Robustness to Incorrect System Models in Stochastic Control and Application to Data-Driven Learning

Ali Devran Kara and Serdar Yüksel

***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. Toward this end, we study continuity properties of discrete-time stochastic control problems with respect to system models (i.e., controlled transition kernels) 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 and robustness cannot be established the under weak convergence of transition kernels in general, but that the expected induced cost is robust under total variation in that it is continuous in the mismatch of transition kernels under convergence in total variation. By imposing further assumptions on the measurement models and on the kernel itself, we show that the optimal cost can be made continuous under weak convergence of transition kernels as well. Using these continuity properties, 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 a true model, thus establishing positive and negative results on robustness. Compared to the existing literature, we obtain refined robustness results that are applicable even when the incorrect models can be investigated under weak convergence and setwise convergence criteria (with respect to a true model), in addition to the total variation criteria. These lead to practically important results on empirical learning in (data-driven) stochastic control since often, in many applications, system models are learned through training data.**

I. Introduction

A. Preliminaries

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. In many 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. With this goal, in this paper we study continuity properties of discrete-time stochastic control problems with respect to system models (i.e., controlled transition kernels) and robustness of optimal control policies designed for incorrect models applied to the true system.

We start with defining the probabilistic model of the problem. $\mathbb{X} \subset \mathbb{R}^m$ denotes the state space of a partially observed controlled Markov process. We let $\mathbb{X}$ to be a Borel set. Thus, the state elements of the model, $\{X_t,\, t \in \mathbb{Z}_+\}$, live in $\mathbb{X}$. Here and throughout the paper $\mathbb{Z}_+$ denotes the set of non-negative integers and $\mathbb{N}$ denotes the set of positive integers. $\mathbb{Y} \subset \mathbb{R}^n$ is also a Borel set denoting the observation space of the model. The state is 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) can observe the output of the channel Q, and it only has access to 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 Γ to be the set of all such admissible policies.

The progress of the system is determined by (1) and the following relationships:

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

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

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

where $\mathcal{T}$ is the transition kernel of the model which is a stochastic kernel from $\mathbb{X} \times \mathbb{U}$ to $\mathbb{X}$.

Using stochastic realization results (see Lemma 1.2 in [1], or Lemma 3.1 of [2]), dynamics of the system can also be represented in a functional form equivalent to the above relationships as follows: we can consider a dynamical system described by the discrete-time equations

$$X_{t+1} = f(X_t, U_t, W_t), \quad Y_t = g(X_t, V_t) \tag{2}$$

for some measurable functions f, g, with $\{W_t\}$ being an independent and identically distributed (i.i.d) system noise process and $\{V_t\}$ an i.i.d. disturbance process, which are independent of X_0 and each other. Here, the first equation

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

The authors are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada, Email: {16adk,yuksel}@mast.queensu.ca

represents the transition kernel $\mathcal{T}$ as it gives the relation of the most recent state and action variables to the upcoming state. From this representation it can be seen that the probabilistic nature of the kernel is determined by the function f and the probability model of the noise W_t. The second equation represents the communication channel Q, as it describes the relation between the state and observation variables.

We let the objective of the agent(decision maker) be the minimization of the infinite horizon discounted cost,

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

for some discount factor $\beta \in (0,1)$, over the set of admissible policies $\gamma \in \Gamma$, 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 $P \in \mathcal{P}(\mathbb{X})$, 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 transition kernels as

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

The focus of the paper will be to address the following problems:

Problem P1: Continuity of $J_\beta^*(P,\mathcal{T})$ under the convergence of the transition kernels. Let $\{\mathcal{T}_n, n \in \mathbb{N}\}$ be a sequence of transition kernels which converge in some sense to another transition kernel $\mathcal{T}$. Does that imply that

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

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 $\mathcal{T}_n$, an incorrect model for a true model $\mathcal{T}$. Is it the case that if $\mathcal{T}_n \to \mathcal{T}$ then $J_\beta(P,\mathcal{T},\gamma_n^*) \to J_\beta^*(P,\mathcal{T})$?

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)?

We refer the reader to Section II-B for a flavor of the application models.

In Section II, we introduce the convergence criteria for controlled transition kernels are introduced. In Section III, continuity properties of the optimal cost are studied. Building on these results, we will analyze the robustness of the optimal control problem with respect to incorrectly estimated system models/kernels in Section IV. Finally, we apply our results to the setup where a system model is learned through the collection of empirical data in Section V.

B. Literature review

A common approach to robustness in the literature has been to design controllers that works sufficiently well for all possible uncertain systems under some structured constraints, such as H_∞ norm bounded perturbations (see [3], [4]). The design for robust controllers has often been developed through a game theoretic formulation where the minimizer is the controller and the maximizer is the uncertainty. The connections of this formulation to risk sensitive control were established in [5], [6]. Using Legendre-type transforms, relative entropy constraints came in to the literature to probabilistically model the uncertainties, see e.g. [7, Eqn. (4)].

For distributionally robust stochastic optimization problems, it is assumed that the underlying probability measure of the system lies within an ambiguity set and a worst case single-stage optimization is made considering the probability measures in the ambiguity set. To construct ambiguity sets, [8], [9] use the Wasserstein metric (see Section II), [10] uses the Prokhorov metric which metrizes the weak topology, [11] uses the total variation distance and [12] works with relative entropy. Further related studies include [13] which studies the optimal control of systems with unknown dynamics for a Linear Quadratic Regulator setup. [14] considers stochastic uncertainties while [15] considers deterministic structured uncertainties in robust control; some connections of these with our paper can be seen in the examples presented in Section II-B. Related work also includes [16], [17].

For a more detailed discussion on the literature, we refer the reader to the long version of this paper [18].

Contributions. (i) Compared to the existing literature, we obtain strictly refined robustness results: We show that continuity and robustness cannot always be established under weak convergence of transition kernels (or approximate models) to a true kernel in general, but that the optimal cost is continuous in the transition kernels under the convergence in total variation. By imposing further assumptions on the measurement models and on the actual kernel itself, we also show that the optimal cost can be made continuous under weak convergence of transition kernels. (ii) Using the continuity findings in (i), we establish bounds on the mismatch error that occurs due to the application of a control policy which is designed for an incorrectly estimated system model in terms of a distance measure between true model and the incorrect one; and thus we establish robustness properties due to mismatch. On the other hand, we show that robustness may not hold under weak convergence in the sense that as the assumed model and the true model converge to one another, the loss due to mismatch may not go to zero. (iii) The findings lead to consequential positive results on empirical learning in (data-driven) stochastic control (see Section V) since often, in many applications, system models are learned through empirical data.

II. Some Examples and Convergence Criteria for Transition Kernels

A. Convergence of probability measures and convergence criteria for transition kernels.

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

$$\int_{\mathbb{R}^N} c(x)\mu_n(\mathrm{d}x) \to \int_{\mathbb{R}^N} c(x)\mu(\mathrm{d}x) \qquad (*)$$

for every continuous 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

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

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

We should note that the convergence of two probability measures under relative entropy distance implies the convergence in total variation through Pinsker's inequality [19, Lemma 5.2.8].

Another metric for probability measures is the Wasserstein distance. In general, convergence in the Wasserstein distance of order 1 implies weak convergence (in particular this metric bounds from above the Bounded-Lipschitz metric [20, p.109]). 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

(i) $\mathcal{T}_n \to \mathcal{T}$ weakly if $\mathcal{T}_n(\cdot|x,u) \to \mathcal{T}(\cdot|x,u)$ weakly, for all $x \in \mathbb{X}$ and $u \in \mathbb{U}$.

(ii) $\mathcal{T}_n \to \mathcal{T}$ 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 $u \in \mathbb{U}$.

B. Examples

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. As it is stated earlier, the probabilistic nature of the transition kernel is affected by the function F and the probability model of the noise w_t. In the following, we sometimes assume the function F is not known perfectly or the exact probability model of the noise is unknown or both of them are unknown simultaneously.

(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 $\mathcal{T}_n$ converges weakly to $\mathcal{T}$.

(ii) Much of the robust control literature deals with the following type of uncertain deterministic systems: $\tilde{F}(x_t,u_t) = F(x_t,u_t) + \Delta F(x_t,u_t)$, where $F(\cdot)$ represents the nominal model and $\Delta F(\cdot)$ is the model uncertainties(see e.g. [21], [15], [13]). For these systems, $x_{t+1} = \tilde{F}(x_t,u_t)$, can be viewed to be a special case of the analysis here in view of the discussion in (i): For such deterministic systems, pointwise convergence of $\tilde{F}$ to F, i.e. $\Delta F(x_t,u_t) \to 0, \forall x_t, u_t$, can be viewed as weak convergence for deterministic systems in the context of the discussion in (i). It is evident, however, that total variation convergence would be too strong for such a convergence criterion, since $\delta_{\tilde{F}(\cdot)} \to \delta_{F(\cdot)}$ weakly but $\|\delta_{\tilde{F}} - \delta_F\|_{TV} = 2$ for all $\Delta F(x_t,u_t) > 0$.

(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 $w \sim \mu$ is not known correctly, an incorrect model μ_n is assumed. If $\mu_n \to \mu$ weakly or in total variation then the corresponding transition kernels $\mathcal{T}_n$ converges in the same sense to $\mathcal{T}$.

(iv) Suppose now that F and the probability model of w_t is unknown and they are assumed to be F_n and μ_n. If $\mu_n \to \mu$ weakly and $F_n(x,u,w_n) \to F(x,u,w)$ for all $(x,u) \in \mathbb{X} \times \mathbb{U}$ and for $w_n \to w$, then the transition kernel $\mathcal{T}_n$ corresponding to the model F_n converges weakly to the one of F, $\mathcal{T}$.

(v) These studies will be used and analyzed in detail in Section V, where data-driven stochastic control problems will be considered. We will make the point that, for empirical learning weak convergence is a naturally applicable convergence criterion, whereas the applicability of convergence and robustness under stronger notions such as total variation require non-trivial knowledge on and restrictions for the system model.

The analysis in the paper will thus provide robustness results for a large class of control systems as studied in the aforementioned examples.

III. Continuity of Optimal Cost with respect to Convergence of Transition Kernels

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

A. Absence of continuity under weak convergence

The following result shows that the optimal discounted cost may not be continuous under the weak convergence of transition kernels.

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

Proof. We prove the result with a counter example. Let $\mathbb{X} = \mathbb{U} = \mathbb{Y} = [-1,1]$ 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 $P \sim \delta_1$ that is $X_0 = 1$ and the transition kernels are defined to be,

$$\mathcal{T}(x_1|x,u) = \delta_{-1}(x)[\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}] + \delta_1(x)[\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}]$$

$$+ (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0$$

$$\mathcal{T}_n(x_1|x,u) = \delta_{-1}(x)[\frac{1}{2}\delta_{(1-1/n)} + \frac{1}{2}\delta_{(-1+1/n)}]$$
$$+ \delta_1(x)[\frac{1}{2}\delta_{(1-1/n)} + \frac{1}{2}\delta_{(-1+1/n)}]$$
$$+ (1-\delta_{-1}(x))(1-\delta_1(x))\delta_0.$$

It can be seen that $\mathcal{T}_n \to \mathcal{T}$ weakly according to Definition 1(i).

Under this setup the optimal discounted costs can be found as,

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

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

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

B. A sufficient condition for continuity under weak convergence

In the following, we will establish some regularity properties for the optimal cost in the space of transition kernels.

Assumption 1. *(a) The stochastic kernel $\mathcal{T}(dx_1|x_0 = x, u_0 = u)$ is weakly continuous in (x,u).*
(b) The observation channel $Q(dy|x)$ is continuous in total variation (Density admitting noise additive channels satisfy this property, e.g. AWGN channels).
(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.

Lemma 1. *Suppose a 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 $(x_n,u_n)\to(x,u)$. Under Assumption 1*

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

Proof sketch.

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

Using Assumption 1, it can be shown that (see [18, Appendix A.1]) for any $t\geq 0$:

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

To show this, we use the total variation continuity of the channel, which allows us to work on fixed observation realizations $y\in\mathbb{Y}$. Hence, we use the argument that $\mathcal{T}_n(\cdot|x_t,\gamma(y_[0,t])) \to \mathcal{T}(\cdot|x_t,\gamma(y_[0,t]))$ weakly uniformly over γ as $\gamma(y_[0,t])$ takes values from a compact space ($\mathbb{U}$) for fixed observations.

Using the dominance of the discount factor for large time steps, the result follows. □

Theorem 2. *Under the conditions of Lemma 1*

$$\lim_{n\to\infty}|J^*_\beta(\mathcal{T}_n,P) - J^*_\beta(\mathcal{T},P)| = 0.$$

Proof. We start with the following bound,

$$|J^*_\beta(\mathcal{T}_n) - J^*_\beta(\mathcal{T})|$$
$$\leq \max\Big(J_\beta(\mathcal{T}_n,\gamma^*) - J_\beta(\mathcal{T},\gamma^*), J_\beta(\mathcal{T},\gamma^*_n) - J_\beta(\mathcal{T}_n,\gamma^*_n)\Big)$$

where γ^* and γ^*_n are the optimal policies respectively for $\mathcal{T}$ and $\mathcal{T}_n$. Both terms inside of $\max$ go to 0 by Lemma 1. □

C. Continuity under total variation

We now propose a result that gives an upper bound for the rate of convergence. For the result we will make use of strategic measures . For stochastic control problems, *strategic measures* are defined (see Schäl [22], also [23], [24]) as the set of probability measures induced on the product spaces of the state and action pairs by measurable control policies: Given an initial distribution on the state, and a policy, one can uniquely define a probability measure on the infinite product space consistent with finite dimensional distributions, by Ionescu Tulcea theorem [25]. Now, define a *strategic measure* under a policy $\gamma^n = \{\gamma_0^n, \gamma_1^n, \cdots, \gamma_k^n, \cdots\}$ as a probability measure defined on $\mathcal{B}(\mathbb{X}\times\mathbb{Y}\times\mathbb{U})^{\mathbb{Z}_+}$ by:

$$P_{\mathcal{T}}^{\gamma^n}(\mathrm{d}(x_0,y_0,u_0),\mathrm{d}(x_1,y_1,u_1),\cdots)$$
$$= P(\mathrm{d}x_0)Q(\mathrm{d}y_0|x_0)1_{\{\gamma^n(y_0)\in\mathrm{d}u_0\}}\mathcal{T}(\mathrm{d}x_1|x_0,u_0)$$
$$\times Q(\mathrm{d}y_1|x_1)1_{\{\gamma^n(y_0,y_1)\in\mathrm{d}u_1\}}\cdots$$

Theorem 3. *If the cost function c is bounded,*

$$|J^*_\beta(P,\mathcal{T}_n) - J^*_\beta(P,\mathcal{T})| \leq$$
$$\|c\|_\infty\frac{\beta}{(\beta-1)^2}\sup_{x\in\mathbb{X},u\in\mathbb{U}}\|\mathcal{T}_n(.|x,u) - \mathcal{T}(.|x,u)\|_{TV}.$$

Proof sketch. We start with the following bound as before,

$$|J^*_\beta(P,\mathcal{T}_n) - J^*_\beta(P,\mathcal{T})|$$
$$\leq \max\Big(J_\beta(\mathcal{T}_n,\gamma^*) - J_\beta(\mathcal{T},\gamma^*), J_\beta(\mathcal{T},\gamma^*_n) - J_\beta(\mathcal{T}_n,\gamma^*_n)\Big).$$

Then we have

$$|J_\beta(P,\mathcal{T}_n,\gamma) - J_\beta(P,\mathcal{T},\gamma)| \tag{3}$$
$$\leq \sum_k \beta^k\|c\|_\infty\|P^\gamma_{\mathcal{T}_n}(d(x,y,u)_{[0,k]}) - P^\gamma_{\mathcal{T}}(d(x,y,u)_{[0,k]})\|_{TV}.$$

The result follows from the following relation (see [18, Appendix A.3]):

$$\|P^\gamma_{\mathcal{T}_n}(d(x,y,u)_{[0,k]}) - P^\gamma_{\mathcal{T}}(d(x,y,u)_{[0,k]})\|_{TV}$$
$$\leq k\sup_{x\in\mathbb{X},u\in\mathbb{U}}\|\mathcal{T}(.|x,u) - \mathcal{T}_n(.|x,u)\|_{TV}. \tag{4}$$

□

IV. Robustness to incorrect transition kernels

Suppose we design an optimal policy, γ^*_n, for a transition kernel, $\mathcal{T}_n$, assuming it is the correct model and apply the policy to the true model whose transition kernel is $\mathcal{T}$. In this section,we ask the following question: Does the cost caused by γ^*_n converge to the true optimal cost as $\mathcal{T}_n$ converges in some sense to $\mathcal{T}$?

A. Total variation

We propose a result for bounding our loss that arises because of applying the incorrectly calculated policy to the actual model.

Theorem 4. *Suppose the stage-wise cost function $c(x,u)$ is bounded in $\mathbb{X}\times\mathbb{U}$, then*

$$|J_\beta(P,\mathcal{T},\gamma_n^*) - J_\beta^*(P,\mathcal{T})| \leq 2\|c\|_\infty \frac{\beta}{(\beta-1)^2} \sup_{x\in\mathbb{X},u\in\mathbb{U}} \|\mathcal{T}(.|x,u) - \mathcal{T}_n(.|x,u)\|_{TV}$$

for a fixed prior distribution $P\in\mathcal{P}(\mathbb{X})$, where γ_n^* is the optimal policy designed for the transition kernel $\mathcal{T}_n$.

Proof. We begin with the following,

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

The second term is bounded using Theorem 3. For the first term, we use the inequalities 3 and 4. □

B. Weak convergence

Theorem 5. *Under the conditions of Lemma 1*

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

for a fixed prior distribution $P\in\mathcal{P}(\mathbb{X})$, where γ_n^* is the optimal policy designed for the transition kernel $\mathcal{T}_n$.

Proof. We write

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

The first term goes to 0 by Lemma 1. For the second term we use Theorem 2. □

C. Fully observed models

We now present results for the case where the controller has access to state directly. Notice also that fully observed models can be viewed as partially observed with the measurement channel thought to be

$$Q(\cdot|x) = \delta_x(\cdot), \tag{5}$$

which is only weakly continuous, thus it does not satisfy Assumption 1(b). The following result shows that the sufficient conditions for partially observed case cannot guarantee robustness for the fully observed case under weak convergence and also shows that the robustness is not a direct consequence of continuity.

Theorem 6. *(i) Under the conditions of Lemma 1 (weak convergence of kernels), $J_\beta^*(\mathcal{T}_n)\to J_\beta^*(\mathcal{T})$.*

(ii) Even if the conditions of Lemma 1 holds, the model may not be robust, i.e. it is not always true that $J_\beta(\mathcal{T},\gamma_n^*)\to J_\beta^*(\mathcal{T})$.

(iii) The continuity result in Theorem 3 and the robustness result in Theorem 4 apply to this case since the fully observed model may be viewed as a partially observed model with the measurement channel Q given in (5).

For a proof, please see [18].

V. Implications for Data-Driven Learning Methods in Stochastic Control

In practice, one might try to learn the transition kernel of the chain from the observed data. This can be done using the empirical history of the process, i.e. up to a finite time horizon N, the state and action sequence $\{x_1,u_1,x_2,u_2,\cdots,x_N,u_N\}$ can be used to infer some information about the transition laws of the process.

Let us briefly discuss the case where a random variable is repeatedly observed, buts 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 set $B\subset\mathbb{X}$, 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 (a.s.) by the strong law of large numbers. Also, $\mu_n\to\mu$ weakly with probability one ([26], Theorem 11.4.1). However, μ_n can not converge to μ in total variation, in general. On the other hand, if we know that μ admits a density, we can find estimators to estimate μ under total variation [27].

Corollary 1 (to Theorem 3 and Theorem 4)**.** Suppose we are given the following dynamics for finite state space $\mathbb{X}$, and finite action space $\mathbb{U}$,

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

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

$$J_\beta^*(\mathcal{T}_n)\to J_\beta^*(\mathcal{T}),\quad J_\beta(\mathcal{T},\gamma_n^*)\to J_\beta^*(\mathcal{T})$$

with probability 1, where γ_n^* is the optimal policy designed for $\mathcal{T}_n$.

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 have $\mathcal{T}_n(\cdot|x,u)\to\mathcal{T}(\cdot|x,u)$ under total variation. Therefore, by Theorem 3 and Theorem 4, the result follows. □

The following holds for more general spaces.

Corollary 2 (to Theorem 5 and Theorem 2)**.** Suppose we are given the following dynamics with state space $\mathbb{X}$ and compact action space $\mathbb{U}$,

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

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown. Suppose that $f(x,u,w):\mathbb{W}\to\mathbb{X}$

is invertible for all (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 $B_w \subset \mathbb{W}$, and for every $n \in \mathbb{N}$, the empirical occupation measures are

$$\mu_n(B_w) = \frac{1}{n}\sum_{t=0}^{n} \mathbb{1}_{\{f^{-1}_{x_{i-1},u_{i-1}}(x_i)\in B_w\}} \tag{6}$$

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 empirical measures, we can construct the empirical transition kernels such that for any (x_0,u_0) and every (fixed) Borel $B_x \subset \mathbb{X}$

$$\mathcal{T}_n(B_x|x_0,u_0) = \mu_n(f^{-1}_{x_0,u_0}(B_x)). \tag{7}$$

If the measurement channel represented by the function g is continuous in total variation then

$$J^*_\beta(\mathcal{T}_n) \to J^*_\beta(\mathcal{T}), \quad J_\beta(\mathcal{T},\gamma^*_n) \to J^*_\beta(\mathcal{T})$$

with probability 1, where γ^*_n is the optimal policy designed for $\mathcal{T}_n$.

Proof. We have $\mu_n \to \mu$ weakly with probability one where μ is the true model. It can be shown that the transition kernels are such that $\mathcal{T}_n(\cdot|x_n,u_n) \to \mathcal{T}(\cdot|x,u)$ weakly for any $(x_n,u_n)\to(x,u)$. 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, by Theorem 5 and Theorem 2 if $c(x,u)$ is bounded and $\mathbb{U}$ is compact, the result follows. □

Corollary 3 (to Theorem 3 and Theorem 4)**.** Suppose we are given the following dynamics for a general state space, $\mathbb{X}$, and action space, $\mathbb{U}$,

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

where $\{w_t\}$ and $\{v_t\}$ are i.i.d.noise processes and the noise models are unknown however it is known that the noise w_t admits a continuous probability density function.

Suppose that $f(x,u,w):\mathbb{W}\to\mathbb{X}$ is invertible for all (x,u) . It can be shown that we can construct estimators for the noise process w_t which estimates the true model consistently under total variation. We can then also construct the estimators for the kernels, $\tilde{\mathcal{T}}_n$, as in 7.

Then independent of the channel

$$J^*_\beta(\tilde{\mathcal{T}}_n) \to J^*_\beta(\mathcal{T}), \quad J_\beta(\mathcal{T},\gamma^*_n) \to J^*_\beta(\mathcal{T})$$

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

Proof. By [27] we can estimate μ in total variation so that almost surely $\lim_{n\to\infty}\|\mu_n-\mu\|_{TV}=0$. One can show that the convergence of μ_n to μ under total variation metric implies the convergence of $\tilde{\mathcal{T}}_n$ to $\mathcal{T}$ in total variation uniformly over all $x\in\mathbb{X}$ and $u\in\mathbb{U}$ i.e. $\lim_{n\to\infty}\sup_{x,u}\|\tilde{\mathcal{T}}_n(\cdot|x,u)-\mathcal{T}(\cdot|x,u)\|_{TV}=0$.Thus, by Theorem 3 and Theorem 4, the result follows. □

VI. Conclusion

We studied regularity properties of optimal stochastic control on the space of transition kernels, and practical implications of these on robustness of optimal control policies designed for an incorrect model applied to an actual system, and applications to empirical learning.

References

[1] I. I. Gihman and A. V. Skorohod, *Controlled stochastic processes*. Springer Science & Business Media, 2012.

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

[3] T. Başar and P. Bernhard, *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Boston, MA: Birkhäuser, 1995.

[4] K. Zhou, J. C. Doyle, and K. Glover, *Robust and optimal control*. Prentice-Hall, 1996, vol. 40.

[5] D. Jacobson, "Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games," *IEEE Transactions on Automatic control*, vol. 18, no. 2, pp. 124–131, 1973.

[6] P. Dupuis, M. R. James, and I. Petersen, "Robust properties of risk-sensitive control," *Mathematics of Control, Signals and Systems*, vol. 13, no. 4, pp. 318–332, 2000.

[7] P. D. Pra, L. Meneghini, and W. J. Runggaldier, "Connections between stochastic control and dynamic games," *Mathematics of Control, Signals and Systems*, vol. 9, no. 4, pp. 303–326, 1996.

[8] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *SSRN Electronic Journal*, 04 2016.

[9] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, pp. 1–52, 2017.

[10] E. Erdoğan and G. N. Iyengar, "Ambiguous chance constrained problems and robust optimization," *Mathematical Programming*, vol. 107, no. 1-2, pp. 37–61, 2005.

[11] H. Sun and H. Xu, "Convergence analysis for distributionally robust optimization and equilibrium problems," *Mathematics of Operations Research*, vol. 41, pp. 377–401, 07 2015.

[12] H. Lam, "Robust sensitivity analysis for stochastic systems," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1248–1275, 2016.

[13] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *arXiv preprint arXiv:1710.01688v2*, 2018.

[14] V. A. Ugrinovskii, "Robust H-infinity control in the presence of stochastic uncertainty," *International Journal of Control*, vol. 71, no. 2, pp. 219–237, 1998.

[15] A. V. Savkin and I. R. Petersen, "Robust control of uncertain systems with structured uncertainty," *Journal of Mathematical Systems, Estimation, and Control*, vol. 6, no. 3, pp. 1–14, 1996.

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

[17] A. D. Kara and S. Yüksel, "Robustness to incorrect priors in partially observed stochastic control," *arXiv preprint arXiv :1803.05103*, 2018.

[18] ——, "Robustness to incorrect system models and application to data-driven learning in stochastic control," *arXiv preprint arXiv:1803.06046*, 2018.

[19] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[20] C. Villani, *Optimal transport: old and new*. Springer, 2008.

[21] M. A. Rotea and P. Khargonekar, "Stabilization of uncertain systems with norm bounded uncertaintya control lyapunov function approach," *SIAM Journal on Control and Optimization*, vol. 27, no. 6, pp. 1462–1476, 1989.

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

[23] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov processes*. Springer, 1979, vol. 235.

[24] E. A. Feinberg, "On measurability and representation of strategic measures in Markov decision processes," *Lecture Notes-Monograph Series*, pp. 29–43, 1996.

[25] O. Hernandez-Lerma and J. Lasserre, *Discrete-time Markov control processes*. Springer, 1996.

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

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