Modern Trends in Controlled Stochastic Processes : - Theory and Applications, V.III

Alexey B. Piunovskiy and Yi Zhang
editors

Table of Contents

Robustness to Approximations and Model Learning in MDPs and POMDPs

Ali Devran Kara and Serdar Yüksel

Queen's University
Department of Mathematics and Statistics, Kingston, ON, Canada
16adk@queensu.ca; yuksel@queensu.ca

Abstract. In stochastic control applications, typically only an ideal model (controlled transition kernel) is assumed and the control design is based on the given model, raising the problem of performance loss due to the mismatch between the assumed model and the actual model. In some further setups, an exact model may be known, but this model may entail computationally challenging optimality analysis leading to the solution of some approximate model being implemented. With such a motivation, we study continuity properties of discrete-time stochastic control problems with respect to system models and robustness of optimal control policies designed for incorrect models applied to the true system. We study both fully observed and partially observed setups under an infinite horizon discounted expected cost criterion. We show that continuity can be established under total variation convergence of the transition kernels under mild assumptions and with further restrictions on the dynamics and observation model under weak and setwise convergence of the transition kernels. Using these, we establish convergence results and error bounds due to mismatch that occurs by the application of a control policy which is designed for an incorrectly estimated system model to the actual system, thus establishing results on robustness. These entail implications on empirical learning in (data-driven) stochastic control since often system models are learned through empirical training data where typically the weak convergence criterion applies but stronger convergence criteria do not. We finally view and establish approximation as a particular instance of robustness.

Keywords: Markov decision processes, robust stochastic control, approximate models, empirical learning, POMDPs
AMS(2020) subject classification: 93E20, 90C40, 90C39.

1 Introduction and Problem Definition

In this article, we study the robustness problem of Markov Decision Processes (MDPs) and partially observed Markov decision processes (POMDPs) with incomplete/incorrect characterization, and view learning and approximate modeling as instances of the robustness problem. The article builds on some recent work of the authors but the models considered here are more general (involving changing cost functions also in the MDP models), and the unifying relationship between robustness and finite model approximations involving standard Borel models has not been studied elsewhere, to our knowledge.

Let $\mathbb{X} \subset \mathbb{R}^m$ denote a Borel set which is the state space of a partially observed controlled Markov process. Here and throughout the paper $\mathbb{Z}_+$ denotes the set of non-negative integers and $\mathbb{N}$ denotes the set of positive integers. Let $\mathbb{Y} \subset \mathbb{R}^n$ be a Borel set denoting the observation space of the model, and let the state be observed through an observation channel Q. The observation channel, Q, is defined as a stochastic kernel (regular conditional probability) from $\mathbb{X}$ to $\mathbb{Y}$, such that $Q(\,\cdot\,|x)$ is a probability measure on the (Borel) σ-algebra $\mathcal{B}(\mathbb{Y})$ of $\mathbb{Y}$ for every $x \in \mathbb{X}$, and $Q(A|\,\cdot\,) : \mathbb{X} \to [0,1]$ is a Borel measurable function for every $A \in \mathcal{B}(\mathbb{Y})$. A decision maker (DM) is located at the output of the channel Q, and hence it only sees the observations $\{Y_t,\, t \in \mathbb{Z}_+\}$ and chooses its actions from $\mathbb{U}$, the action space which is a Borel subset of some Euclidean space. An *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$ such that γ_t is measurable with respect to the σ-algebra generated by the information variables

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

where

$$U_t = \gamma_t(I_t), \quad t \in \mathbb{Z}_+, \tag{1}$$

are the $\mathbb{U}$-valued control actions and

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

We define $\mathit{\Gamma}$ to be the set of all such admissible policies. The update rules of the system are determined by (1) and the following:

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

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

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

where $\mathcal{T}$ is the transition kernel of the model. The objective of the agent (decision maker) is the minimization of the infinite horizon discounted cost,

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

for some discount factor $\beta \in (0,1)$, over the set of admissible policies $\gamma \in \mathit{\Gamma}$, where $c : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ is a Borel-measurable stage-wise cost function and $E_P^{\mathcal{T},\gamma}$ denotes the expectation with initial state probability measure P and transition kernel $\mathcal{T}$ under policy γ. Note that we write the infinite horizon discounted cost as a function of the transition kernels and the stage-wise cost function since we will analyze the cost under the changes on those variables.

We define the optimal cost for the discounted infinite horizon setup as a function of the stage-wise cost function and the transition kernels as

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

Problem P1: Continuity of $J_\beta^*(c,\mathcal{T})$ under the Convergence of the Models. Let $\{\mathcal{T}_n, n\in\mathbb{N}\}$ be a sequence of transition kernels which converges in some sense to another transition kernel $\mathcal{T}$ and $\{c_n, n\in\mathbb{N}\}$ be a sequence of stage-wise cost functions corresponding to $\mathcal{T}_n$ which converge in some sense to another cost function c. Does that imply that

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

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

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

Problem P4: Approximation by Finite MDPs as an Instance of Robustness to Incorrect Models. Can we view the approximation problem of a continuous space MDP model with a finite model (in particular [22, Theorem 2.2], [22, Theorem 4.1] or [23, Theorem 3.2]) as an instance of the robustness problem?

Brief Literature Review. Robustness is a desired property for the optimal control of stochastic or deterministic systems when a given model does not reflect the actual system perfectly, as is usually the case in practice. This is a classical problem, and there is a very large literature on robust stochastic control and its application to learning-theoretic methods; see e.g. [1, 2, 7, 8, 14, 16, 18, 20, 21, 25, 26]. A rather comprehensive literature review is presented in [18]. The article builds on [16, 18], but the models considered considered here are more general (involving changing cost functions also in the MDP models), and the unifying relationship between robustness and finite model approximations involving standard Borel models has not been studied elsewhere, to our knowledge.

1.1 Some Examples and Convergence Criteria for Transition Kernels

Convergence Criteria for Transition Kernels. Before presenting convergence criteria for controlled transition kernels, we first review the convergence of probability measures. Three important notions of convergences for sets of probability measures to be studied in the paper are weak convergence, setwise convergence, and convergence under total variation. For $N \in \mathbb{N}$, a sequence $\{\mu_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathbb{R}^N)$ is said to converge to $\mu \in \mathcal{P}(\mathbb{R}^N)$ *weakly* if

$$\int_{\mathbb{R}^N} c(x)\mu_n(dx) \to \int_{\mathbb{R}^N} c(x)\mu(dx) \tag{$*$}$$

for every continuous and bounded $c : \mathbb{R}^N \to \mathbb{R}$. $\{\mu_n\}$ is said to converge *setwise* to $\mu \in \mathcal{P}(\mathbb{R}^N)$ if $(*)$ holds for all measurable and bounded $c : \mathbb{R}^N \to \mathbb{R}$. For probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^N)$, the *total variation* metric is given by

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

where the supremum is taken over all measurable real f such that $\|f\|_\infty = \sup_{x \in \mathbb{R}^N} |f(x)| \le 1$. A sequence $\{\mu_n\}$ is said to converge in total variation to $\mu \in \mathcal{P}(\mathbb{R}^N)$ if $\|\mu_n - \mu\|_{TV} \to 0$. Total variation defines a stringent metric for convergence; for example, a sequence of discrete probability measures does not converge in total variation to a probability measure which admits a density function. Setwise convergence, though, induces a topology on the space of probability measures which is not metrizable [10, p. 59]. However, the space of probability measures on a complete, separable, metric (Polish) space endowed with the topology of weak convergence is itself complete, separable, and metric [19]. We also note here that relative entropy convergence, through Pinsker's inequality [11, Lemma 5.2.8], is stronger than even total variation convergence, which has also been studied in robust stochastic control. Another metric for probability measures is the Wasserstein distance: For compact spaces, the Wasserstein distance of order 1 metrizes the weak topology and for non-compact spaces convergence in the W_1 metric implies weak convergence. Considering these relations, our results in this paper can be directly generalized to the relative entropy distance or the Wasserstein distance. Building on the above, we introduce the following convergence notions for (controlled) transition kernels.

Definition 1. *For a sequence of transition kernels $\{\mathcal{T}_n, n \in \mathbb{N}\}$, we say that*

- *$\mathcal{T}_n \to \mathcal{T}$ weakly if $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ weakly, for all $x \in \mathbb{X}$ and $u \in \mathbb{U}$,*
- *$\mathcal{T}_n \to \mathcal{T}$ setwise if $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ setwise, for all $x \in \mathbb{X}$ and $u \in \mathbb{U}$,*
- *$\mathcal{T}_n \to \mathcal{T}$ under the total variation distance if $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ under total variation for all $x \in \mathbb{X}$ and $u \in \mathbb{U}$.*

Examples [18]. Let a controlled model be given as $x_{t+1} = F(x_t, u_t, w_t)$, where $\{w_t\}$ is an i.i.d. noise process. The uncertainty on the transition kernel for such a system may arise from lack of information on F or the i.i.d. noise process w_t or both:

(i) Let $\{F_n\}$ denote an approximating sequence for F, so that $F_n(x,u,w) \to F(x,u,w)$ pointwise. Assume that the probability measure of the noise is known. Then, corresponding kernels $\mathcal{T}_n$ converge weakly to $\mathcal{T}$: If we denote the probability measure of w with μ, for any $g \in C_b(\mathbb{X})$ and for any $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$ using the dominated convergence theorem we have

$$\lim_{n\to\infty} \int g(x_1)\mathcal{T}_n(dx_1|x_0,u_0) = \lim_{n\to\infty} \int g(F_n(x_0,u_0,w))\mu(dw)$$
$$= \int g(F(x_0,u_0,w))\mu(dw) = \int g(x_1)\mathcal{T}(dx_1|x_0,u_0).$$

(ii) Much of the robust control literature deals with deterministic systems where the nominal model is a deterministic perturbation of the actual model (see e.g. [24]). The considered model is in the following form: $\tilde{F}(x_t,u_t) = F(x_t,u_t)$ $+\Delta F(x_t,u_t)$, where F represents the nominal model and ΔF is the model uncertainty satisfying some norm bounds. For such deterministic systems, pointwise convergence of $\tilde{F}$ to the nominal model F, i.e. $\Delta F(x_t,u_t) \to 0$, can be viewed as weak convergence for deterministic systems by the discussion in (i). It is evident, however, that total variation convergence would be too strong for such a convergence criterion, since $\delta_{\tilde{F}(x_t,u_t)} \to \delta_{F(x_t,u_t)}$ weakly but $\|\delta_{\tilde{F}(x_t,u_t)} - \delta_{F(x_t,u_t)}\|_{TV} = 2$ for all $\Delta F(x_t,u_t) \neq 0$ where δ denotes the Dirac measure.

(iii) Let $F(x_t,u_t,w_t) = f(x_t,u_t) + w_t$ be such that the function f is known and $w_t \sim \mu$ is not known correctly and an incorrect model μ_n is assumed. If $\mu_n \to \mu$ weakly, setwise, or in total variation, then the corresponding transition kernels $\mathcal{T}_n$ converge in the same sense to $\mathcal{T}$. Observe the following:

$$\int g(x_1)\mathcal{T}_n(dx_1|x_0,u_0) - \int g(x_1)\mathcal{T}(dx_1|x_0,u_0)$$
$$= \int g(w_0 + f(x_0,u_0))\mu_n(dw_0) - \int g(w_0 + f(x_0,u_0))\mu(dw_0). \qquad (2)$$

(a) Suppose $\mu_n \to \mu$ weakly. If g is a continuous and bounded function, then $g(\cdot + f(x_0,u_0))$ is a continuous and bounded function for all $(x_0,u_0) \in \mathbb{X} \times \mathbb{U}$. Thus, (2) goes to 0. Note that f does not need to be continuous. (b)

Suppose $\mu_n \to \mu$ setwise. If g is a measurable and bounded function, then $g(\cdot + f(x_0, u_0))$ measurable and bounded for all $(x_0, u_0) \in \mathbb{X} \times \mathbb{U}$. Thus, (2) goes to 0. (c) Finally, assume $\mu_n \to \mu$ in total variation. If g is bounded, (2) converges to 0, as in item (b). As a special case, assume that μ_n and μ admit densities h_n and h, respectively; then the pointwise convergence of h_n to h implies the convergence of μ_n to μ in total variation by Scheffé's theorem.

(iv) Suppose now neither F nor the probability model of w_t is known perfectly. It is assumed that w_t admits a measure μ_n and $\mu_n \to \mu$ weakly. For the function F we again have an approximating sequence $\{F_n\}$. If $F_n(x, u, w_n) \to F(x, u, w)$ for all $(x, u) \in \mathbb{X} \times \mathbb{U}$ and for any $w_n \to w$, then the transition kernel $\mathcal{T}_n$ corresponding to the model F_n converges weakly to the one of F, $\mathcal{T}$: For any $g \in C_b(\mathbb{X})$,

$$\lim_{n\to\infty} \int g(x_1)\mathcal{T}_n(dx_1|x_0, u_0) = \lim_{n\to\infty} \int g(F_n(x_0, u_0, w))\mu_n(dw)$$
$$= \int g(F(x_0, u_0, w))\mu(dw) = \int g(x_1)\mathcal{T}(dx_1|x_0, u_0).$$

(v) Let again $\{F_n\}$ denote an approximating sequence for F and suppose now $F_{x_0,u_0,n}(\cdot) := F_n(x_0, u_0, \cdot) : \mathbb{W} \to \mathbb{X}$ is invertible for all $x_0, u_0 \in \mathbb{X} \times \mathbb{U}$ and $F^{-1}_{(x_0,u_0),n}(\cdot)$ denotes the inverse for fixed (x_0, u_0). It is assumed that $F^{-1}_{(x_0,u_0),n}(x_1) \to F^{-1}_{x_0,u_0}(x_1)$ pointwise for all (x_0, u_0). Suppose further that the noise process w_t admits a continuous density $f_W(w)$. The Jacobian matrix, $\frac{\partial x_1}{\partial w}$, is the matrix whose components are the partial derivatives of x_1, i.e. with $x_1 \in \mathbb{X} \subset \mathbb{R}^m$ and $w \in \mathbb{W} \subset \mathbb{R}^m$, it is an $m \times m$ matrix with components $\frac{\partial (x_1)_i}{\partial w_j}$, $1 \leq i, j \leq m$. If the Jacobian matrix of derivatives $\frac{\partial x_1}{\partial w}(w)$ is continuous in w and nonsingular for all w, then we have that the density of the state variables can be written as

$$f_{X_1,n,(x_0,u_0)}(x_1) = f_W(F^{-1}_{x_0,u_0,n}(x_1))\Big|\frac{\partial x_1}{\partial w}(F^{-1}_{x_0,u_0,n}(x_1))\Big|^{-1},$$
$$f_{X_1,(x_0,u_0)}(x_1) = f_W(F^{-1}_{x_0,u_0}(x_1))\Big|\frac{\partial x_1}{\partial w}(F^{-1}_{x_0,u_0}(x_1))\Big|^{-1}.$$

With the above, $f_{X_1,n,(x_0,u_0)}(x_1) \to f_{X_1,(x_0,u_0)}(x_1)$ pointwise for all fixed (x_0, u_0). Therefore, by Scheffé's theorem, the corresponding kernels $\mathcal{T}_n(\cdot|x_0, u_0) \to \mathcal{T}(\cdot|x_0, u_0)$ in total variation for all (x_0, u_0).

(vi) These examples will be utilized in Section 5.1, where data-driven stochastic control problems will be considered where estimated models are obtained through empirical measurements of the state action variables.

1.2 Summary

We now introduce the main assumptions that will be occasionally used for our technical results in the article.

Assumption 1 *(a) The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\{\mathcal{T}_n(\cdot|x_n, u_n), n \in \mathbb{N}\}$ converges weakly to $\mathcal{T}(\cdot|x, u)$ for any sequence $\{x_n, u_n\} \subset \mathbb{X} \times \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $(x_n, u_n) \to (x, u)$.*

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

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

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

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

Assumption 2 *The observation channel $Q(\cdot|x)$ is continuous in total variation i.e., if $x_n \to x$, then $Q(\,\cdot\,|x_n) \to Q(\,\cdot\,|x)$ in total variation (only for partially observed models).*

Assumption 3 *(a) The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\{\mathcal{T}_n(\cdot|x, u_n), n \in \mathbb{N}\}$ converges setwise to $\mathcal{T}(\cdot|x, u)$ for any sequence $\{u_n\} \subset \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $u_n \to u$.*

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

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

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

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

Assumption 4 *(a) The sequence of transition kernels $\mathcal{T}_n$ satisfies the following: $\|\mathcal{T}_n(\cdot|x, u_n) - \mathcal{T}(\cdot|x, u)\|_{TV} \to 0$ for any sequence $\{u_n\} \subset \mathbb{U}$ and $x, u \in \mathbb{X} \times \mathbb{U}$ such that $u_n \to u$.*

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

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

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

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

In Sections 2 and 3 we study continuity (Problem P1) and robustness (Problem P2) for partially observed models. In particular we show the following:

(a) Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 1).

(b) Under Assumptions 1 and 2, continuity and robustness hold (Theorem 4, Theorem 8).

(c) Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 5).

(d) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 1).

(f) Under Assumption 4, continuity and robustness hold (Theorem 6, Theorem 7).

In Section 4, we study continuity (Problem P1) and robustness (Problem P2) for fully observed models. In particular we show the following

(a) Continuity and robustness do not hold in general under weak convergence of kernels (Theorem 9, Example 1).
(b) Under Assumption 1, continuity holds (Theorem 10), under Assumption 1, robustness holds if the optimal policies for every initial point are identical (Theorem 11).
(c) Continuity and robustness do not hold in general under setwise convergence of the kernels (Theorem 12, Theorem 14).
(d) Under Assumption 3, continuity holds (Theorem 13), and under Assumption 3, robustness holds if the optimal policies for every initial point are identical (Theorem 15).
(e) Continuity and robustness do not hold in general under total variation convergence of the kernels (Example 1).
(f) Under Assumption 4, continuity and robustness hold (subsection 4.3).

In Section 5, we study applications to empirical learning (in Section 5.1) where we establish the positive relevance of Theorem 10, and then applications to finite model approximations under the perspective of robustness in in Section 5.2. Here, we restrict the analysis to the case with weakly continuous kernels.

2 Continuity of Optimal Cost in Convergence of Models (POMDP Case)

In this section, we will study continuity of the optimal discounted cost under the convergence of transition kernels and cost functions.

2.1 Weak Convergence

Absence of Continuity under Weak Convergence. The following shows that the optimal cost may not be continuous under weak convergence of transition kernels.

Theorem 1. *[18]. Let $\mathcal{T}_n \to \mathcal{T}$ weakly, then it is not necessarily true that $J_\beta^*(c,\mathcal{T}_n) \to J_\beta^*(c,\mathcal{T})$ even when the prior distributions are the same, the measurement channel Q is continuous in total variation, and $c(x,u)$ is continuous and bounded on $\mathbb{X}\times\mathbb{U}$.*

We prove the result with a counterexample [18]. Letting $\mathbb{X}=\mathbb{U}=\mathbb{Y}=[-1,1]$ and $c(x,u)=(x-u)^2$, the observation channel is chosen to be uniformly distributed over [-1,1], $Q\sim U([-1,1])$), the initial distributions of the state variable are chosen to be same as $P\sim\delta_1$, where $\delta_x(A):=1_{\{x\in A\}}$ for Borel A, and the transition kernels are:

$$\mathcal{T}(\cdot|x,u) = \delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] \\ + (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot)$$

$$\mathcal{T}_n(\cdot|x,u) = \delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) \\ + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot).$$

It can be seen that $\mathcal{T}_n \to \mathcal{T}$ weakly according to Definition 1(i). Note that the cost function is continuous, and the measurement channel is continuous in total variation. The optimal discounted costs can be found as

$$J^*_\beta(c,\mathcal{T}) = \sum_{k=1}^{\infty} E_P^{\mathcal{T}}[\beta^k X_k^2] = \sum_{k=1}^{\infty} \beta^k = \frac{\beta}{1-\beta}$$

$$J^*_\beta(c,\mathcal{T}_n) = \sum_{k=1}^{\infty} E_P^{\mathcal{T}_n}[\beta^k X_k^2] = \beta[\frac{1}{2}(1-\frac{1}{n})^2 + \frac{1}{2}(-1+\frac{1}{n})^2].$$

Then we have $J^*_\beta(c,\mathcal{T}_n) \to \beta \neq \frac{\beta}{1-\beta}$.

2.2 A Sufficient Condition for Continuity under Weak Convergence

In the following, we will establish and utilize some regularity properties for the optimal cost with respect to the convergence of transition kernels.

Assumption 5 *(a) The stochastic kernel $\mathcal{T}(\cdot|x,u)$ is weakly continuous in (x,u), i.e. if $(x_n,u_n) \to (x,u)$, then $\mathcal{T}(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly.*

(b) The observation channel $Q(\cdot|x)$ is continuous in total variation, i.e., if $x_n \to x$, then $Q(\,\cdot\,|x_n) \to Q(\,\cdot\,|x)$ in total variation.

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

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

It is a well known result that, any POMDP can be reduced to a (completely observable) MDP, whose states are the posterior state distributions or *beliefs* of the observer; that is, the state at time t is $Z_t(\,\cdot\,) := \mathsf{Pr}\{X_t \in \,\cdot\,|Y_0,\dots,Y_t,U_0,\dots,U_{t-1}\} \in \mathcal{P}(\mathbb{X})$. We call this equivalent MDP the belief-MDP . The belief-MDP has state space $Z = \mathcal{P}(\mathbb{X})$ and action space $\mathbb{U}$. Under the topology of weak convergence, since $\mathbb{X}$ is a Borel space, Z is metrizable with the Prokhorov metric which makes Z into a Borel space [19]. The transition probability η of the belief-MDP can be constructed through non-linear filtering equations.

The one-stage cost function c of the belief-MDP is given by $\tilde{c}(z,u) := \int_{\mathbb{X}} c(x,u)z(dx)$. Under the regularity of the belief-MDP, we have that the *discounted cost optimality operator* $T : C_b(Z) \to C_b(Z)$

$$(T(f))(z) = \min_u(\tilde{c}(z,u) + \beta E[f(z_1)|z_0=z,u_0=u]) \tag{3}$$

is a contraction from $C_b(Z)$ to itself under the supremum norm. As a result, there exists a fixed point, the value function, and an optimal control policy exists. In view of this existence result, in the following we will consider optimal policies.

The following result is key to proving the main result of this section whose detailed analysis can be found in [18].

Theorem 2. *Suppose we have a uniformly bounded family of functions* $\{f_n^\gamma : \mathbb{X} \to \mathbb{R}, \gamma \in \Gamma, n > 0\}$ *such that* $\|f_n^\gamma\|_\infty < C$ *for all* $\gamma \in \Gamma$ *and for all* $n > 0$ *for some* $C < \infty$.

Further suppose we have another uniformly bounded family of functions $\{f^\gamma : \mathbb{X} \to \mathbb{R}, \gamma \in \Gamma\}$ *such that* $\|f^\gamma\|_\infty < C$ *for all* $\gamma \in \Gamma$ *for some* $C < \infty$. *Under the following assumptions,*

(i) For any $x_n \to x$

$$\sup_{\gamma\in\Gamma} \left| f_n^\gamma(x_n) - f^\gamma(x) \right| \to 0, \quad \sup_{\gamma\in\Gamma} \left| f^\gamma(x_n) - f^\gamma(x) \right| \to 0,$$

(ii) $\sup_\gamma \rho(\mu_n^\gamma, \mu^\gamma) \to 0$ *where* ρ *is some metric for the weak convergence topology,*

we have

$$\sup_{\gamma\in\Gamma} \left| \int f_n^\gamma(x)\mu_n^\gamma(dx) - \int f^\gamma(x)\mu^\gamma(dx) \right| \to 0.$$

Theorem 3. *Under Assumptions 1 and 2,*

$$\sup_{\gamma\in\Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \to 0.$$

Proof sketch.

$$\begin{aligned}
&\sup_{\gamma\in\Gamma} |J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma)| \\
&= \sup_{\gamma\in\Gamma} \left| \sum_{t=0}^{\infty} \beta^t \left(E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c\big(X_t, \gamma(Y_{[0,t]})\big)\right] \right) \right| \\
&\leq \sum_{t=0}^{\infty} \beta^t \sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c\big(X_t, \gamma(Y_{[0,t]})\big)\right] \right|.
\end{aligned}$$

Recall that an *admissible policy* γ is a sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$. In the last step above, we make a slight abuse of notation; the sup at the first step is over all sequence of control functions $\{\gamma_t,\, t \in \mathbb{Z}_+\}$ whereas the sup at the last step is over all sequence of control functions $\{\gamma_{t'},\, t' \leq t\}$, but we will use the same notation, γ, in the rest of the proof.

For any $\epsilon > 0$, we choose a $K < \infty$ such that $\sum_{t=K+1}^{\infty} \beta^k 2\|c\|_\infty \leq \epsilon/2$. For the chosen K, we choose an $N < \infty$ such that

$$\sup_{\gamma\in\Gamma} \left| E_P^{\mathcal{T}_n}\left[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\right] - E_P^{\mathcal{T}}\left[c\big(X_t, \gamma(Y_{[0,t]})\big)\right] \right| \leq \epsilon/2K$$

for all $t \leq K$ and for all $n > N$. We note that in [18] a fixed c function was considered, but by considering the additional term

$$\sup_{\gamma \in \Gamma} \Big| E_P^{\mathcal{T}_n}\Big[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\Big] - E_P^{\mathcal{T}}\Big[c_n\big(X_t, \gamma(Y_{[0,t]})\big)\Big]\Big|$$

and noting that $\sup_\gamma |\int Q(dy|x_n) c_n(x_n, \gamma(y)) - \int Q(dy|x) c(x, \gamma(y))| \to 0$, for every $x_n \to x$, by a generalized dominated convergence theorem as Q is continuous in total variation, a triangle inequality argument shows that the same result applies. This follows from a generalized dominated convergence theorem as stated in Theorem 2 whose detailed analysis can be found in [18]. Thus, $\sup_{\gamma \in \Gamma} \big| J_\beta(c_n, \mathcal{T}_n, \gamma) - J_\beta(c, \mathcal{T}, \gamma) \to 0$ as $n \to \infty$. □

Theorem 4. *Suppose the conditions of Theorem 3 hold. Then* $\lim_{n\to\infty} |J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})| = 0$.

Proof sketch. We start with the following bound:

$$\begin{aligned} &|J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})| \\ \leq &\max\Big(J_\beta(c_n, \mathcal{T}_n, \gamma^*) - J_\beta(c, \mathcal{T}, \gamma^*), J_\beta(c, \mathcal{T}, \gamma_n^*) - J_\beta(c_n, \mathcal{T}_n, \gamma_n^*)\Big), \end{aligned} \tag{4}$$

where γ^* and γ_n^* are the optimal policies, respectively, for $\mathcal{T}$ and $\mathcal{T}_n$. Both terms go to 0 by Theorem 3. □

2.3 Absence of Continuity under Setwise Convergence

We now show that continuity of optimal costs may fail under the setwise convergence of transition kernels. Theorem 12 in the next section establishes this result for fully observed models, which serves as a proof for this setup also.

Theorem 5. *Let $\mathcal{T}_n \to \mathcal{T}$ setwise. Then, it is not true in general that $J_\beta^*(c, \mathcal{T}_n) \to J_\beta^*(c, \mathcal{T})$, even when $\mathbb{X}, \mathbb{Y}$, and $\mathbb{U}$ are compact and $c(x,u)$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}$.*

2.4 Continuity under Total Variation

Theorem 6. *Under Assumption 4, $J_\beta^*(c_n, \mathcal{T}_n) \to J_\beta^*(c, \mathcal{T})$.*

Proof sketch. We start with the following bound:

$$\begin{aligned} |J_\beta^*(c_n, \mathcal{T}_n) - J_\beta^*(c, \mathcal{T})| \leq &\max\Big(J_\beta(c_n, \mathcal{T}_n, \gamma^*) - J_\beta(c, \mathcal{T}, \gamma^*), J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \\ &- J_\beta(c, \mathcal{T}, \gamma_n^*)\Big), \end{aligned}$$

where γ^* and γ_n^* are the optimal policies, respectively, for $\mathcal{T}$ and $\mathcal{T}_n$.

We now study the following:

$$\begin{aligned}
&\sup_{\gamma\in\mathit{\Gamma}} |J_\beta(c_n,\mathcal{T}_n,\gamma)-J_\beta(c,\mathcal{T},\gamma)|\\
&=\sup_{\gamma\in\mathit{\Gamma}}\left|\sum_{t=0}^{\infty}\beta^t\left(E_P^{\mathcal{T}_n}\left[c_n\big(X_t,\gamma(Y_{[0,t]})\big)\right]-E_P^{\mathcal{T}}\left[c\big(X_t,\gamma(Y_{[0,t]})\big)\right]\right)\right|\\
&\leq\sum_{t=0}^{\infty}\beta^t\sup_{\gamma\in\mathit{\Gamma}}\left|E_P^{\mathcal{T}_n}\left[c_n\big(X_t,\gamma(Y_{[0,t]})\big)\right]-E_P^{\mathcal{T}}\left[c\big(X_t,\gamma(Y_{[0,t]})\big)\right]\right|.
\end{aligned}$$

It can be shown that ([18])

$$\sup_{\gamma\in\mathit{\Gamma}}\left|E_P^{\mathcal{T}_n}\left[c_n\big(X_t,\gamma(Y_{[0,t]})\big)\right]-E_P^{\mathcal{T}}\left[c\big(X_t,\gamma(Y_{[0,t]})\big)\right]\right|\to 0. \tag{5}$$

This was shown in [18] for fixed c. The extension to varying c_n follows from a triangle inequality step with the assumption that $\mathcal{T}_n(\cdot|x,u_n)\to\mathcal{T}(\cdot|x,u)$ setwise, and $c_n(x,u_n)\to c(x,u)$ for any $u_n\to u$. Therefore, using identical steps as in the proof of Theorem 3 we have $\sup_{\gamma\in\mathit{\Gamma}}\big|J_\beta(c_n,\mathcal{T}_n,\gamma)-J_\beta(c,\mathcal{T},\gamma)\big|\to 0$. □

3 Robustness to Incorrect Models (POMDP Case)

Here, we consider the robustness problem **P2**: Suppose we design an optimal policy, γ_n^*, for a transition kernel, $\mathcal{T}_n$ and a cost function c_n, assuming they are the correct model and apply the policy to the true model whose transition kernel is $\mathcal{T}$ and whose cost function is c. We study the robustness of the sub-optimal policy γ_n^*.

3.1 Total Variation

The next theorem gives an asymptotic robustness result.

Theorem 7. *Under Assumption 4*

$$|J_\beta(c_n,\mathcal{T},\gamma_n^*)-J_\beta^*(c,\mathcal{T})|\to 0,$$

where γ_n^* is the optimal policy designed for the kernel $\mathcal{T}_n$.

Proof sketch. We write the following:

$$|J_\beta(c,\mathcal{T},\gamma_n^*)-J_\beta^*(c,\mathcal{T})|\leq|J_\beta(c,\mathcal{T},\gamma_n^*)-J_\beta^*(c_n,\mathcal{T}_n)|+|J_\beta^*(c_n,\mathcal{T}_n)-J_\beta^*(c,\mathcal{T})|.$$

Both terms can be shown to go to 0 using (5). □

3.2 Setwise Convergence

Theorem 14 in the next section establishes the lack of robustness under the setwise convergence of kernels. As we note later, a fully observed system can be viewed as a partially observed system with the measurement being the state itself, (see (6)).

3.3 Weak Convergence

Theorem 8. *Under Assumptions 1 and 2, $|J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta^*(c,\mathcal{T})| \to 0$, where γ_n^* is the optimal policy designed for the transition kernel $\mathcal{T}_n$.*

Proof sketch. We write

$$\begin{aligned}|J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta^*(c,\mathcal{T})| \leq &|J_\beta(c,\mathcal{T},\gamma_n^*) - J_\beta(c_n,\mathcal{T}_n,\gamma_n^*)| + |J_\beta(c_n,\mathcal{T}_n,\gamma_n^*) \\ &- J_\beta(\mathcal{T},\gamma^*)|.\end{aligned}$$

The first term goes to 0 by Theorem 3. For the second term we use Theorem 4. □

4 Continuity and Robustness in the Fully Observed Case

In this section, we consider the fully observed case where the controller has direct access to the state variables. We present the results for this case separately, since here we cannot utilize the regularity properties of measurement channels which allows for stronger continuity and robustness results. Under measurable selection conditions due to weak or strong (setwise) continuity of transition kernels [13, Section 3.3], for infinite horizon discounted cost problems optimal policies can be selected from those which are stationary and deterministic. Therefore we will restrict the policies to be stationary and deterministic so that $U_t = \gamma(X_t)$ for some measurable function γ. Notice also that fully observed models can be viewed as partially observed with the measurement channel thought to be

$$Q(\cdot|x) = \delta_x(\cdot), \tag{6}$$

which is only weakly continuous, thus it does not satisfy Assumption 2.

4.1 Weak Convergence

Absence of Continuity under Weak Convergence. We start with a negative result.

Theorem 9. *For $\mathcal{T}_n \to \mathcal{T}$ weakly, it is not necessarily true that $J_\beta^*(c,\mathcal{T}_n) \to J_\beta^*(c,\mathcal{T})$ even when the prior distributions are the same and $c(x,u)$ is continuous and bounded in $\mathbb{X}\times\mathbb{U}$.*

Proof. We prove the result with a counterexample, similar to the model used in the proof of Theorem 1 Letting $\mathbb{X} = [-1,1]$, $\mathbb{U} = \{-1,1\}$ and $c(x,u) = (x-u)^2$, the initial distributions are given by $P \sim \delta_1$, that is, $X_0 = 1$, and the transition kernels are

$$\begin{aligned}\mathcal{T}(\cdot|x,u) =&\delta_{-1}(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot)] \\ &+ (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot), \\ \mathcal{T}_n(\cdot|x,u) =&\delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) + \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)}(\cdot) \\ &+ \frac{1}{2}\delta_{(-1+1/n)}(\cdot)] + (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0(\cdot).\end{aligned}$$

It can be seen that $\mathcal{T}_n \to \mathcal{T}$ weakly according to Definition 1(i). Under this setup we can calculate the optimal costs as follows:

$$J_\beta^*(c,\mathcal{T}_n) = \frac{1}{n^2} + \sum_{k=2}^{\infty} \beta^k = \frac{1}{n^2} + \frac{\beta^2}{1-\beta},$$

and $J_\beta^*(c,\mathcal{T}) = 0$. Thus, continuity does not hold. □

We now present another counter example emphasizing the importance of continuous convergence in the actions. The following counter example shows that without the continuous convergence and regularity assumptions on the kernel $\mathcal{T}$, continuity fails even when $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ pointwise (for x,u) in total variation (also setwise and weakly) and even when the cost function $c(x,u)$ is continuous and bounded. Notice that this example also holds for both setwise and weak convergence.

Example 1. Assume that the kernels are given by

$$\begin{aligned} \mathcal{T}_n(\cdot|x,u) &\sim U([u^n, 1+u^n]), \\ \mathcal{T}(\cdot|x,u) &\sim \begin{cases} U([0,1]) & \text{if } u \neq 1, \\ U([1,2]) & \text{if } u = 1, \end{cases} \end{aligned}$$

where $\mathbb{U} = [0,1]$ and $\mathbb{X} = \mathbb{R}$. We note first that $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ in total variation for every fixed x and u.

The cost function is in the following form:

$$c(x,u) = \begin{cases} 2 & \text{if } x \leq \frac{1}{e}, \\ 2 - \frac{x-\frac{1}{e}}{0.1} & \text{if } \frac{1}{e} < x \leq 0.1 + \frac{1}{e}, \\ 1 & \text{if } 0.1 + \frac{1}{e} < x \leq 1 + \frac{1}{e} - 0.1, \\ 2 - \frac{1+\frac{1}{e}-x}{0.1} & \text{if } 1 + \frac{1}{e} - 0.1 < x \leq 1 + \frac{1}{e}, \\ 2 & \text{if } 1 + \frac{1}{e} < x. \end{cases}$$

Notice that $c(x,u)$ is a continuous function.

With this setup, $\gamma^*(x) = 0$ is an optimal policy for $\mathcal{T}$ since on the $[0,1]$ interval the induced cost is less than the cost induced on the $[1,2]$ interval. The cost under this policy is

$$J_\beta^*(c,\mathcal{T}) = \sum_{t=0}^{\infty} \beta^t \left(2 \times \frac{1}{e} + \frac{0.3}{2} + 0.9 - \frac{1}{e}\right) = \frac{1}{1-\beta}\left(1.05 + \frac{1}{e}\right).$$

For $\mathcal{T}_n$, $\gamma_n^*(x) = e^{-\frac{1}{n}}$ is an optimal policy for every n as $e^{-\frac{1}{n}\times n} = \frac{1}{e}$ and thus the state is distributed between $\frac{1}{e} < x \leq 1 + \frac{1}{e}$ in which interval the cost is the least. Hence, we can write

$$\begin{aligned} \lim_{n\to\infty} J_\beta(c,\mathcal{T}_n,\gamma_n^*) &= \sum_{t=0}^{\infty} \beta^t \left(0.3 + 1 - 0.2\right) = \frac{1.1}{1-\beta} \neq \frac{1}{1-\beta}\left(1.05 + \frac{1}{e}\right) \\ &= J_\beta^*(c,\mathcal{T}). \end{aligned}$$

◇

A Sufficient Condition for Continuity under Weak Convergence. We will now establish that if the kernels and the model components have some further regularity, continuity does hold. The assumptions of the following result are the same as the assumptions for the partially observed case (Theorem 4) except for the assumption on the measurement channel Q.

Theorem 10. *Under Assumption 1, $J_\beta(c_n, \mathcal{T}_n, \gamma_n^*) \to J_\beta(c, \mathcal{T}, \gamma^*)$ for any initial state x_0, as $n \to \infty$.*

Proof. We will use the successive approximations for an inductive argument.

Recall *discounted cost optimality operator* $T : C_b(Z) \to C_b(Z)$ from (3)

$$(T(v))(x) = \inf_{u \in \mathbb{U}} \left(c(x, u) + \beta E[v(x_1)|x_0 = x, u_0 = u] \right),$$

which is a contraction from $C_b(\mathbb{X})$ to itself under the supremum norm and has a fixed point, the value function. For the kernel $\mathcal{T}$, we will denote the approximation functions by

$$v^k(x) = T(v^{k-1})(x),$$

and for the kernel $\mathcal{T}_n$ we will use $v_n^k(x)$ to denote the approximation functions, notice that the operator T also depends on n for the model $\mathcal{T}_n$, but we will continue using it as T in what follows.

We wish to show that the approximation functions for $\mathcal{T}_n$ continuously converge to the ones for $\mathcal{T}$. Then, for the first step of the induction we have

$$v^1(x) = c(x, u^*), \quad v_n^1(x_n) = c_n(x_n, u_n^*),$$

and thus we can write,

$$|v^1(x) - v_n^1(x_n)| \leq \sup_{u \in \mathbb{U}} \big| c(x, u) - c_n(x_n, u) \big|$$

since $c_n(x_n, u_n) \to c(x, u)$ for all $(x_n, u_n) \to (x, u)$ and the action space, $\mathbb{U}$, is compact, the first step of the induction holds, i.e. $\lim_{n\to\infty} |v^1(x) - v_n^1(x_n)| = 0$.

For the k^{th} step we have

$$v^k(x) = T(v^{k-1})(x) = \inf_u \big[c(x, u) + \beta \int_{\mathbb{X}} v^{k-1}(x^1) \mathcal{T}(dx^1|x, u) \big],$$
$$v_n^k(x_n) = T(v_n^{k-1})(x_n) = \inf_u \big[c_n(x_n, u) + \beta \int_{\mathbb{X}} v_n^{k-1}(x^1) \mathcal{T}_n(dx^1|x_n, u) \big].$$

Note that the assumptions of the theorem satisfy the measurable selection criteria and hence we can choose minimizing selectors ([13, Section 3.3]). If we denote the selectors by u^* and u_n^*, we can write

$$\begin{aligned}
&|v^k(x) - v_n^k(x_n)| \\
\leq &\max\Bigg(\bigg[|c(x,u^*) - c_n(x_n,u^*)| \\
&\quad + \beta|\int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u^*)|\bigg], \\
&\quad \bigg[|c(x,u_n^*) - c_n(x_n,u_n^*)| \\
&\quad + \beta|\int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u_n^*) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u_n^*)|\bigg]\Bigg).
\end{aligned}$$

Hence, we can write

$$\begin{aligned}
&|v^k(x) - v_n^k(x_n)| \qquad (7)\\
\leq &\sup_{u\in\mathbb{U}}\bigg[|c(x,u) - c_n(x_n,u)| \\
&\quad + \beta|\int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u)|\bigg],
\end{aligned}$$

above, the first term goes to 0 as $c_n(x_n,u_n) \to c(x,u)$ for all $(x_n,u_n) \to (x,u)$ and the action space, $\mathbb{U}$, is compact. For the second term we write,

$$\begin{aligned}
&\sup_{u\in\mathbb{U}}|\int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v_n^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u)| \\
\leq &\sup_{u\in\mathbb{U}}|\int_{\mathbb{X}} \big(v^{k-1}(x^1) - v_n^{k-1}(x^1)\big)\mathcal{T}_n(dx^1|x_n,u)| \\
&+ \sup_{u\in\mathbb{U}}|\int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}(dx^1|x,u) - \int_{\mathbb{X}} v^{k-1}(x^1)\mathcal{T}_n(dx^1|x_n,u)|
\end{aligned}$$

above, for the first term, by the induction argument for any $x_n^1 \to x^1$, $\big|v^{k-1}(x^1) - v_n^{k-1}(x_n^1)\big| \to 0$ (i.e., we have continuous convergence). We also have that $\mathcal{T}_n(\cdot|x_n,u) \to \mathcal{T}(\cdot|x,u)$ weakly uniformly over $u \in \mathbb{U}$ as $\mathbb{U}$ is compact. Therefore, using Theorem 2 the first term goes to 0. For the second term we again use that $\mathcal{T}_n(\cdot|x_n,u)$ converges weakly to $\mathcal{T}(\cdot|x,u)$ uniformly over $u \in \mathbb{U}$. With an almost identical induction argument it can also be shown that $v^{k-1}(x^1)$ is continuous in x^1, thus the second term also goes to 0.

So far, we have showed that for any $k \in \mathbb{N}$, $\lim_{n\to\infty}\big|v_n^k(x_n) - v^k(x)\big| = 0$ for any $x_n \to x$, in particular it is also true that $\lim_{n\to\infty}\big|v_n^k(x) - v^k(x)\big| = 0$ for any x.

As we have stated earlier, it can be shown that the approximation operator, T is a contractive operator under supremum norm with modulus β and it converges to a fixed point which is the value function. Thus, we have

$$\left|J_\beta(c,\mathcal{T},\gamma^*) - v^k(x)\right| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}, \quad \left|J^*_\beta(c_n,\mathcal{T}_n,\gamma^*_n) - v^k_n(x)\right| \leq \|c\|_\infty \frac{\beta^k}{1-\beta}. \tag{8}$$

Combining the results,

$$\begin{aligned}|J_\beta(c_n,\mathcal{T}_n,\gamma^*_n) - |J_\beta(c,\mathcal{T},\gamma^*)| \leq &|J_\beta(c_n,\mathcal{T}_n,\gamma^*_n) - v^k_n(x)| + |v^k_n(x) - v^k(x)| \\ &+ |J_\beta(c,\mathcal{T},\gamma^*) - v^k(x)|.\end{aligned}$$

Note that the first and the last term can be made arbitrarily small since (8) holds for all $k \in \mathbb{N}$; the second term goes to 0 with an inductive argument for all $k \in \mathbb{N}$. □

A Sufficient Condition for Robustness under Weak Convergence. We now present a result that establishes robustness if the optimal policies for every initial point are identical. That is, for every n, γ^*_n is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$). A sufficient condition for this property is that γ^*_n solves the discounted cost optimality equation (DCOE) for every initial point.

A policy $\gamma^* \in \varGamma$ solves the discounted cost optimality equation and is optimal if it satisfies

$$J^*_\beta(c,\mathcal{T},x) = c(x,\gamma^*(x)) + \beta \int J^*_\beta(c,\mathcal{T},x_1)\mathcal{T}(dx_1|x,\gamma^*(x)).$$

Thus, a policy is optimal for every initial point if it satisfies the DCOE for all initial points $x \in \mathbb{X}$. The following generalizes [18].

Theorem 11. *Under Assumption 1, $J_\beta(c,\mathcal{T},\gamma^*_n) \to J_\beta(c,\mathcal{T},\gamma^*)$ for any initial point x_0 if γ^*_n is optimal for any initial point for the kernel $\mathcal{T}_n$ and for the stage-wise cost function c_n.*

Remark 1. For the partially observed case, the proof approach we use makes use of policy exchange (e.g. (4)) and for this approach the total variation continuity of channel $Q(\cdot|x)$ is a key step to deal with the uniform convergence over policies. As we stated before, the channel for fully observed models can be considered in the form of (6) which is only weakly continuous and not continuous in total variation. Thus, obtaining a result uniformly over all policies may not be possible. However, for the fully observed models we can reach continuity and robustness (Theorem 10, Theorem 11) using a value iteration approach. With this approach, instead of exchanging policies and analyzing uniform convergence over all policies, we can exchange control actions (e.g. (7)) and analyze uniform convergence over the action space $\mathbb{U}$ by using the discounted optimality operator (3). Hence, we are only able to show convergence over optimal policies for the fully observed case, i.e. $J_\beta(c_n,\mathcal{T}_n,\gamma^*_n) \to J_\beta(c,\mathcal{T},\gamma^*)$ or $J_\beta(c,\mathcal{T},\gamma^*_n) \to J_\beta(c,\mathcal{T},\gamma^*)$ where γ^*_n and γ^* are optimal policies, whereas, for partially observed models, regularity of the channel allows us to show convergence over any sequence of policies, i.e. $\sup_{\gamma\in\varGamma} |J_\beta(c_n,\mathcal{T}_n,\gamma) - J_\beta(c,\mathcal{T},\gamma)| \to 0$.

Remark 2. As we have discussed in subsection 2.2, a partially observed model can be reduced to a fully observed process where the state process (beliefs) becomes probability measure valued. Consider the partially observed models with transition kernels $\mathcal{T}_n$ and $\mathcal{T}$ (with a channel Q) and their corresponding fully observed transition kernels η_n and η: following the discussions and techniques in [9] and [15], one can show that η_n and η satisfy the conditions of Theorem 11 and Theorem 10 that is $\eta_n(\cdot|z_n, u_n) \to \eta(\cdot|z, u)$ for any $(z_n, u_n) \to (z, u)$ under the following set of assumptions

- $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ for any $(x_n, u_n) \to (x, u)$,
- $Q(\cdot|x)$ is continuous on total variation in x.

We remark that these conditions also agree with the conditions presented for continuity and robustness of the partially observed models (Theorem 4 and Theorem 8).

4.2 Setwise Convergence

Absence of Continuity under Setwise Convergence. We give a negative result similar to Theorem 5, via Example 1:

Theorem 12. *Letting $\mathcal{T}_n \to \mathcal{T}$ setwise, then it is not necessarily true that $J^*_\beta(c, \mathcal{T}_n) \to J^*_\beta(c, \mathcal{T})$ even when $c(x, u)$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}$.*

A Sufficient Condition for Continuity under Setwise Convergence.

Theorem 13. *Under Assumption 3 $J_\beta(c_n, \mathcal{T}_n, \gamma^*_n) \to J_\beta(c, \mathcal{T}, \gamma^*)$, for any initial state x_0, as $n \to \infty$.*

Proof. We use the same value iteration technique that we used to prove Theorem 10. See [18]. □

Absence of Robustness under Setwise Convergence. Now, we give a result showing that even if the continuity holds under the setwise convergence of the kernels, the robustness may not be satisfied (see [18, Theorem 4.7]).

Theorem 14. *Supposing $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x, u)$ setwise for every $x \in \mathbb{X}$ and $u \in \mathbb{U}$ and $(x_n, u_n) \to (x, u)$, then it is not true in general that $J_\beta(c, \mathcal{T}, \gamma^*_n) \to J_\beta(c, \mathcal{T}, \gamma^*)$, even when $\mathbb{X}$ and $\mathbb{U}$ are compact and $c(x, u)$ is continuous and bounded in $\mathbb{X} \times \mathbb{U}$.*

A Sufficient Condition for Robustness under Setwise Convergence. We now present a similar result to Theorem 11 that is we show that under the conditions of Theorem 13, if further for every n, γ^*_n is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$) then robustness holds under setwise convergence.

Theorem 15. *Supposing Assumption 3 holds, if further we have that for every n, γ^*_n is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$) then $J_\beta(c, \mathcal{T}, \gamma^*_n) \to J_\beta(c, \mathcal{T}, \gamma^*)$.*

4.3 Total Variation

The continuity result in Theorem 6 and the robustness result in Theorem 7 apply to this case since the fully observed model may be viewed as a partially observed model with the measurement channel Q given in (6).

5 Applications to Data-Driven Learning and Finite Model Approximations

5.1 Application of Robustness Results to Data-Driven Learning

In practice, one may estimate the kernel of a controlled Markov chain using empirical data; see e.g. [3, 12] for some related literature in the control-free and controlled contexts.

Let us briefly review the basic case where an i.i.d. sequence of random variables is repeatedly observed, but its probability measure is not known apriori. Let $\{(X_i),\ i \in \mathbb{N}\}$ be an $\mathbb{X}$-valued i.i.d. random variable sequence generated according to some distribution μ. Defining for every (fixed) Borel $B \subset \mathbb{X}$, and $n \in \mathbb{N}$, the empirical occupation measures $\mu_n(B) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in B\}}$, one has $\mu_n(B) \to \mu(B)$ almost surely by the strong law of large numbers. It then follows that $\mu_n \to \mu$ weakly with probability one ([6], Theorem 11.4.1). However, μ_n does not converge to μ in total variation or setwise, in general. On the other hand, if we know that μ admits a density, we can find estimators to estimate μ under total variation [5, Chapter 3]. For a more detailed discussion, see [17, p. 1950-1951]. In the previous sections, we established robustness results under the convergence of transition kernels in the topology of weak convergence and total variation. We build on these observations.

Corollary 1 (to Theorem 6 and Theorem 7). *Suppose we are given the following dynamics for finite state space, $\mathbb{X}$, and finite action space, $\mathbb{U}$,*

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t)$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown. Suppose that there is an initial training period so that under some policy, every x, u pair is visited infinitely often if training were to continue indefinitely, but that the training ends at some finite time. Let us assume that, through this training, we empirically learn the transition dynamics such that for every (fixed) Borel $B \subset \mathbb{X}$, for every $x \in \mathbb{X}$, $u \in \mathbb{U}$ and $n \in \mathbb{N}$, the empirical occupation measures are

$$\mathcal{T}_n(B|x_0 = x, u_0 = u) = \frac{\sum_{i=1}^{n} 1_{\{X_i \in B, X_{i-1}=x, U_{i-1}=u\}}}{\sum_{i=1}^{n} 1_{\{X_{i-1}=x, U_{i-1}=u\}}}.$$

Then we have that $J_\beta^*(\mathcal{T}_n) \to J_\beta^*(\mathcal{T})$ and $J_\beta(\mathcal{T}, \gamma_n^*) \to J_\beta^*(\mathcal{T})$, where γ_n^* is the optimal policy designed for $\mathcal{T}_n$. Since the channel model g has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t, v_t) = x_t$.

Proof. We have that $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ weakly for every $x \in \mathbb{X}$, $u \in \mathbb{U}$ almost surely by law of large numbers. Since the spaces are finite, we also have $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ under total variation. By Theorem 6 and Theorem 7, the results follow. □

The following holds for more general spaces.

Corollary 2 (to Theorems 8, 4, 10 and 11). *Suppose we are given the following dynamics with state space $\mathbb{X}$ and action space $\mathbb{U}$,*

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t),$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown. Suppose that $f(x,u,\cdot) : \mathbb{W} \to \mathbb{X}$ is invertible for all fixed (x,u) and $f(x,u,w)$ is continuous and bounded on $\mathbb{X} \times \mathbb{U} \times \mathbb{W}$. We construct the empirical measures for the noise process w_t such that for every (fixed) Borel $B \subset \mathbb{W}$, and for every $n \in \mathbb{N}$, the empirical occupation measures are

$$\mu_n(B) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{f^{-1}_{x_{i-1},u_{i-1}}(x_i) \in B\}} \tag{9}$$

where $f^{-1}_{x_{i-1},u_{i-1}}(x_i)$ denotes the inverse of $f(x_{i-1},u_{i-1},w) : \mathbb{W} \to \mathbb{X}$ for given (x_{i-1},u_{i-1}). Using the noise measurements, we construct the empirical transition kernel estimates for any (x_0,u_0) and Borel B as

$$\mathcal{T}_n(B|x_0,u_0) = \mu_n(f^{-1}_{x_0,u_0}(B)).$$

(i) If the measurement channel (represented by the function g) is continuous in total variation then $J^*_\beta(\mathcal{T}_n) \to J^*_\beta(\mathcal{T})$ and $J_\beta(\mathcal{T},\gamma^*_n) \to J^*_\beta(\mathcal{T})$, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$ for all initial points.

(ii) If the measurement channel is in the form $g(x_t,v_t) = x_t$ (i.e. fully observed) then $J^*_\beta(\mathcal{T}_n) \to J^*_\beta(\mathcal{T})$ and if further for every n, γ^*_n is optimal for every $x_0 \in \mathbb{X}$ (under the model $\mathcal{T}_n$) then $J_\beta(\mathcal{T},\gamma^*_n) \to J^*_\beta(\mathcal{T})$.

Proof. We have $\mu_n \to \mu$ weakly with probability one where μ is the model. We claim that the transition kernels are such that $\mathcal{T}_n(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n) \to (x,u)$. To see that observe the following for $h \in C_b(\mathbb{X})$

$$\begin{aligned}&\int h(x_1)\mathcal{T}_n(dx_1|x_n,u_n) - \int h(x_1)\mathcal{T}(dx_1|x,u)\\ &= \int h(f(x_n,u_n,w))\mu_n(dw) - \int h(f(x,u,w))\mu(dw) \to 0,\end{aligned}$$

where μ_n is the empirical measure for w_t and μ is the true measure again. For the last step, we used that $\mu_n \to \mu$ weakly and $h(f(x_n,u_n,w))$ continuously converge to $h(f(x,u,w))$ i.e. $h(f(x_n,u_n,w_n)) \to h(f(x,u,w)$ for some $w_n \to w$ since f and g are continuous functions. Similarly, it can be also shown that $\mathcal{T}_n(\cdot|x,u)$ and $\mathcal{T}(\cdot|x,u)$ are weakly continuous on (x,u). Thus, for the case where the channel is continuous in total variation by Theorem 8 and Theorem 4 if $c(x,u)$ is bounded and $\mathbb{U}$ is compact the result follows.

For the fully observed case, $J^*_\beta(\mathcal{T}_n) \to J^*_\beta(\mathcal{T})$ by Theorem 10 and $J_\beta(\mathcal{T}, \gamma^*_n) \to J^*_\beta(\mathcal{T})$ by Theorem 11. □

Remark 3. We note here that the moment estimation method can also lead to consistency. Suppose that the distribution of W is determined by its moments, such that estimate models W_n have moments of all orders and $\lim_n = E[W_n^r] = E[W^r]$ for all $r \in \mathbb{Z}_+$. Then, we have that [4, Thm 30.2] $W_n \to W$ weakly and thus $\mathcal{T}_n(\cdot|x_n, u_n) \to \mathcal{T}(\cdot|x,u)$ weakly for any $(x_n, u_n) \to (x,u)$ under the assumptions of above corollary. Hence, we reach continuity and robustness using the same arguments as in the previous result (Corollary 2).

Now, we give a similar result with the assumption that the noise process of the dynamics admits a continuous probability density function.

Corollary 3 (to Theorem 6 and Theorem 7). *Suppose we are given the following dynamics for real vector state space $\mathbb{X}$ and action space $\mathbb{U}$*

$$x_{t+1} = f(x_t, u_t, w_t), \qquad y_t = g(x_t, v_t),$$

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown but it is known that the noise w_t admits a continuous probability density function. Suppose that $f(x,u,\cdot) : \mathbb{W} \to \mathbb{X}$ is invertible for all (x,u). We collect i.i.d. samples of $\{w_t\}$ as in (9) and use them to construct an estimator, $\tilde{\mu}_n$, as described in [5] which consistently estimates μ in total variation. Using these empirical estimates, we construct the empirical transition kernel estimates for any (x_0,u_0) and Borel B as

$$\mathcal{T}_n(B|x_0,u_0) = \tilde{\mu}_n(f^{-1}_{x_0,u_0}(B)).$$

Then independent of the channel, $J^*_\beta(\mathcal{T}_n) \to J^*_\beta(\mathcal{T})$ and $J_\beta(\mathcal{T},\gamma^*_n) \to J^*_\beta(\mathcal{T})$, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$. Since the channel model g has no restrictions, this result also applies to the fully observed model setup by taking $g(x_t, v_t) = x_t$.

Proof. By [5] we can estimate μ in total variation so that almost surely $\lim_{n\to\infty} \|\tilde{\mu}_n - \mu\|_{TV} = 0$. We claim that the convergence of $\tilde{\mu}_n$ to μ under total variation metric implies the convergence of $\mathcal{T}_n$ to $\mathcal{T}$ in total variation uniformly over all $x \in \mathbb{X}$ and $u \in \mathbb{U}$ i.e. $\lim_{n\to\infty} \sup_{x,u} \|\mathcal{T}_n(\cdot|x,u) - \mathcal{T}(\cdot|x,u)\|_{TV} = 0$. Observe the following:

$$\begin{aligned}
&\sup_{x,u} \|\mathcal{T}_n(\cdot|x,u) - \mathcal{T}(\cdot|x,u)\|_{TV} \\
=&\sup_{x,u} \sup_{||h||_\infty \le 1} \Big| \int h(x_1)\mathcal{T}_n(dx_1|x,u) - \int h(x_1)\mathcal{T}(dx_1|x,u)\Big| \\
=&\sup_{x,u} \sup_{||h||_\infty \le 1} \Big| \int h(f(x,u,w))\tilde{\mu}_n(dw) - \int h(f(x,u,w))\mu(dw)\Big| \\
\le&\|\tilde{\mu}_n - \mu\|_{TV} \to 0.
\end{aligned}$$

Thus, by Theorem 6 and Theorem 7, the result follows. □

The following example presents some system and channel models which satisfy the requirements of the above corollaries.

Example 2. Let $\mathbb{X}, \mathbb{Y}$, $\mathbb{U}$ be real vector spaces with

$$x_{t+1} = f(x_t, u_t) + w_t, \qquad y_t = h(x_t, v_t)$$

for unknown i.i.d. noise processes $\{w_t\}$ and $\{v_t\}$.

1. Suppose the channel is in the following form; $y_t = h(x_t, v_t) = x_t + v_t$ where v_t admits a density (e.g. Gaussian density). It can be shown by an application of Scheffé's theorem that the channels in this form are continuous in total variation. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 2 hold for partially observed models.
2. If the channel is in the following form; $x_t = h(x_t, v_t)$ then the system is fully observed. If further $f(x_t, u_t)$ is continuous and bounded then the requirements of Corollary 2 holds for fully observed models.
3. Suppose the function $f(x_t, u_t)$ is known, if the noise process w_t admits a continuous density, then one can estimate the noise model in total variation in a consistent way (see [5]). Hence, the conditions of Corollary 3 holds independent of the channel model.

◇

5.2 Application to Approximations of MDPs and POMDPs with Weakly Continuous Kernels

We now discuss **Problem P4**, that is whether approximation of an MDP model with a standard Borel space with a finite MDPs can be viewed an instance of robustness problem to incorrect models and whether our results can be applied.

Review of Finitely Quantized Approximations to Standard Borel MDPs. Consider an MDP which is quantized as follows.

Finite State Approximate MDP: Quantization of the State Space. *Let $d_{\mathbb{X}}$ denote the metric on $\mathbb{X}$. For each $n \geq 1$, there exists a finite subset $\{x_{n,i}\}_{i=1}^{k_n}$ of $\mathbb{X}$ such that*

$$\min_{i \in \{1,\ldots,k_n\}} d_{\mathbb{X}}(x, x_{n,i}) < 1/n \textit{ for all } x \in \mathbb{X}.$$

Let $\mathbb{X}_n := \{x_{n,1}, \ldots, x_{n,k_n}\}$ and define Q_n mapping any $x \in \mathbb{X}$ to the nearest element of $\mathbb{X}_n$, i.e.,

$$Q_n(x) := \arg\min_{x_{n,i} \in \mathbb{X}_n} d_{\mathbb{X}}(x, x_{n,i}).$$

For each n, a partition $\{\mathbb{S}_{n,i}\}_{i=1}^{k_n}$ of the state space $\mathbb{X}$ is induced by Q_n by setting

$$\mathbb{S}_{n,i} = \{x \in \mathbb{X} : Q_n(x) = x_{n,i}\}.$$

Let ψ be a probability measure on $\mathbb{X}$ which satisfies

$$\psi(\mathbb{S}_{n,i}) > 0 \text{ for all } i, n,$$

and define probability measures $\psi_{n,i}$ on $\mathbb{S}_{n,i}$ by restricting ψ to $\mathbb{S}_{n,i}$:

$$\psi_{n,i}(\,\cdot\,) := \psi(\,\cdot\,)/\psi(\mathbb{S}_{n,i}).$$

Using $\{\psi_{n,i}\}$, we define a sequence of finite-state MDPs, denoted as f-MDP$_m$, to approximate the compact-state MDP.

For each m, f-MDP$_m$ is defined as: $\big(\mathbb{X}_n, \mathbb{U}, \mathcal{T}_n, c_n\big)$, and the one-stage cost function $c_n : \mathbb{X}_n \times \mathbb{U} \to [0, \infty)$ and the transition probability $\mathcal{T}_n$ on $\mathbb{X}_n$ given $\mathbb{X}_n \times \mathbb{U}$ are given by

$$c_n(x_{n,i}, a) := \int_{\mathbb{S}_{n,i}} c(x, a)\psi_{n,i}(dx)$$
$$\mathcal{T}_n(\,\cdot\,|x_{n,i}, a) := \int_{\mathbb{S}_{n,i}} Q_n * \mathcal{T}(\,\cdot\,|x, a)\psi_{n,i}(dx),$$

where $Q_n * \mathcal{T}(\,\cdot\,|x, a) \in \mathcal{P}(\mathbb{X}_n)$ is the pushforward of the measure $\mathcal{T}(\,\cdot\,|x, a)$ with respect to Q_n; that is,

$$Q_n * \mathcal{T}(z_{n,j}|x, a) = \mathcal{T}\big(\{y \in \mathbb{X} : Q_n(y) = x_{n,j}\}|x, a\big),$$

for all $x_{n,j} \in \mathbb{X}_n$.

Finite Action Approximate MDP: Quantization of the Action Space. *Let $d_{\mathbb{U}}$ denote the metric on $\mathbb{U}$. Since the action space $\mathbb{U}$ is compact and thus totally bounded, one can find a sequence of finite sets $\Lambda_n = \{a_{n,1}, \ldots, a_{n,k_n}\} \subset \mathbb{U}$ such that for all n,*

$$\min_{i \in \{1,\ldots,k_n\}} d_{\mathbb{U}}(a, a_{n,i}) < 1/n \text{ for all } a \in \mathbb{U}.$$

In other words, Λ_n is a $1/n$-net in $\mathbb{U}$. Let us assume that the sequence $\{\Lambda_n\}_{n\geq 1}$ is fixed. To ease the notation in the sequel, let us define the mapping Υ_n

$$\Upsilon_n(f)(x) := \operatorname*{arg\,min}_{a \in \Lambda_n} d_{\mathbb{U}}(f(x), a), \tag{10}$$

where ties are broken so that $\Upsilon_n(f)(x)$ is measurable.

It is known that finite quantization policies are nearly optimal under the conditions to be presented below, see [23, Theorem 3.2]. Thus, to make the presentation shorter, we will either assume that the action set is finite, or it has been approximated by a finite action space through the construction above. Assuming finite action sets will help us avoid measurability issues (see universal measurability discussions in [22]) as well as issues with existence of optimal policies.

Assumption 6 *(a) The one stage cost function c is nonnegative and continuous.*
(b) The stochastic kernel $\mathcal{T}(\,\cdot\,|x,a)$ is weakly continuous in $(x,a) \in \mathbb{X} \times \mathbb{U}$.
(c) $\mathbb{U}$ is finite.
(d) $\mathbb{X}$ is compact.

We note that condition (d) in Assumption 6 as presented in [22] was more general, but we have used the simpler version here for clarity in exposition.

One can write the following fixed point equation for the finite MDP

$$J_\beta^n(x) = \min_{a\in\mathbb{U}} \left\{ c_n(x,a) + \beta \sum_{x_1 \in \mathbb{X}_n} J_\beta^n(x_1)\mathcal{T}_n(x_1|x,a) \right\}$$

where $\mathcal{T}_n$ is the transition model for the finite MDP and c_n is the cost function defined on the finite model. Since the acton space is finite, we can find an optimal policy, say f_n^* for this fixed point equation. One can also simply extend J_β^n and f_n^*, which are defined on $\mathbb{X}_n$ to the entire state space $\mathbb{X}$ by taking them constant over the quantization bins $\mathbb{S}_{n,i}$. If we call the extended versions $\hat{J}_\beta^n$ and $\hat{f}_n$, the following result can be established:

Theorem 16. *[22, Theorem 2.2 and 4.1] Suppose Assumption 6 holds. Then, for any $\beta \in (0,1)$ the discounted cost of the deterministic stationary policy $\hat{f}_n$, obtained by extending the discounted optimal policy f_n^* of f-MDP$_m$ to $\mathbb{X}$ (i.e., $\hat{f}_n = f_n^* \circ Q_n$), converges to the discounted value function J^* of the compact-state MDP:*

$$\lim_{n\to\infty} \|\hat{J}_\beta^n(\cdot\,) - J_\beta^*(\cdot\,)\| = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \|J_\beta(\hat{f}_n,\,\cdot\,) - J_\beta^*\| = 0. \tag{11}$$

Theorems 16 shows that under Assumption 6, an optimal solution can be approximated via the solutions of finite models. We now show that the above approximation scheme can be viewed in relation to our robustness results.

Proof sketch of Theorem 16 via results from Section 4. With the introduced setup, one can see that the extended value function and optimal policy for the finite model satisfy the following:

$$\hat{J}_\beta^n(x) = \min_{a\in\mathbb{U}} \left\{ \hat{c}_n(x,u) + \beta \int \hat{J}_\beta^n(x_1)\hat{\mathcal{T}}_n(dx_1|x,u) \right\}$$

where $\hat{c}_n$ is the extended version of c_n to the state space $\mathbb{X}$ by making it constant over the quantization bins $\{\mathbb{S}_{n,i}\}_i$ and $\hat{\mathcal{T}}_n$ is such that for any function f

$$\int f(x_1)\hat{\mathcal{T}}_n(dx_1|x,u) := \int_{x_1\in\mathbb{X}} \int_{z\in\mathbb{S}_{n,i}} f(x_1)\mathcal{T}(dx_1|z,u)\psi_{n,i}(dz)$$

where $\mathbb{S}_{n,i}$ is the quantization bin that x belongs to.

With this setup, one can see that for any $x_n \to x$ we have $\hat{c}_n(x_n, u) \to c(x, u)$ and for any continuous and bounded f

$$\int f(x_1)\hat{\mathcal{T}}_n(dx_1|x_n, u) := \int_{x_1 \in \mathbb{X}} \int_{z \in \mathbb{S}_{n,i}} f(x_1)\mathcal{T}(dx_1|z, u)\psi_{n,i}(dz)$$
$$\to \int f(x_1)\mathcal{T}(dx_1|x, u).$$

Hence, Assumption 1 holds under Assumption 6, and we can conclude the proof using Theorem 11 and Theorem 10. □

6 Concluding Remarks

We studied regularity properties of optimal stochastic control on the space of transition kernels, and applications to robustness of optimal control policies designed for an incorrect model applied to an actual system. We also presented applications to data-driven learning and related the robustness problem to finite MDP approximation techniques. For the problems presented in this article, our focus was on infinite horizon discounted cost setup. However, we note that the results can be extended to the infinite horizon average cost setup under various forms of ergodicity properties on the state process.

References

1. Backhoff-Veraguas, J., Bartl, D., Beiglböck, M., Eder, M.: Adapted wasserstein distances and stability in mathematical finance. Finance and Stochastics **24**, 3601–632 (2020)
2. Bayraktar, E., Dolinsky, Y., Guo, J.: Continuity of utility maximization under weak convergence. Mathematics and Financial Economics **14(4)**, 1–33 (2020)
3. Billingsley, P.: Statistical methods in Markov chains. The Annals of Mathematical Statistics, **32**, 12–40 (1961)
4. Billingsley, P.: Probability and Measure. Wiley, 3rd ed., New York (1995)
5. Devroye, L., Györfi, L.: Non-parametric Density Estimation: The L_1 View. John Wiley, New York, (1985)
6. Dudley, R.M.: Real Analysis and Probability, 2nd ed. Cambridge University Press, Cambridge (2002)
7. Dupuis, P., James, M.R., Petersen, I.: Robust properties of risk-sensitive control. Mathematics of Control, Signals and Systems **13(4)**, 318–332 (2000).
8. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming **171(1)**, 1–52 (2018)
9. Feinberg, E., Kasyanov, P., Zgurovsky, M.: Partially observable total-cost Markov decision process with weakly continuous transition probabilities. Math. Oper. Res. **41(2)**, 656–681 (2016)
10. Ghosh, J. K., Ramamoorthi, R.V.: Bayesian Nonparametrics. Springer, New York (2003)
11. Gray, R.M.: Entropy and Information Theory. Springer-Verlag, New York (1990)

12. Györfi, L., Kohler, M.: Nonparametric estimation of conditional distributions. IEEE Trans. Information Theory **53(5)**, 1872–1879 (2007)
13. Hernandez-Lerma, O., and Lasserre, J.: Discrete-Time Markov Control Processes. Springer, New York (1996)
14. Jacobson, D.: Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. IEEE Trans. on Automatic control **18(2)**, 124–131 (1973)
15. Kara, A.D., Saldi, N., Yüksel, S.: Weak Feller property of non-linear filters. Systems & Control Letters **134**, 104–512 (2019)
16. Kara, A. D., Yüksel, S.: Robustness to incorrect system models in stochastic control and application to data-driven learning. 2018 IEEE Conf. on Decision and Control (CDC), 2753–2758 (2018)
17. Kara, A.D., Yüksel, S.: Robustness to incorrect priors in partially observed stochastic control. SIAM J. Control Optim. **57(3)**, 1929–1964 (2019)
18. Kara, A.D., Yüksel, S.: Robustness to incorrect system models in stochastic control. SIAM J. Control Optim. **58**(2), 1144–1182 (2020)
19. Parthasarathy, K.: Probability Measures on Metric Spaces. AMS, Providence, Rhode Island (2005)
20. Petersen, I., James, M.R., Dupuis, P.: Minimax optimal control of stochastic uncertain systems with relative entropy constraints. IEEE Trans. on Automatic Control **45(3)**, 398–412 (2000)
21. Pra, P.D., Meneghini, L., Runggaldier, W.J.: Connections between stochastic control and dynamic games. Mathematics of Control, Signals and Systems **9(4)**, 303–326 (1996)
22. Saldi, N., Yüksel, S., Linder, T.: On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. Math. Oper. Res. **42(4)**, 945-978 (2017)
23. Saldi, N., Yüksel, S., Linder, T.: Near optimality of quantized policies in stochastic control under weak continuity conditions. J. Math. Anal. Appl. **435(1)**, 321–337 (2015)
24. Savkin, A.V., Petersen, I.R.: Robust control of uncertain systems with structured uncertainty. J. of Mathematical Systems, Estimation, and Control **6(3)**, 1–14 (1996)
25. Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Math. Oper. Res **41(2)**, 377–401 (2016)
26. Ugrinovskii, V.A:. Robust H-infinity control in the presence of stochastic uncertainty. Int. J. of Control **71(2)**, 219–237 (1998)