$$= \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \mathcal{T}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right.$$

$$\left. - \int \widehat{F}_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) \widehat{O}(dy_{t-k+1}|x_{t-k+1}) \widehat{\mathcal{T}}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right|$$

$$\leq \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \mathcal{T}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right.$$

$$\left. - \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \widehat{\mathcal{T}}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right|$$

$$+ \left| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \widehat{\mathcal{T}}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right.$$

$$\left. - \int \widehat{F}_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) \widehat{O}(dy_{t-k+1}|x_{t-k+1}) \widehat{\mathcal{T}}(dx_{t-k+1}|x_{t-k}, \gamma(y_{[0,t-k]})) \right|$$

$$\leq \left\| \int F_{k-1}(x_{t-k+1}, y_{[0,t-k+1]}) O(dy_{t-k+1}|x_{t-k+1}) \right\|_{Lip} \sup_{x,u} W_1(\mathcal{T}(\cdot|x,u), \widehat{\mathcal{T}}(\cdot|x,u))$$

$$+ \sup_x \left| \int F_{k-1}(x, y_{[0,t-k+1]}) O(dy_{t-k+1}|x) - \int \widehat{F}_{k-1}(x, y_{[0,t-k+1]}) \widehat{O}(dy_{t-k+1}|x) \right|$$

$$\leq \sup_{x,u} W_1(\mathcal{T}(\cdot|x,u), \widehat{\mathcal{T}}(\cdot|x,u)) (\|F_{k-1}\|_{Lip} + K_O \|c\|_\infty) + \|c\|_\infty \sup_x \|O(\cdot|x) - \widehat{O}(\cdot|x)\|_{TV}$$

$$+ \|F_{k-1} - \widehat{F}_{k-1}\|_\infty$$

$$\leq K_T L_{\mathbb{X}} \|F_{k-1}\|_{Lip} + K_T K_O \|c\|_\infty L_{\mathbb{X}} + K_O \|c\|_\infty L_{\mathbb{X}} + \|F_{k-1} - \widehat{F}_{k-1}\|_\infty,$$

where we used Lemma 3.2. Using (A.2), we can find a further upper-bound as:

$$
\begin{aligned}
\left\|F_k - \widehat{F}_k\right\|_\infty \leq & K_T L_\mathbb{X} \left(K_O \|c\|_\infty \left(K_T + \cdots + K_T^{k-1}\right) + K_c K_T^{k-1}\right) + K_T K_O \|c\|_\infty L_\mathbb{X} \\
& + K_O \|c\|_\infty L_\mathbb{X} + \|F_{k-1} - \widehat{F}_{k-1}\|_\infty \\
\leq & K_O \|c\|_\infty L_\mathbb{X} \left(K_T^2 + \cdots + K_T^k\right) + K_c K_T^k L_\mathbb{X} + K_O \|c\|_\infty K_T L_\mathbb{X} + K_O \|c\|_\infty L_\mathbb{X} \\
& + \|F_{k-1} - \widehat{F}_{k-1}\|_\infty \\
= & K_O \|c\|_\infty L_\mathbb{X} \sum_{m=0}^{k} (K_T)^m + K_c (K_T)^k L_\mathbb{X} + \|F_{k-1} - \widehat{F}_{k-1}\|_\infty .
\end{aligned}
$$

Denoting by $A_k := \sum_{m=0}^{k} K_T^m$, we can then write:

$$
\begin{aligned}
\left\|F_k - \widehat{F}_k\right\|_\infty & \leq K_O \|c\|_\infty L_\mathbb{X} \sum_{n=1}^{k} A_n + K_c L_\mathbb{X} \sum_{n=1}^{k} (K_T)^n + \|F_0 - \widehat{F}_0\|_\infty \\
& \leq K_O \|c\|_\infty L_\mathbb{X} \sum_{n=1}^{k} A_n + K_c L_\mathbb{X} \sum_{n=1}^{k} (K_T)^n + K_c L_\mathbb{X} \\
& \leq K_O \|c\|_\infty L_\mathbb{X} \sum_{n=0}^{k} A_n + K_c L_\mathbb{X} \sum_{n=0}^{k} (K_T)^n \\
& = K_O \|c\|_\infty L_\mathbb{X} \sum_{n=0}^{k} \sum_{m=0}^{n} K_T^m + K_c L_\mathbb{X} \sum_{n=0}^{k} K_T^n .
\end{aligned}
$$

The proof is completed.   ∎