Data-driven approximate dynamic programming

A linear programming approach

Sutter, Tobias; Kamoutsi, Angeliki; Esfahani, Peyman Mohajerin; Lygeros, John

DOI
10.1109/CDC.2017.8264426

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings of the 2017 IEEE 56th Annual Conference on Decision and Control

Data-driven approximate dynamic programming: A linear programming approach

Tobias Sutter, Angeliki Kamoutsi, Peyman Mohajerin Esfahani, and John Lygeros

***Abstract*— This article presents an approximation scheme for the infinite-dimensional linear programming formulation of discrete-time Markov control processes via a finite-dimensional convex program, when the dynamics are unknown and learned from data. We derive a probabilistic explicit error bound between the data-driven finite convex program and the original infinite linear program. We further discuss the sample complexity of the error bound which translates to the number of samples required for an a priori approximation accuracy. Our analysis sheds light on the impact of the choice of basis functions for approximating the true value function. Finally, the relevance of the method is illustrated on a truncated LQG problem.**

I. Introduction

We are concerned with discrete-time Markov control processes (MCPs) with Borel (general uncountable) state and action spaces and the long run expected average cost optimality criterion. These stochastic optimal control problems are key tools of mathematical modelling and appear in many fields such as engineering and operations research. A unified theoretical framework, consisting mainly of dynamic programming techniques, has been developed over the years to solve them [1], [2], [3]. However, oftentimes it is impossible to obtain a solution in closed form, which motivates the task of finding tractable approximations. Such approximation schemes are the core of a methodology known as *approximate dynamic programming*, which has been extensively studied in the literature from different perspectives [4], [5], [6], [7]; see [8] for a comprehensive survey on this field.

In addition, in many realistic applications the underlying dynamics are unknown and the decision maker needs to learn the optimal policy by trial-and-error, through interaction with the environment. In such a setting of unknown dynamics, the problem is particularly difficult and a prevalent approach in the existing literature consists of dynamic-programming-based reinforcement learning methods, also known as neuro-dynamic programming [9], [10], [4]. The two most common types of such reinforcement learning algorithms are Q-learning and actor-critic algorithms. Q-learning algorithms [11] are simulation-based schemes derived from value iteration, while actor-critic methods [12] are simulation-based, two-time scale variants of policy iteration. Q-learning comes with asymptotic convergence guarantees but it is mostly considered in the case that state and action spaces are both discrete. On the other hand, while actor-critic algorithms can tackle continuous action and state spaces, since they are gradient-based, one can prove convergence only to a local optimum.

In this work, we present a data-driven algorithm that is based on the linear programming (LP) approach to MCPs. The LP approach to finite state/finite action MCPs has been studied in the pioneering work [13]. Later extensions of the approach to discrete-time MCPs with uncountable state and action spaces for several cost criteria, have been investigated, e.g., [1], [14], [15], [16]. In particular, MCPs can be recast as abstract "static" optimization problems over a closed convex set of measures as infinite-dimensional linear programs. This reformulation allows the use of tools from the well-established field of convex programming to tackle them. Furthermore, the LP approach to MCPs is particularly appealing from the perspective of dealing with unconventional problems involving additional constrains or secondary costs, where traditional dynamic programming techniques are not applicable [17], [18], [19].

This article presents an approximation scheme for the infinite-dimensional LP formulation of MCPs via a finite-dimensional convex program and can be seen as an extension of [7], to the case where the transition kernel is unknown but information on it is obtained by simulation. More specifically, in response to the current state and action, data about the next state is received. We derive a probabilistic explicit error bound between the data-driven finite convex program and the original infinite LP (Theorem 2) and discuss the sample complexity of the error bound, i.e., how many data are required for a certain approximation accuracy. Moreover, our analysis provides insight on what is a good choice of basis functions that are used to approximate the value function.

Notation. For $p \in [1,\infty]$, we denote by $\|\cdot\|_p$ the p-norm in $\mathbb{R}^n$. Let (X,ρ) be a metric space. Given a function $u : X \to \mathbb{R}$, its sup-norm is given by $\|u\|_\infty := \sup_{x\in X} |u(x)|$, and its Lipschitz norm by $\|u\|_L := \max\{\|u\|_\infty, \sup_{x\neq x'} \frac{|u(x)-u(x')|}{\rho(x,x')}\}$. The space of real-valued Lipschitz functions on a set X is denoted by $\mathcal{L}(X)$. Let $\mathcal{B}(X)$ be the Borel σ-algebra on X. Measurability is always understood in the sense of Borel measurability. Products of metric spaces are assumed to be endowed with the product topology and the corresponding product σ-algebra. Given

Research was supported by the European Union 7th Framework Program "Scalable Proactive Event-Driven Decision-making (SPEEDD)"

The authors are with the Automatic Control Laboratory, ETH Zürich, Switzerland and the Delft Center for Systems and Control, TU Delft, Netherlands; Emails: {sutter, kamoutsa, lygeros}@control.ee.ethz.ch, P.MohajerinEsfahani@tudelft.nl.

a compact subset $A \subset X$, we consider the projection multimapping $\Pi_A(x) := \arg\min_{x' \in A} \rho(x, x')$.

The outline of this paper is as follows. Section II states the problem under consideration, namely the constrained average cost MCP and introduces the infinite-dimensional linear program characterizing it. In Section III, we consider the case of unknown dynamics and we present the approximation of the infinite LP via a finite data-driven convex program. In Section IV, we prove the main theoretical result of the paper, i.e., a probabilistic error bound and its sample complexity between the finite convex program and the original infinite LP. To illustrate the proposed methodology, in Section V, the theoretical results are applied to a truncated LQG problem. We conclude in Section VI with a summary of our work and comment on possible subjects of further research.

II. Infinite LP Characterization

We briefly recall some standard definitions and refer interested readers to [1], [17], [20] for further details. Consider a *Markov control model* $\big(X, A, \{A(x) : x \in X\}, Q, c\big)$, where X (resp. A) is a Borel space (i.e., a Borel subset of a complete and separable metric space) called the *state space* (resp. *action space*). For each $x \in X$ the measurable set $A(x) \subseteq A$ denotes the set of *feasible actions* when the system is in state $x \in X$ and has the property that the set of feasible state-action pairs $\mathbb{K} := \{(x, a) : x \in X, a \in A(x)\}$ is a measurable subset in $X \times A$. The *transition law* is a stochastic kernel Q on X given $\mathbb{K}$. A stochastic kernel acts on bounded measurable functions $u : X \to \mathbb{R}$ from the left as

$$Qu(x,a) := \int_X u(y) Q(\,\mathrm{d}y|x,a), \quad \forall (x,a) \in \mathbb{K},$$

and on probability measures μ on K from the right as

$$\mu Q(B) := \int_K Q(B|x,a)\mu\big(\,\mathrm{d}(x,a)\big), \quad \forall B \in \mathcal{B}(X).$$

Finally $c : \mathbb{K} \to \mathbb{R}_+$ denotes a measurable function called the *one-stage cost function*. The *admissible history spaces* are defined recursively as $H_0 := X$ and $H_t := H_{t-1} \times \mathbb{K}$ for $t \in \mathbb{N}$ and the canonical sample space is defined as $\Omega := (X \times A)^\infty$. All random variables will be defined on the measurable space $(\Omega, \mathcal{B})$ where $\mathcal{B}$ denotes the corresponding product σ-algebra. A generic element $\omega \in \Omega$ is of the form $\omega = (x_0, a_0, x_1, a_1, \ldots)$, where $x_i \in X$ are the states and $a_i \in A$ the action variables. An *admissible policy* is a sequence $\pi = (\pi_t)_{t \in \mathbb{N}_0}$ of stochastic kernels π_t on A given $h_t \in H_t$, satisfying the constraints $\pi_t(A(x_t)|h_t) = 1$. The set of admissible policies will be denoted by Π. Given a probability measure $\nu \in \mathcal{P}(X)$ and a policy $\pi \in \Pi$, by the Ionescu Tulcea theorem [21, p. 140-141] there exists a unique probability measure $\mathbb{P}^\pi_\nu$ on $(\Omega, \mathcal{B})$ such that for all measurable sets $B \subset X$, $C \subset A$, $h_t \in H_t$, and $t \in \mathbb{N}_0$

$$\mathbb{P}^\pi_\nu\big(x_0 \in B\big) = \nu(B)$$
$$\mathbb{P}^\pi_\nu\big(a_t \in C|h_t\big) = \pi_t(C|h_t)$$
$$\mathbb{P}^\pi_\nu\big(x_{t+1} \in B|h_t, a_t\big) = Q(B|x_t, a_t).$$

The expectation operator with respect to $\mathbb{P}^\pi_\nu$ is denoted by $\mathbb{E}^\pi_\nu$. The stochastic process $\big(\Omega, \mathcal{B}, \mathbb{P}^\pi_\nu, (x_t)_{t\in\mathbb{N}_0}\big)$ is called a *discrete-time MCP*. In this article we consider optimal control problems where the aim is to minimize the long run *average cost* (AC) over the set of admissible policies and initial state measures. We define the optimal value of our AC optimal control problem by

$$J^{\mathrm{AC}} := \inf_{(\pi,\nu)\in\Pi\times\mathcal{P}(X)} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi_\nu\left(\sum_{t=0}^{T-1} c(x_t, a_t)\right). \quad (1)$$

We emphasize, however, that the results presented also apply to other performance objectives, including problems with *discounted payoff*.

We impose the following assumptions on the control model which hold throughout the article.

Assumption 1 (Control model):

(i) the set of feasible state-action pairs is the unit hypercube $\mathbb{K} := [0,1]^{\dim(X\times A)}$;

(ii) the transition law Q is Lipschitz continuous, i.e., there exists $L_Q > 0$ such that for all $k, k' \in \mathbb{K}$ and all continuous functions $u : X \to \mathbb{R}$

$$|Qu(k) - Qu(k')| \le L_Q \|u\|_\infty \|k - k'\|_\infty;$$

(iii) the cost function c is non-negative and Lipschitz continuous on $\mathbb{K}$ with respect to the ∞-norm.

Consider the (infinite) linear program

$$J := \left\{ \begin{array}{ll} \sup\limits_{\rho,u} & \rho \\ \text{s.t.} & \rho + \mathcal{T}u(x,a) \le c(x,a), \ \forall (x,a) \in \mathbb{K} \\ & \rho \in \mathbb{R}, \ u \in \mathcal{L}(X), \end{array} \right. \quad (2)$$

where $\mathcal{T} : \mathcal{L}(X) \to \mathcal{L}(X \times A)$, defined by

$$\begin{aligned} \mathcal{T}u(x,a) :=& u(x) - \int_X u(y) Q(\,\mathrm{d}y|x,a) \\ =& u(x) - \mathbb{E}^{Q(\cdot|x,a)}\big[u(y)\big] \end{aligned} \quad (3)$$

denotes a linear, weakly continuous operator [14]. The linear programming formulation (2) is an alternative characterization of the problem (1) in the sense of the following theorem.

Theorem 1 ([22, Proposition 2.4]): Under Assumption 1, the LP (2) is solvable (i.e., the supremum in (2) is attained) and $J^{\mathrm{AC}} = J$.

We denote the optimizer of problem (2) by $u^\star$. The focus of our study is on providing an approximation for the linear program (2) via a finite dimensional convex program. Moreover, we treat the setting where the transition kernel Q is unknown but information on it is obtained by simulation. In response to the current state and action the next state is received.

III. Finite Approximation

Let $\{(x_j, a_j)\}_{j\leq N}$ be i.i.d. samples generated with respect to some probability measure $\mathbb{P}_1$ supported on $\mathbb{K}$. We propose the following finite-dimensional convex program as an approximation to the infinite LP (2) and hence to the optimal control problem (1)

$$J^m_{n,N} = \begin{cases} \sup\limits_{(\rho,\alpha)\in\mathbb{R}^{n+1}} & \rho \\ \text{s.t.} & \rho + \sum\limits_{i=1}^{n} \alpha_i \mathcal{T}_m u_i(x_j, a_j) \\ & \quad \leq c(x_j, a_j),\ \forall j \in \{1, \cdots, N\} \\ & \|\alpha\|_2 \leq \theta, \end{cases} \tag{4}$$

where $\{u_i\}_{i=1}^n \subset \mathcal{L}(X)$ is a family of linearly independent elements, called the basis functions and $\theta > 0$ is a regularization parameter. Moreover, we use the following notation

$$\mathcal{T}_m u(x,a) := u(x) - \frac{1}{m}\sum_{i=1}^{m} u(y_i), \quad y_i \overset{\text{i.i.d.}}{\sim} Q(\cdot|x,a).$$

Note that the program (4) does not require knowledge of the transition kernel Q, but instead, it uses simulations to learn Q via the samples y_i in the operator $\mathcal{T}_m$. In the following, we quantify the approximation error of (4) with respect to (2). To this end, another assumption is needed.

Assumption 2 (Approximation method): The basis functions satisfy $\|u_i\|_L \leq 1$, for all $i = 1, \ldots, n$.

We set $\mathbb{U}_n := \{\sum_{i=1}^n \alpha_i u_i : \|\alpha\|_2 \leq \theta\}$ and

$$\mathsf{N}(n, \varepsilon, \beta) := \min\left\{N \in \mathbb{N} : \sum_{i=0}^{n-1} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i} \leq \beta\right\}.$$

The following theorem is the main theoretical result of this article.

Theorem 2: Given Assumptions 1 and 2, let $\varepsilon, \beta \in (0,1)$ and consider the finite convex program (4) where the number of sampled constraints satisfies $N \geq \mathsf{N}(n+1, (\frac{\varepsilon z_n}{2})^{\dim(\mathbb{K})}, \frac{\beta}{2})$, where $z_n := \big(\theta\sqrt{n}(\max\{L_Q, 1\}+1) + \|c\|_L\big)^{-1}$ and $m \geq \frac{8Cn\theta^2 \log(4nN/\beta)}{\varepsilon^2}$. Then, with probability $1-\beta$

$$\left|J - J^m_{n,N}\right| \leq \big(1 + \max\{L_Q, 1\}\big)\left\|u^\star - \Pi_{\mathbb{U}_n}(u^\star)\right\|_L + \varepsilon.$$

The proof is given in Section IV. Note that that $J^m_{n,N}$ is a real valued random variable on the space $(\mathbb{K} \times X^m)^N$. Strictly speaking, the error bound of Theorem 2 has to be interpreted with respect to $\mathbb{P}_2^N$, where $\mathbb{P}_2$ is a probability measure on $\mathbb{K} \times X^m$ defined by $\mathbb{P}_2[\,\mathrm{d}(x, a, y_1, \ldots, y_m)] := Q^m(\,\mathrm{d}y|x,a)\mathbb{P}_1[\,\mathrm{d}(x,a)]$ and $\mathbb{P}_2^N$ stands for the N-fold product probability measure. For simplicity we slightly abuse the notation and use $\mathbb{P}$ instead of $\mathbb{P}_2^N$, and will be doing so hereinafter.

Remark 1 (Projection residual): The residual error $\left\|u^\star - \Pi_{\mathbb{U}_n}(u^\star)\right\|_L$ can be approximated by leveraging results from the literature on universal function approximation. Prior information about the value function $u^\star$ may offer explicit quantitative bounds. For instance, for MCPs satisfying Assumption 1, we know that $u^\star$ is Lipschitz continuous. For appropriate choice of basis functions, we can therefore ensure a convergence rate of $n^{-1/\dim(X)}$, see for instance [23] for polynomials and [24] for the Fourier basis functions.

Remark 2 (Curse of dimensionality): As explained in [25, Remark 3.9] and [7, Remark 4.5], the number N of sampled constraints grows linearly in n and logarithmically in $1/\beta$. It, however, has an exponential growth as $\varepsilon^{\dim(X\times A)}$. To mitigate this inherent computational complexity, one may resort to a more elegant sampling approach so that the required number of samples N has a sublinear rate in the second argument, see for instance [26].

To select θ, one may minimize the complexity of the a priori bound in Theorem 2, which is reflected through the required number of samples (with respect to the state-action space and the state space). At the same time, the impact of the bound θ through the projection residual (cf. Remark 1) should also be taken into account. The first factor is monotonically growing with respect to θ, i.e., the smaller the parameter θ, the lower the number of the required samples. The second factor, i.e., the projection residual, is monotonically decreasing with respect to θ. Therefore, an acceptable choice of θ is an upper bound for the projection error of the optimal solution onto the $\operatorname{span}\{u_1, \ldots, u_n\}$ uniformly in $n \in \mathbb{N}$, i.e.,

$$\theta \geq \sup\left\{\|\alpha^\star\|_2 : \Pi_{\operatorname{span}\{u_1,\ldots,u_n\}}(u^\star) = \sum_{i=1}^{n} \alpha_i^\star u_i,\ n \in \mathbb{N}\right\},$$

where the projection is with respect to the Lipschitz norm. In case that the basis functions are L_2-orthonormal

$$\|\alpha^\star\|_2 \leq \|u^\star\|_L \leq \max\{L_Q, 1\}\|c\|_\infty, \tag{5}$$

where L_Q is the Lipschitz constant in Assumption 1(ii). We note that the first inequality in (5) follows since X is a unit hypercube, and the second inequality follows from [22, Lemma 2.3], see also [22, Section 5] for further detailed analysis.

IV. Proof or Theorem 2

Some preliminaries are needed in order to prove Theorem 2. Consider the finite convex program

$$J_{n,N} := \begin{cases} \sup\limits_{(\rho,\alpha)\in\mathbb{R}^{n+1}} & \rho \\ \text{s.t.} & \rho + \sum\limits_{i=1}^{n} \alpha_i \mathcal{T} u_i(x_j, a_j) \\ & \quad \leq c(x_j, a_j), \forall j \in \{1, \cdots, N\} \\ & \|\alpha\|_2 \leq \theta. \end{cases} \tag{6}$$

Lemma 1: Given Assumption 2, for any $\varepsilon > 0$

$$\mathbb{P}\Big[\left|J^m_{n,N} - J_{n,N}\right| \leq \varepsilon\Big] \geq 1 - 2nN\exp\left(\frac{-\varepsilon^2 m}{2n\theta^2}\right).$$

Proof: As the first step, we invoke the Hoeffding inequality [27] together with the subadditivity of probability

measures[1] which states that for any $\varepsilon > 0$

$$\mathbb{P}\Big[\forall i = 1,\dots,n, j = 1,\dots,N, |\mathcal{T}u_i(x_j,a_j) - \mathcal{T}_m u_i(x_j,a_j)| \leq \varepsilon\Big] \geq 1 - 2nN\exp\left(\frac{-\varepsilon^2 m}{2}\right).$$

Hence, for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\Big[\forall j = 1,\dots,N \sup_{\|\alpha\|_2 \leq \theta} \left|\sum_{i=1}^n \alpha_i \mathcal{T}u_i(x_j,a_j) \right. \\
&\quad \left. - \left|\sum_{i=1}^n \alpha_i \mathcal{T}_m u_i(x_j,a_j)\right| \leq \varepsilon\Big] \\
&\geq \mathbb{P}\Big[\forall i = 1,\dots,n, j = 1,\dots,N \ \|\alpha\|_1 |\mathcal{T}u_i(x_j,a_j) \\
&\quad - \mathcal{T}_m u_i(x_j,a_j)| \leq \varepsilon\Big] \geq 1 - 2nN\exp\left(\frac{-\varepsilon^2 m}{2n\theta^2}\right),
\end{aligned}$$

where we have used Assumption 2 leading to $\|\sum_{i=1}^n \alpha_i u_i\|_L \leq \|\alpha\|_1 \leq \sqrt{n}\theta$. Therefore, with confidence $1 - 2N\exp\left(\frac{-\varepsilon^2 m}{2n\theta^2}\right)$ we have

$$\begin{aligned}
J_{n,N}^m &= \left\{ \begin{array}{ll} \sup\limits_{(\rho,\alpha)\in\mathbb{R}^{n+1}} & \rho \\ \text{s.t.} & \rho + \sum\limits_{i=1}^n \alpha_i \big(\mathcal{T}u_i(x_j,a_j) \\ & \quad + \mathcal{T}_m u_i(x_j,a_j) - \mathcal{T}u_i(x_j,a_j)\big) \\ & \quad \leq c(x_j,a_j), \quad \forall j \in \{1,\cdots,N\} \\ & \|\alpha\|_2 \leq \theta. \end{array} \right. \\
&\geq \left\{ \begin{array}{ll} \sup\limits_{(\rho,\alpha)\in\mathbb{R}^{n+1}} & \rho - \varepsilon \\ \text{s.t.} & \rho + \sum\limits_{i=1}^n \alpha_i \mathcal{T}u_i(x_j,a_j) \\ & \quad \leq c(x_j,a_j), \quad \forall j \in \{1,\cdots,N\} \\ & \|\alpha\|_2 \leq \theta. \end{array} \right. \\
&= J_{n,N} - \varepsilon
\end{aligned}$$

and similarly one can show $J_{n,N}^m \leq J_{n,N} + \varepsilon$, which completes the proof. ■

Proof of Theorem 2: The proof consists of combining three results. First, recall that [7, Corollary 3.9] for the given setting of Theorem 2

$$0 \leq J - J_n \leq \big(1 + \max\{L_Q, 1\}\big)\big\|u^\star - \Pi_{\mathbb{U}_n}(u^\star)\big\|_L, \quad (7)$$

where

$$J_n := \left\{ \begin{array}{ll} \sup\limits_{(\rho,\alpha)\in\mathbb{R}^{n+1}} & \rho \\ \text{s.t.} & \rho + \sum\limits_{i=1}^n \alpha_i \mathcal{T}u_i(x,a) \\ & \quad \leq c(x,a), \quad \forall (x,a) \in X \times A \\ & \|\alpha\|_2 \leq \theta. \end{array} \right.$$

Next, [7, Corollary 3.9] states that for $N \geq \mathsf{N}(n+1, (\varepsilon z_n)^{\dim(K)}, \beta)$, where $z_n := \big(\theta\sqrt{n}(\max\{L_Q,1\}+1) + \|c\|_L\big)^{-1}$

$$\mathbb{P}^N\Big[|J_n - J_{n,N}| \leq \varepsilon\Big] \geq \beta, \quad (8)$$

where $J_{n,N}$ is defined in (6). Finally, a simple union bound of (7), (8) and Lemma 1 concludes the proof. ■

[1] i.e., $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$.

V. Numerical Example

Consider the linear system

$$x_{t+1} = \vartheta x_t + \rho a_t + \xi_t, \quad t \in \mathbb{N},$$

with quadratic stage cost $c(x,a) = qx^2 + ra^2$, where $q \geq 0$ and $r > 0$ are given constants. We assume that $X = A = [-L, L]$ and the parameters $\vartheta, \rho \in \mathbb{R}$ are known. The disturbances $\{\xi_t\}_{t\in\mathbb{N}}$ are i.i.d. random variables generated by a truncated normal distribution with known parameters μ and σ, independent of the initial state x_0. Thus, the process ξ_t has a distribution density

$$f(x,\mu,\sigma,L) = \left\{ \begin{array}{ll} \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{L-\mu}{\sigma}\right) - \Phi\left(\frac{-L-\mu}{\sigma}\right)}, & x \in [-L, L] \\ 0 & \text{o.w.}, \end{array} \right.$$

where ϕ is the probability density function of the standard normal distribution, and Φ is its cumulative distribution function. The transition kernel Q has a density function $q(y|x,a)$, i.e., $Q(B|x,a) = \int_B q(y|x,a)\,\mathrm{d}y$ for all $B \in \mathcal{B}(X)$, that is given by

$$q(y|x,a) = f(y - \vartheta x - \rho a, \mu, \sigma, L).$$

In the special case that $L = +\infty$ the above problem represents the classical LQG problem, whose solution can be obtained via the algebraic Riccati equation [28, p. 372]. By a simple change of coordinates it can be seen that the presented system fulfills Assumptions 1 and 2. Moreover, the following lemma provides the technical parameters required for the proposed error bounds.

Lemma 2 (Truncated LQG properties): The error bounds provided by Theorem 2 hold with the norms $\|c\|_\infty = L^2(q+r)$, $\|c\|_L = 4L^2\sqrt{q^2+r^2}$, and the Lipschitz constant of the kernel is

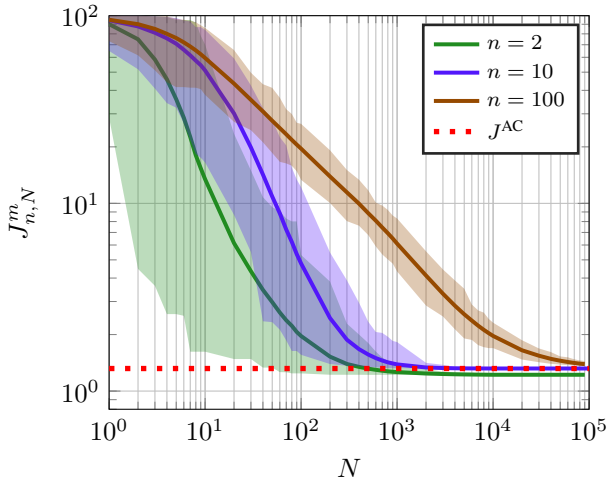
$$L_Q = \frac{2L\max\{\vartheta,\rho\}}{\sigma^2\sqrt{2\pi}\left(\Phi\left(\frac{L-\mu}{\sigma}\right) - \Phi\left(\frac{-L-\mu}{\sigma}\right)\right)}.$$

Proof: In regard to Assumption 1(i), we consider the change of coordinates $\bar{x}_t := \frac{x_t}{2L} + \frac{1}{2}$ and $\bar{a}_t := \frac{a_t}{2L} + \frac{1}{2}$. In the new coordinates, the constants of Lemma 2 follow from a standard computation. ■

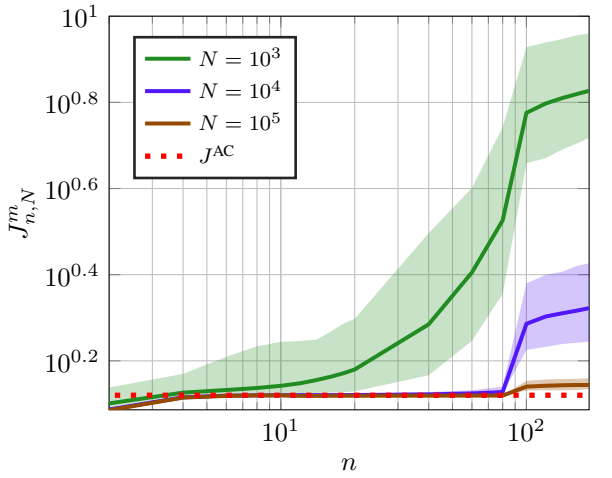
a) Simulation details: For the simulation results we choose the numerical values $\vartheta = 0.8$, $\rho = 0.5$, $\sigma = 1$, $\mu = 0$, $q = 1$, $r = 0.5$, and $L = 10$. Throughout this section we used the Fourier basis $u_{2k-1}(s) = \frac{L}{k\pi}\cos\left(\frac{k\pi s}{L}\right)$ and $u_{2k}(s) = \frac{L}{k\pi}\sin\left(\frac{k\pi s}{L}\right)$ and the uniform distribution on $K = X \times A = [-L, L]^2$ to draw the random samples $\{x_j, a_j\}_{j=1}^N$ in program (4).

b) Simulation results: The simulation results are shown in Figure 1. Figure 1(a) suggests three interesting features concerning n, the number of basis functions: The higher the number of basis functions,

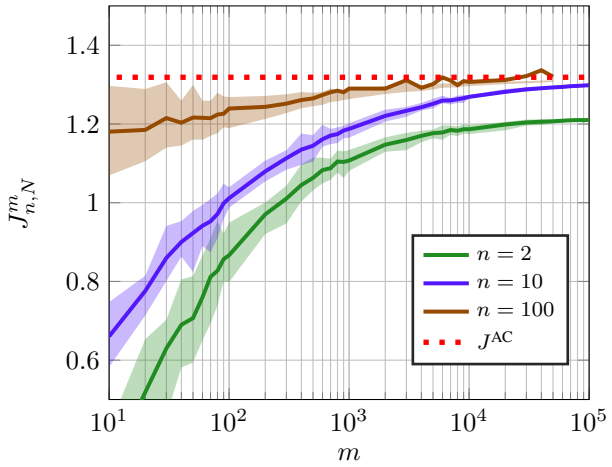
(i) the smaller the approximation error (i.e., asymptotic distance for $N \to \infty$ to the red dotted line),



(a) varying constraint samples N, where $m = 10^6$



(b) varying number of basis functions n, where $m = 10^6$



(c) varying kernel-learning samples m, where $N = 10^3$

Fig. 1. The objective performance $J^m_{n,N}$ is computed according to (4). The colored tubes represent the results between $[10\%, 90\%]$ quantiles (shaded areas) as well as the means (solid lines) across 200 independent experiments of the objective performance $J^m_{n,N}$. The red dotted line denoted by J^{AC} is the optimal solution approximated by $n = 10^3$, $m = 10^6$ and $N = 10^6$.

(ii) the lower the variance of approximation with respect to the sampling distribution for each N, and

(iii) the slower the convergence behavior with respect to the sample size N.

The feature (iii), namely that a high number of basis functions requires a large number of sampled constraints N to produce reasonable approximation errors can also be seen in Figure 1(b). Moreover, the higher the number of sampled constraints N the lower the variance of the approximation. Figure 1(b) suggests that there a sweet spot, namely given a certain number n of basis functions, there is a minimum number of sampled constraints N required for an acceptable approximation accuracy. Finally, Figure 1(c) indicates that the more basis functions n, the less samples from the kernel m are required for $J^m_{n,N}$ to be close to the optimal value.

VI. Conclusion

In this paper we presented an approximation scheme for the infinite-dimensional LP formulation of of discrete-time Markov control processes via a finite-dimensional convex program, in the case the dynamics of the system are unknown and learned from data. We derived a probabilistic explicit error bound between the data-driven finite convex program and the original infinite LP, that is equivalent to the optimal control problem.

For future work, there are several interesting directions. First, even though we discuss the sample complexity of the error bound in this paper, i.e., how many constraint-samples are required for an a priori approximation accuracy, we do not provide any insight in what would be a good distribution to draw these samples from. One would intuitively expect that certain regions of the state-action space are more "informative" than others. Another open question is, given such an approximating scheme, how to synthesize ε-approximating policies, i.e., policies whose corresponding cost is ε away from the optimal value.

References

[1] O. Hernández-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, ser. Applications of Mathematics Series. Springer, 1996.

[2] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, Mar. 1993. [Online]. Available: http://dx.doi.org/10.1137/0331018

[3] O. Hernndez-Lerma, J. Hennet, and J. Lasserre, "Average cost markov decision processes: Optimality conditions," *Journal of Mathematical Analysis and Applications*, vol. 158, no. 2, pp. 396 – 406, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0022247X9190244T

[4] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2007.

[5] D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations Research*, vol. 51, no. 6, pp. 850–865, 2003. [Online]. Available: http://or.journal.informs.org/content/51/6/850.abstract

[6] D. P. De Farias and B. Van Roy, "On constraint sampling in the linear programming approach to approximate dynamic programming," *Mathematics of operations research*, vol. 29, no. 3, pp. 462–478, 2004.

[7] P. Mohajerin Esfahani, T. Sutter, D. Kuhn, and J. Lygeros, "From Infinite to Finite Programs: Explicit Error Bounds with Applications to Approximate Dynamic Programming," *ArXiv e-prints*, Jan. 2017.

[8] D. Bertsekas, "Dynamic programming and suboptimal control: A survey from adp to mpc," *European Journal of Control*, vol. 11, no. 45, pp. 310 – 334, 2005.

[9] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[10] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[11] C. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge,England, 1989.

[12] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003. [Online]. Available: http://dx.doi.org/10.1137/S0363012901385691

[13] A. S. Manne, "Linear programming and sequential decisions," *Management Science*, vol. 6, no. 3, pp. 259–267, 1960. [Online]. Available: http://dx.doi.org/10.1287/mnsc.6.3.259

[14] O. Hernández-Lerma and J. Lasserre, *Further topics on discrete-time Markov control processes*, ser. Applications of Mathematics Series. Springer, 1999.

[15] ——, *Handbook of Markov decision processes: methods and applications*, ser. International Series in Operations Research & Management Science, 40. Kluwer Academic Publishers, 2002, ch. The linear programming approach.

[16] V. Borkar, "A convex analytic approach to markov decision processes," *Probability Theory and Related Fields*, vol. 78, no. 4, pp. 583–602, 1988.

[17] O. Hernández-Lerma, J. González-Hernández, and R. López-Martínez, "Constrained average cost markov control processes in borel spaces," *SIAM Journal on Control and Optimization*, vol. 42, no. 2, pp. 442–468, 2003.

[18] F. Dufour and T. Prieto-Rumeau, "Finite linear programming approximations of constrained discounted markov decision processes," *SIAM Journal on Control and Optimization*, vol. 51, no. 2, pp. 1298–1324, 2013. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/120867925

[19] E. Shafieepoorfard, M. Raginsky, and S. Meyn, "Rational inattention in controlled markov processes," in *American Control Conference (ACC), 2013*, June 2013, pp. 6790–6797.

[20] A. Arapostathis, V. Borkar, E. Fernndez-Gaucherand, M. Ghosh, and S. Marcus, "Discrete-time controlled markov processes with average cost criterion: A survey," *SIAM Journal on Control and Optimization*, vol. 31, no. 2, pp. 282–344, 1993.

[21] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control*. Academic Press, Inc., 1978, vol. 139.

[22] F. Dufour and T. Prieto-Rumeau, "Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities," *Stochastics*, vol. 87, no. 2, pp. 273–307, 2015.

[23] R. T. Farouki, "The Bernstein polynomial basis: A centennial retrospective," *Computer Aided Geometric Design*, vol. 29, no. 6, pp. 379 – 419, 2012.

[24] S. Olver, "On the convergence rate of a modified Fourier series," *Mathematics of Computation*, vol. 78, no. 267, pp. 1629–1645, 2009.

[25] P. Mohajerin Esfahani, T. Sutter, and J. Lygeros, "Performance bounds for the scenario approach and an extension to a class of non-convex programs," *Automatic Control, IEEE Transactions on*, 2014. [Online]. Available: http://dx.doi.org/10.1109/TAC.2014.2330702

[26] A. Nemirovski and A. Shapiro, "Scenario approximations of chance constraints," in *Probabilistic and Randomized Methods for Design under Uncertainty*, G. Calafiore and F. Dabbene, Eds. Springer, 2006, pp. 3–47.

[27] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities*. Oxford University Press, 2013. [Online]. Available: http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001

[28] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II*, 4th ed. Athena Scientific, 2012.