Data-driven dynamic multi-objective optimal control: A Hamiltonian-inequality driven satisficing reinforcement learning approach

Majid Mazouchi [*] **Yongliang Yang** [**] **Hamidreza Modares** [*]

[*] *Michigan State University, Department of Mechanical Engineering, East Lansing, MI 48824 USA (e-mail: Mazouchi, Modaresh@msu.edu)*
[**] *School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 10083, China (e-mail: yangyongliang@ieee.org)*

Abstract: This paper presents an iterative data-driven algorithm for solving dynamic multi-objective (MO) optimal control problems arising in control of nonlinear continuous-time systems with multiple objectives. It is first shown that the Hamiltonian function corresponding to each objective can serve as a comparison function to compare the performance of admissible policies. Relaxed Hamilton-Jacobi-bellman (HJB) equations in terms of HJB inequalities are then solved in a dynamic constrained MO framework to find Pareto-optimal solutions. Relation to satisficing (good enough) decision-making framework is shown. A Sum-of-Square (SOS)-based iterative algorithm is developed to solve the formulated MO optimization with HJB inequalities. To obviate the requirement of complete knowledge of the system dynamics, a data-driven satisficing reinforcement learning approach is proposed to solve the SOS optimization problem in real-time using only the information of the system trajectories measured during a time interval without having full knowledge of the system dynamics. Finally, a simulation example is provided to show the effectiveness of the proposed algorithm.

Keywords: Multi-objective optimization; Pareto optimality; Reinforcement learning; Sum-of-Square theory.

1. INTRODUCTION

In most of the real-world control systems, the system designer must account for multiple objectives (such as safety, control effort, transient performance, comfort, etc.) to evaluate candidate control policies. However, since there usually exist conflicts between objectives (i.e., reaching the best value for one objective needs some reconciliation on other objectives), a control policy is best realized by finding an appropriate context-dependent trade-off among objectives. Multi-objective (MO) optimization has been widely utilized to find a diverse set of efficient solutions, each corresponding to a trade-off alternative, in optimization problems with multiple objectives Marler and Arora (2004); Gambier and Jipp (2011); Peitz and Dellnitz (2018) and Roijers et al. (2013). The decision-making process is then performed posteriori as the decision-maker identifies the most suitable alternative depending on the context.

There are at least two challenges in control of dynamical systems with multiple objectives that are not well addressed in most MO optimization approaches. First, most of the existing MO optimization frameworks are built from the premise that the objective functions to be optimized are static. In the control engineering systems, however, we are dealing with dynamical MO functions Toivonen (1986); Toivonen and Makila (1989); Logist et al. (2010); Ober-Blobaum et al. (2012) and utilizing static optimization frameworks for control of dynamic systems results in myopic short-sighted decisions that do not possess the capability of proactively responding to changes and uncertainties and adapting to novel scenarios. Second, the next-generation autonomous systems such as self-driving cars must autonomously and without any human intervention decide on a suitable trade-off between objectives.

Reinforcement Learning (RL) techniques have been used to solve optimal control problems for system with uncertain dynamics. Most of existing RL algorithms are presented to solve single-objective optimal control problems Lewis and Vrabie (2009); Modares et al. (2016); Jiang and Jiang (2015); Kamalapurkar et al. (2018). Recently, there has been a surge of interest in the study of MO Reinforcement Learning (MORL) problems Kang and Bien (2004); Logist et al. (2010); Caramia and Dell'Olmo (2008); Lopez and Lewis (2019). Most of existing MORL algorithms assume a given preference and find a single best policy corresponding to it based on the weighted sum of the objective functions. However, to successfully operate in a changing and uncertain environment, inspired by human cognitive psychology experiments, which indicates that humans can learn multiple potential solutions for different situational objectives and apply only one at a time, it is desired to learn multiple optimal solutions. Once a diverse set of solutions is found, as mission scenario develops, the system must then decide, without a priori specification of preferences, which policy provides an appropriate trade-off. A higher level of decision-making can decide on the preferences and the most relevant calculated optimal solution can be used as a warm start to avoid learning from scratch in a novel scenario. One might argue that solving several optimal control problems for a diverse set of preferences using a weighted sum of objectives can produce diverse solutions. However, these methods cannot learn control policies in the nonconvex parts of the Pareto optimal set Das and Dennis (1997); Caramia and Dell'Olmo (2008). Moreover, since different objectives have different physical meanings and units, their scales are incomparable and the weighted-sum approach cannot capture the aspiration level (i.e., level of satisfaction) of each objective function for each context.

An interesting solution approach to MO static optimization is the ε-constraint method Carmichael (1980), for which the MO optimization problem is converted into a single-objective constrained optimization problem. That is, it converts all objectives, except for the one with the highest preference, to constraints by imposing some satisficing bounds on them. Doing so, it can deal efficiently with MO problems with nonconvex Pareto fronts. To our knowledge, no attention has been paid to solving dynamic MO optimal control problems using ε-constraint method. This method, however, can lay the cornerstone to solving MO optimal control problems for which there is no a priori preferences and preferences and objective's aspirations can change as the mission developed. The aspiration level of each objective can be explicitly specified by the bounds imposed on long-horizon objectives.

The main motivation of this paper is to develop a novel satisficing control framework that enables us to find a diverse set of solutions to a MO optimal control problem, without knowing the complete knowledge about the system dynamics. A challenge in solving dynamic MO problems is that one needs to find a control policy that satisfies a number of Hamilton-Jacobi Bellman (HJB) equations Lopez and Lewis (2019), which is hard or even impossible due to conflict between objectives. As shown in this paper, a control policy, however, can simultaneously satisfy several Hamiltonian inequalities encoding the objective functions and their aspiration levels. To this end, it is first shown that the Hamiltonian function corresponding to each objective can serve as a comparison function to compare the performance of admissible policies. Using this fact, the MO optimal control problem is formulated as a dynamic ε-constraint problem with relaxed HJB equations as constraints. This formulation can be interpreted as a satisficing MO decision-making framework, for which, instead of optimizing some objective functions, an aspiration level is set for them. An SOS-based iterative algorithm is then developed to find a finite number of solutions of MO optimal control problems with relaxed HJB equations offline. This SOS-based iterative algorithm needs knowledge of the system dynamic, which may not be available. To obviate the requirement of complete knowledge of the system dynamics, a data-driven reinforcement learning method is proposed for finding a Pareto optimal solution using only the information of the system trajectories measured during a time interval online in real-time.

Notations: The following notations are needed throughout the paper. Let $\Re^n$ and $\Re^{n\times m}$ denote the n dimensional real vector space, and the $n \times m$ real matrix space, respectively. Let $\mathbb{Z}^+$ and $\Re^+$ denote the sets of all positive integers and real numbers, respectively. The set of all continuously differentiable polynomial functions is denoted by C^1. $\mathcal{P}$ denotes the set of all positive definite and proper polynomial functions in C^1. Let $0_k \in \Re^k$ be the vector with all zeros and $1_k \in \Re^k$ the vector with all ones. Assume that $y^1, y^2 \in \Re^m$. Then, $y^1 \leq y^2$ denotes weak componentwise order which implies $y^1_k \leq y^2_k$, $k = 1, ..., m$. $y^1 \prec y^2$ denotes Pareto order, which implies $y^1_k \leq y^2_k$, $k = 1, ..., m$, $y^1 \neq y^2$. $y^1 \nprec y^2$ denotes that y^1 is not Pareto dominated by y^2. Assume that $d_1, d_2 \in \mathbb{Z}^+$, and $d_2 \geq d_1$, then $\overrightarrow{m}^{(d_1,d_2)}(x) \in \Re^{\theta n}$ is the arranged in lexicographic order vector of distinct monic monomials in terms of $x \in \Re^n$ with degree κ where $\theta := \begin{pmatrix} n+d_2 \\ d_2 \end{pmatrix} - \begin{pmatrix} n+d_1-1 \\ d_1-1 \end{pmatrix}$ and $d_1 \leq \kappa \leq d_2$. Moreover, the set of all polynomials in $x \in \Re^n$ with degree κ is denoted by $\mathcal{R}[x]_{d_1,d_2}$.

2. PROBLEM FORMULATION

Consider the following continues-time nonlinear system

$$\dot{x} = f(x) + g(x)u \tag{1}$$

where $x \in \Re^n$ and $u \in \Re^m$ are the state and control input of the system, respectively. In this work, we assume that $f(.) : \Re^n \rightarrow \Re^n$ and $g(.) : \Re^n \rightarrow \Re^{n\times m}$ are polynomial mappings and $f(0) = 0$.

For simplicity, throughout the paper, we assume the system has only two objectives. The proposed approach, however, can be readily extended to more than two objectives. The two cost or objective functions associated with the system (1) are defined as

$$J_i\left(x,u\right) = \int\limits_0^\infty r_i\left(x\left(t\right), u(x)\right) dt, i = 1, 2, \tag{2}$$

where $r_i(x, u) = Q_i(x) + u^T R_i(x)u$, with $Q_i(x) \geq 0$ as the penalty on the states, and $R_i \in \Re^{m\times m}$ as a symmetric positive definite matrix.

Definition 1. A control policy $u = \mu(x)$ is said to be admissible with respect to the cost functions $J_i(.)$, $i = 1, 2$, if it is continuous, $\mu(0) = 0$, and it globally stabilizes the dynamics (1) and makes $J_i(.)$, $i = 1, 2$ finite. The set of admissible policies is denoted by Φ in this paper.

Define the value function for a control policy $u \in \Phi$ as

$$V_i(x(t)) = \int\limits_t^\infty r_i(x(\tau),\ u)d\tau,\ i = 1, 2, \tag{3}$$

where $V_i(x(\infty)) = 0$.

Next, for an associated admissible policy $u \in \Phi$, define the Hamilton functions corresponding to the value functions (3) as

$$\mathcal{H}_i\ (x,\ u, V_i) = Q_i(x) + u^T R_i(x)u + \nabla V_i^T\ (f(x) + g(x)u), \tag{4}$$

for $i = 1, 2$, where ∇V_j is the gradient of V_j.

Definition 2. For the system (1) with multiple objectives given by (2), a control solution u^1 is said to dominate a control solution u^2 in a Pareto sense, if and only if $V_i(u^1) \leq V_i(u^2)$, $\forall i \in \{1, 2\}$ and $V_i(u^1) < V_i(u^2)$, $\exists i \in \{1, 2\}$.

Problem 1. Consider the nonlinear system (1). Design an admissible control policy $u(x) \in \Phi$ that minimizes the cost functions (2) in a Pareto sense.

Minimizing each cost function independently while ignoring the other cost functions can be performed using standard optimal control techniques Lewis and Vrabie (2009). However, for dynamic MO optimal control, it is rarely possible to design a controller that optimizes all objective function simultaneously and independently. Therefore, normally, utopian point, i.e., $J^{utopian} := [J_1^{utopian}\ J_2^{utopian}]^T$ where $J_i^{utopian} \leq J_i\left(x\left(0\right), u\right)$, $\forall x \in \Re^n$, $\forall u \in \Re^m$, $\forall i = 1, 2$, is unattainable. However, it is of great importance to find solutions that are as close as possible to a utopian point. Such solutions are called Pareto optimal solutions.

3. A HAMILTONIAN-DRIVEN SATISFICING MO OPTIMAL CONTROL FRAMEWORK

In this section, it is shown that the Hamiltonian function corresponding to each objective can serve as a comparison function to compare the performance of admissible policies in a Pareto sense. The following theorem shows that minimizing one objective function while converting the other objective as a constraint resembles the satisficing (good enough) decision making framework for which the constraint bound is an indication of the aspiration level (the level of satisfaction) of the objective 2.

Theorem 1. Let $u^j(.)$, $j = 1, 2$ be two different admissible policies, with their value function vectors given as $V^j(x) = [\,V_1^j(x)\ V_2^j(x)\,]^T$, $j = 1, 2$, being the solution to (4), i.e., $\mathcal{H}_i\ (.) = 0$. Consider now the following conditioned dynamic optimization problem

$$\bar{u}^j := \arg\min \mathcal{H}_1(x, u(.),\ V_1^j) \tag{5}$$

$$s.t. \qquad -\,\delta^j(x) \leq \mathcal{H}_2(x, u(.),\ V_2^j) \leq 0 \tag{6}$$

with $\delta^j(x) > 0$ as the aspiration for objective 2. Let also $\mathcal{H}_{\min}^j := \left[\,\mathcal{H}_1^j\ \mathcal{H}_2^j\,\right]^T$ where $\mathcal{H}_1^j := \mathcal{H}_1(x, \bar{u}^j(x),\ V_1^j)$ and $\mathcal{H}_2^j := \mathcal{H}_2(x, \bar{u}^j(x),\ V_2^j)$. Then, the following properties hold, $\forall x \in \Re^n$.
1) $\mathcal{H}_{\min}^j \leq 0_2$, $j = 1, 2$.
2) If $-\,\delta^j(x) \leq \mathcal{H}_2^j \leq 0, j = 1, 2$, and $\mathcal{H}_1^1 < \mathcal{H}_1^2$, then $V_1^2 < V_1^1$ and consequently $V^1 \not\prec V^2$, $\forall x \in \Re^n$.
3) If $\delta^2(x) < \delta^1(x)$ and $\mathcal{H}_1^1 < \mathcal{H}_1^2$, then $V^1 \not\prec V^2$ and $V^2 \not\prec V^1$.

Proof. The proof has three parts. It follows from (5)-(6) that $-\,\delta^j(x) \leq \mathcal{H}_2^j = \mathcal{H}_2(x, \bar{u}^j(x),\ V_2^j) \leq 0$ and $\mathcal{H}_1^j = \mathcal{H}_1(x, \bar{u}^j(.),\ V_1^j) \leq \mathcal{H}_1(x, u^j(.),\ V_1^j) = 0$, $\forall j = 1, 2$. This proves part 1. We now prove part 2. Let $V_1^2(x) = V_1^1(x) + \Lambda(x)$ for some $\Lambda(x) > 0$. Based on the Hamiltonian (4) for $V_1^1(.)$ and the stationary condition Lewis et al. (2012), one has

$$\begin{aligned}\mathcal{H}_1^2 &= Q_1(x) + \nabla V_1^{2T}(.) f(x) - \tfrac{1}{4}\nabla {V_1^2}^T g(x) R_1^{-1} g^T(x) \nabla V_1^2 \\ &= \ \mathcal{H}_1^1 + \tfrac{d\Lambda}{dt} - \tfrac{1}{4}\nabla {\Lambda_1}^T g(x) R_1^{-1} g^T(x) \nabla \Lambda_1 \end{aligned} \tag{7}$$

After some manipulation, (7) can be rewritten as

$$\tfrac{d\Lambda}{dt} = \mathcal{H}_1^2 - \mathcal{H}_1^1 + \tfrac{1}{4}\nabla {\Lambda_1}^T g(x) R_1^{-1} g^T(x) \nabla \Lambda_1 \tag{8}$$

If $\mathcal{H}_1^2 - \mathcal{H}_1^1 \geq 0$, (8) implies that $d\Lambda/dt \geq 0$. Based on (3), $\Lambda(x(\infty)) = 0$, so (8) implies that $\Lambda(x) \leq 0, \forall x \in \Re^n$. Thus, $\mathcal{H}_1^1 < \mathcal{H}_1^2$ implies that $V_1^2 < V_1^1$ and consequently $V^1 \not\prec V^2$, $\forall x \in \Re^n$. This completes the proof of part 2. To prove part 3, considering the inequality condition (6), the Lagrangian is $\Gamma^j = \mathcal{H}_1(x, \bar{u}^j(x), V_1^j(x)) + \lambda_{12}^j[-\mathcal{H}_2(x, \bar{u}^j(x), V_2^j(x)) - \delta^j(x)]$ where λ_{12}^j is Lagrange multiplier. Provided that $\delta^1(x)$ and $\delta^2(x)$ are sufficiently small, from the Kuhn-Tucker condition Lewis et al. (2012), one can see that constraint (6) will be active, i.e., $\mathcal{H}_2(x, \bar{u}^j(x), V_2^j) = -\delta^j(x)$, $\lambda_{12}^j > 0$. Moreover, $\lambda_{12}^j = -\partial H_1(x, \bar{u}^j(x), V_1^j(x))/\partial H_2(x, \bar{u}^j(x), V_2^j(x))$ which based on property 2 indicates that an improvement in $\mathcal{H}_1(x, \bar{u}^j(x), V_1^j(x))$ may only be obtained at the cost of degradation in $\mathcal{H}_2(x, \bar{u}^j(x), V_2^j(x))$. Therefore, the inequality condition (6) is active, i.e., $\mathcal{H}_2^j = \mathcal{H}_2(x, \bar{u}^j(.), V_2^j) = -\,\delta^j(x)$ for $j = 1, 2$. Thus, using property 2, $\delta^2(x) < \delta^1(x)$ implies that $\mathcal{H}_2^2 > \mathcal{H}_2^1$ which implies that $V_2^1 < V_2^2$ and consequently $V^2 \not\prec V^1$, $\forall x \in \Re^n$. Moreover, from property 2, one has $\mathcal{H}_1^1 < \mathcal{H}_1^2$ implies that $V_1^2 < V_1^1$ and consequently $V^1 \not\prec V^2$. This completes the proof. □

Remark 1. Theorem 1 implies that active constraint correspond to Pareto optimal solutions. Therefore, by tightening or loosing the aspiration level, i.e., δ^j, one can find different Pareto optimal solutions on the Pareto front, each corresponding on different demands on the objective function 2. The desired aspiration level might depend on the circumstance the system is encountering. Using this sense, in the next section, the problem in hand will be formulated as an ε-constraint problem with relaxed HJB equations as constraints. This framework allows to find variety of solutions, each for different circumstances, and as the mission scenario develops, apply the appropriate solution without calculating it from scratch.

4. MULTI-OBJECTIVE SUBOPTIMAL CONTROL WITH RELAXED HJB EQUATION

In this section, we formulate Problem 1 as an ε-constraint problem with relaxed HJB equations as constraints. To this end, MO optimal control Problem 1 can be reformulated as the following ε-constraint problem.

Problem 2. Consider the nonlinear system (1) associated with the cost functions (2). Design the control policy $u(x)$, to solve the following constrained minimization problem (9)-(12).

$$\min_{V_1}\ \textstyle\int_\Omega V_1(x)dx \tag{9}$$

$$s.t. \quad \mathcal{H}_1(x, u(.), V_1) \leq 0 \tag{10}$$

$$-\delta \leq \mathcal{H}_2(x, u(.),\ V_2) \leq 0 \tag{11}$$

$$V_i \in \mathcal{P},\ i = 1, 2 \tag{12}$$

where $\delta > 0$ is a variable that implicitly indicates the aspiration on optimizing objective V_2. Moreover, $\Omega \in \Re^n$ is an arbitrary closed compact set containing the origin that describes the region in which the objective function $V_1(x)$ is expected to be minimized the most.

Remark 2. Based on (4), (10) implies that the closed-loop system (1) converges to the origin. Moreover, based on Theorem 1, (9) -(11) are equivalent to (5)-(6) which indicates that the cost functions (2) are minimized in a Pareto sense.

Assumption 1. Consider the nonlinear system (1). There exist feedback control policy $u(.)$ and functions $V_{01}(u(.)) \in$

$\mathcal{P}$ and $V_{02}(u(.)) \in \mathcal{P}$, and δ such that

$$\mathcal{L}_2(V_{02}(.), u(V_{01}(.))) \leq \delta, \forall x \in \Re^n \tag{13}$$

where, for any $V_i \in C^1$ and $u \in \Phi$

$$\begin{aligned}\mathcal{L}_i(V_i, u) &= -\nabla V_i^T(x)(f(x) + g(x)u) - r_i(x, u), \ i = 1, 2 \\ &= -\mathcal{H}_i(x, u; V_i) \end{aligned} \tag{14}$$

Theorem 1. Let $V_{01} \in \mathcal{P}$ and its corresponding control policy $u_{01} := \bar{u}(V_{01})$ be the solution to (4). Let Assumption 1 hold for the cost function $V_{01}(u_{01}(.)) \in \mathcal{P}$ and $V_{02}(u_{01}(.)) \in \mathcal{P}$, and control policy u_{01}. For a fixed δ, the following hold. 1) The constrained optimization Problem 2 has a nonempty feasible set. 2) Let $\bar{V}_1(\bar{u}_1(.)) \in \mathcal{P}$ and $\bar{V}_2(\bar{u}_1(.)) \in \mathcal{P}$ be a feasible solution to the relaxed constrained optimization Problem 2. Then, the control policy

$$\bar{u}_1(.) = -\tfrac{1}{2} R_1^{-1} g^T(x) \nabla \bar{V}_1 \tag{15}$$

is globally stabilizing.

Proof. The proof is omitted due to the limited space. □

5. SOS-BASED MULTI-OBJECTIVE CONTROL

In this section, a novel iterative method is developed to find the solution of Problem 2 and accordingly Problems 1 based on the SOS-based methods Ahmadi (2018). To do so, the following definition is needed.

Definition 3. A polynomial $p(x)$ is an SOS polynomial, i.e., $p(x) \in \mathcal{P}^{SOS}$ where $\mathcal{P}^{SOS}$ is a set of SOS polynomial, if $p(x) = \sum_1^m p_i^2(x)$ where $p_i(x) \in \mathcal{P}$, $i = 1, ..., m$.

Let $V_i(x) = \sum_{j=1}^{N} c_{ij} m_{1j}(x) = {C_i}^T \overrightarrow{m}_i^{(2,2d)}(x)$, $i = 1, 2$ where $m_{ij}(x)$, $i = 1, 2$ are predefined monomials in x and c_{ij}, $i = 1, 2$ are coefficients to be determined. Denote $V_i^k(x) := {C_i^k}^T \overrightarrow{m}_i^{(2,2d)}(x)$, $i = 1, 2$.

Assumption 2. For system (1), there exist polynomial functions $V_{01}(u_1(.))$ and $V_{02}(u_1(.))$ and control policy $u_1(.)$ such that $V_{0i}(u_1(.)) \in \mathcal{R}[x]_{2,2d} \cap \mathcal{P}^{SOS}$, $\mathcal{L}_i(V_{0i}(.), u(V_{01}(.))) \in \mathcal{P}^{SOS}$, and $\delta^r - \mathcal{L}_2(V_{02}(.), u(V_{01}(.))) \in \mathcal{P}^{SOS}$, $i = 1, 2$ where $\delta^r > 0$.

Motivated by the work in Jiang and Jiang (2015) Algorithm 1 is given to find the solution of Problem 2.

Theorem 2. Assume that Assumptions 1-2 hold. Then, for a fixed aspiration level δ^r, the following properties hold.
1) The SOS program (16)-(20) has at least one feasible solution;
2) The control policy $u^{(k+1)}(x)$ is globally asymptotically stabilizing the system (1) at the origin;
3) $0 \leq V_1^{k+1} \leq V_1^k$, $\forall k$, where $V_1^k \in \mathcal{P}^{SOS}$;
4) The sequence $\{V_1^k\} \in \mathcal{P}^{SOS}$ is convergent, i.e., $V_1^{\delta^r} := \lim_{k \to \infty} V_1^k \geq V_1^*$.

Proof. The proof is omitted due to the limited space. □

6. DATA-DRIVEN REINFORCEMENT LEARNING IMPLEMENTATION

In this section, a data-driven learning algorithm is developed to implement Algorithm 2 without having the full knowledge of the system dynamics.

Algorithm 1: Relaxed MO SOS-program.

```
1: procedure
2:     Start with $\{V_1^0(.), V_2^0(.), u^{(0)}\}$ that satisfy Assumption 2 and set $r = 1$.
3:     For $k = 1, 2, ...$, if there is a feasible solution then solve the following SOS program:
           $\min_{C_1, K_{C_1}} \int_\Omega V_1(.) dx$    (16)
           $s.t. \quad \mathcal{L}_i(u(V_1), V_i(.)) \in \mathcal{P}^{SOS}, \ i = 1, 2$    (17)
           $\delta^r(x) - \mathcal{L}_2(u(V_1), V_2(u(V_1)) \ \in \mathcal{P}^{SOS},$    (18)
           $V_1^{k-1} - V_1 \ \in \mathcal{P}^{SOS},$    (19)
           $V_i \ \in \mathcal{P}^{SOS}, i = 1, 2,$    (20)
       where $V_i(x) := C_i^T \overrightarrow{m}_i^{(2,2d)}(x)$, $V_i^k(x) := {C_i^k}^T \overrightarrow{m}_i^{(2,2d)}(x)$, $i = 1, 2$, $u(V_1) = K_{C_1} \overrightarrow{m}_1^{(1,\bar{d}^r)}$, $u^{(k)}(V_1^k) = K_{C_1}^k \overrightarrow{m}_1^{(1,\bar{d}^r)}$ and $\delta^1(x)$ is a predefined aspiration level.
4:     If convergence is achieved, or if there is no more feasible solution $u_r^* = u(V_1)$, $U^* = U^* \cup \{u_r^*\}$ where $U^*$ is the set of efficient control policies and go to Step 5 else go back to Step 2 with $k = k + 1$.
5:     Set $r = r + 1$, $\delta^{r+1}(x) = \upsilon \delta^r(x)$, where $0 < \upsilon < 1$ is predefined design parameter go to Step 2.
6: end procedure
```

Now, consider the system (1), after adding an exploratory probing noise, one has

$$\dot{x} = f + g(u^{k+1} + e) \tag{21}$$

where u^{k+1} is a control policy at iteration $k + 1$ and e is an added bounded exploration probing noise.

In the SOS-based Optimization Algorithm 1, under Assumption 2, one has $\forall k, r$, $\mathcal{L}_i(u^k, V_i^k(x, u^k)) \in \mathcal{R}[x]_{2,2\bar{d}^r}$, $i = 1, 2$, $\delta^r - \mathcal{L}_2(u^k, V_2^k(x, u^k)) \in \mathcal{R}[x]_{2,2\bar{d}^r}$, where $\delta^r \in \Re^+$, if the integer $\bar{d}^r$ satisfies

$$\begin{aligned}\bar{d}^r \geq \frac{1}{2} \max\{&\deg(f(.)) + 2d - 1, 2\deg(g(.)) \\ &+2(2d - 1), \deg(Q_1(.)) + \deg(Q_2(.)), \deg(\delta^r(.))\}\end{aligned} \tag{22}$$

where deg(.) represents the degree of the polynomial which is the highest degree of any of the terms. Also, u^{k+1} obtained from the proposed SOS-based Optimization Algorithm 1 satisfies $u^{k+1} \in \mathcal{R}[x]_{1,\bar{d}^r}$, $\forall k, r$.

Hence, there exists a constant matrix $K^{k+1} \in \Re^{m \times n_{\bar{d}^r}}$, with $n_{\bar{d}^r} = \begin{pmatrix} n + \bar{d}^r \\ \bar{d}^r \end{pmatrix} - 1$, such that $u^{k+1} = K^{k+1} \overrightarrow{m}_1^{(1,\bar{d}^r)}$. Also, suppose there exist constant vectors $C_1 \in \Re^{n_{2d}}$ and $C_2 \in \Re^{n_{2d}}$, with $n_{2d} = \begin{pmatrix} n + 2d \\ 2d \end{pmatrix} - n - 1$, such that $V_1(x) := C_1^T \overrightarrow{m}_1^{(2,2d)}(x)$ and $V_2(x) := C_2^T \overrightarrow{m}_2^{(2,2d)}(x)$. It follows then from (21) that

$$\dot{V}_1 = -r_1(x, u^{k+1}) - \mathcal{L}_1(u^{k+1}, V_1(x, u^{k+1})) + (R_1^{-1} g^T \nabla V_1)^T R_1 e \tag{23}$$

$$\dot{V}_2 = -r_2(x, u^{k+1}) - \mathcal{L}_2(u^{k+1}, V_2(x, u^{k+1})) + \nabla V_2^T \nabla V_1^{-T} (R_1^{-1} g^T \nabla V_1)^T R_1 e \tag{24}$$

Notice that the terms $\mathcal{L}_1(u^{k+1}, V_1(x, u^{k+1}))$, $\mathcal{L}_2(u^{k+1}, V_2(x, u^{k+1}))$, $R_1^{-1} g^T \nabla V_1$, and $\nabla V_2^T \nabla V_1^{-T} (R_1^{-1} g^T \nabla V_1)^T R_1 e$ depend on the dynamic of the system. Also, note that constant vectors and matrix $l_{C_1} \in \Re^{n_{2\bar{d}^r}}$ and $l_{C_2} \in \Re^{n_{2\bar{d}^r}}$, and $K_{C_1} \in \Re^{m \times n_{\bar{d}^r}}$ with $\bar{d}^r = \begin{pmatrix} n + 2\bar{d}^r \\ 2\bar{d}^r \end{pmatrix} - \bar{d}^r - 1$ for the tuple (V_1, V_2, u^{k+1}) can be chosen such that:

$$\mathcal{L}_i(u^{k+1}, V_i(x, u^{k+1})) = {l_{C_i}}^T \overrightarrow{m}_i^{(2,2\bar{d}_r)}(x), \; i = 1, 2, \tag{25}$$

$$-\tfrac{1}{2} R_1^{-1} g^T \nabla V_1 = K_{C_1} \overrightarrow{m}_1^{(1,\bar{d}^r)} \tag{26}$$

Therefore, calculating $\mathcal{L}_i(u^{k+1}, V_i(x, u^{k+1}))$, $i = 1, 2$ and $R_1^{-1} g^T \nabla V_1$ amounts to find l_{C_1}, l_{C_2}, and K_{C_1}.

Substituting (25) and (26) in (23)-(24), we have

$$\dot{V}_1 = -r_1(x, u^{k+1}) - {l_{C_1}}^T \overrightarrow{m}_1^{(2,2\bar{d}_r)}(x) - 2(\overrightarrow{m}_1^{(1,\bar{d}^r)})^T K_{C_1}^T R_1 e \tag{27}$$

$$\begin{aligned}\dot{V}_2 = {} & -r_2(x, u^{k+1}) - {l_{C_2}}^T \overrightarrow{m}_2^{(2,2\bar{d}_r)}(x) - 2(\nabla \overrightarrow{m}_2^{(2,2d)}(x(t)))^T \\ & \times C_2(\overrightarrow{m}_1^{(1,\bar{d}^r)})^T((\nabla \overrightarrow{m}_1^{(2,2d)}(x(t)))^T C_1)^T K_{C_1}^T R_1 e\end{aligned} \tag{28}$$

Integrating both sides of (27)-(28) on the interval $[t, t + \delta t]$ yields the following off-policy integral RL Bellman equations

$$\begin{aligned}& C_1^T(\overrightarrow{m}_1^{(2,2d)}(x(t)) - \overrightarrow{m}_1^{(2,2d)}(x(t + \delta t))) = \\ & \int_t^{t+\delta t} (r_1(x, u^{k+1}) + {l_{C_1}}^T \overrightarrow{m}_1^{(2,2\bar{d}_r)}(x) + 2\left(\overrightarrow{m}_1^{(1,\bar{d}^r)}\right)^T K_{C_1}^T R_1 e) d\tau\end{aligned} \tag{29}$$

$$\begin{aligned}& C_2^T(\overrightarrow{m}_2^{(2,2d)}(x(t)) - \overrightarrow{m}_2^{(2,2d)}(x(t + \delta t))) = \\ & \int_t^{t+\delta t} (r_2(x, u^{k+1}) + {l_{C_2}}^T \overrightarrow{m}_2^{(2,2\bar{d}_r)}(x) + 2(\nabla \overrightarrow{m}_2^{(2,2d)}(x(t)))^T \\ & \quad \times C_2(\overrightarrow{m}_1^{(1,\bar{d}^r)})^T((\nabla \overrightarrow{m}_1^{(2,2d)}(x(t)))^T C_1)^T K_{C_1}^T R_1 e) d\tau\end{aligned} \tag{30}$$

It follows from (29)-(30) that l_{C_1}, l_{C_2}, and K_{C_1} can be found by using only the information of the system trajectories measured during a time interval, without requiring any system dynamic information. To this end, we define the following matrices:

$$\sigma_e^1 = [\, \overrightarrow{m}_1^{(2,2d)} \;\; 2\left(\overrightarrow{m}_1^{(1,\bar{d}^r)}\right)^T \otimes e^T R_1 \,]^T, \tag{31}$$

$$\begin{aligned}\sigma_e^2 = [\, \overrightarrow{m}_2^{(2,2d)} \;\; & 2(\nabla \overrightarrow{m}_2^{(2,2d)}(x(t)))^T C_2\left(\overrightarrow{m}_1^{(1,\bar{d}^r)}\right)^T \times \\ & ((\nabla \overrightarrow{m}_1^{(2,2d)}(x(t)))^T C_1)^T \otimes e^T R_1]^T,\end{aligned} \tag{32}$$

$$\phi_i^{k+1} = [\, \int_{t_{0,k+1}}^{t_{1,k+1}} \sigma_e^i d\tau \;\; \cdots \;\; \int_{t_{q_{k+1}-1,k+1}}^{t_{q_{k+1},k+1}} \sigma_e^i d\tau \,]^T, \tag{33}$$

$$\Xi_i^{k+1} = [\, \int_{t_{0,k+1}}^{t_{1,k+1}} r_i(x, u^{k+1}) d\tau \;\; \cdots \;\; \int_{t_{q_{k+1}-1,k+1}}^{t_{q_{k+1},k+1}} r_i(x, u^{k+1}) d\tau \,]^T \tag{34}$$

$$\theta_i^{k+1} = [\, \overrightarrow{m}_i^{(2,2d)} \Big|_{t_{0,k+1}}^{t_{1,k+1}} \;\; \cdots \;\; \overrightarrow{m}_i^{(2,2d)} \Big|_{t_{q_{k+1}-1,k+1}}^{t_{q_{k+1},k+1}} \,]^T, \tag{35}$$

for $i = 1, 2$, where $\phi_i^{k+1} \in \Re^{q_i^{k+1} \times (n_{2\bar{d}^r} + m n_{\bar{d}^r})}$ and $\Xi_i^{k+1} \in \Re^{q_i^{k+1}}$.

It follows from (29)-(30) that

$$\phi_1^{k+1} \begin{bmatrix} l_{C_1} \\ Vec(K_{C_1}) \end{bmatrix} = \Xi_1^{k+1} + \theta_1^{k+1} C_1 \tag{36}$$

$$\phi_2^{k+1} \begin{bmatrix} l_{C_2} \\ Vec(K_{C_1}) \end{bmatrix} = \Xi_2^{k+1} + \theta_2^{k+1} C_2 \tag{37}$$

Assumption 3. At each iteration k, there exists a lower-bound $q_0^{k+1} \in \mathbb{Z}^+$ such that if q_1^{k+1}, $q_2^{k+1} \geq q_0^{k+1}$ where q_1^{k+1} and q_2^{k+1} are dimensional of vectors Ξ_1^{k+1} and Ξ_2^{k+1}, respectively, then $rank(\phi_1^{k+1}) = n_{2\bar{d}^r} + m n_{\bar{d}^r}$ and $rank(\phi_2^{k+1}) = n_{2\bar{d}^r} + m n_{\bar{d}^r}$.

Algorithm 2: Data-driven relaxed MO SOS-based algorithm.

```
1: procedure
2:     Find the tuple ${V_1^0, V_2^0, u^0}$ such that Assumption 2 be satisfied. Choose $C_1^0$ and $C_2^0$ such that $V_1^0(x) := (C_1^0)^T \overrightarrow{m}_1^{(2,2d)}(x)$ and $V_2^0(x) := (C_2^0)^T \overrightarrow{m}_2^{(2,2d)}(x)$.
3:     Employ $u = u^k + e$ as the input to the system (1), where $e$ is the probing noise and calculate and construct $\Xi_1$, $\Xi_2$, $\theta_1$, and $\theta_2$ as (31)-(35), till $\phi_1$, $\phi_2$ be of full column rank.
4:     Solve the following SOS program to find an optimal solution ${C_1^k, C_2^k, K_{C_1}^k}$:
           $\min_{C_1, K_{c_1}} \; (\int_\Omega \overrightarrow{m}_1^{(2,2d)}(x) dx)^T C_1$     (38)
           $s.t. \; \phi_1^{k+1} \begin{bmatrix} l_{C_1} \\ Vec(K_{C_1}) \end{bmatrix} = \Xi_1^{k+1} + \theta_1^{k+1} C_1$     (39)
           $\phi_2^{k+1} \begin{bmatrix} l_{C_2} \\ Vec(K_{C_1}) \end{bmatrix} = \Xi_2^{k+1} + \theta_2^{k+1} C_2$     (40)
           ${l_{C_i}}^T \overrightarrow{m}_i^{(2,2\bar{d}_r)}(x) \in \mathcal{P}^{SOS}, \; i = 1, 2$     (41)
           $\delta^r - {l_{C_2}}^T \overrightarrow{m}_2^{(2,2\bar{d}_r)}(x) \in \mathcal{P}^{SOS},$     (42)
           $(C_1^{k-1} - C_1)^T \overrightarrow{m}_1^{(2,2d)}(x) \in \mathcal{P}^{SOS},$     (43)
5:     Update the value functions and control policy as follows:
           $V_i^k(x) := {C_i^k}^T \overrightarrow{m}_i^{(2,2d)}(x), \; i = 1, 2$     (44)
           $u^{(k+1)}(x) = K_{C_1}^{k+1} \overrightarrow{m}_1^{(1,\bar{d}^r)}$     (45)
6:     If $\left\| C_1^k - C_1^{k-1} \right\| \nleq \gamma$, $\gamma$ which is a predefined threshold, go back to Step 2 with $k = k + 1$ else $u_r^* = u^{(k+1)}(x)$ and go to Step 7.
7: end procedure
```

Now, assume that q_1^{k+1}, $q_2^{k+1} \geq q_0^{k+1}$, $\forall k$. It follows from (36)-(37) that the values of $l_{C_1} \in \Re^{n_{2\bar{d}^r}}$, $l_{C_2} \in \Re^{n_{2\bar{d}^r}}$, and $K_{C_1} \in \Re^{m \times n_{\bar{d}^r}}$ are determined as follows:

$$\begin{cases} \begin{bmatrix} l_{C_1} \\ Vec(K_{C_1}) \end{bmatrix} = ((\phi_1^{k+1})^T \phi_1^{k+1})^{-1} (\phi_1^{k+1})^T (\Xi_1^{k+1} + \theta_1^{k+1} C_1) \\ \begin{bmatrix} l_{C_2} \\ Vec(K_{C_1}) \end{bmatrix} = (\phi_2^{k+1})^T \phi_2^{k+1})^{-1} (\phi_2^{k+1})^T (\Xi_2^{k+1} + \theta_2^{k+1} C_2) \end{cases} \tag{46}$$

So, an iterative SOS-based data-driven learning algorithm is proposed in Algorithm 2 for online implementation of Algorithm 1.

Theorem 3. Assume that Assumptions 1-3 hold. Then, for a fixed δ^r, the following properties hold. 1) There exists at least one feasible solution for the SOS program (38)-(43) and (44)-(45); 2) The control policy $u^{(k+1)}(x)$ (45) is globally asymptotically stabilizing the system (1) at the origin; 3) $0 \leq V_1^{k+1} \leq V_1^k$, $\forall k$, where V_1^k is given in (44); 4) The sequence $\left\{V_1^k\right\}$ is convergent, i.e., $V_1^{\delta^r} = \lim_{k \to \infty} V_1^k \geq V_1^*$, where V_1^k is given in (44).

Proof. Provided that $\{C_1^k, C_2^k\}$ is a feasible solution to the relaxed MO SOS-program (16)-(20), one can find the

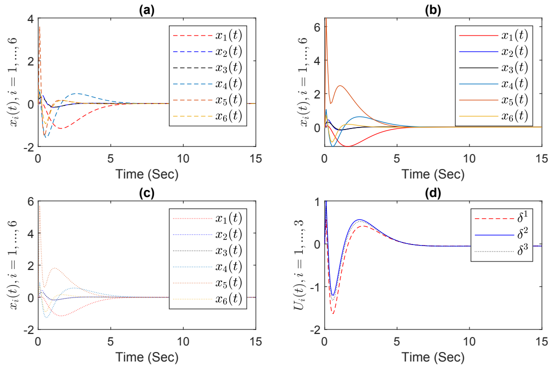

Fig. 1. Comparison of the system state trajectories and control policies for three aspiration levels $\delta^i, i = 1, 2, 3$

corresponding matrix $K^k_{C_1} \in \Re^{m \times n_{\bar{d}^r}}$ such that the tuple $\{C^k_1, C^k_2, K^k_{C_1}\}$ be a feasible solution to the data-driven relaxed MO SOS-program (38)-(43) and (44)-(45), which imply that property 1 holds. Moreover, since the tuple $\{C^k_1, C^k_2, K^k_{C_1}\}$ is a feasible solution to the data-driven relaxed MO SOS-program (38)-(43) and (44)-(45) and the tuple $\{C^k_1, C^k_2\}$ is a feasible solution to the relaxed MO SOS-program (16)-(20) and Algorithms 1 and 2 have the equal objective function, $K^{k+1}_{C_1}\overrightarrow{m}_1^{(1,\bar{d}^r)}$ is an optimal solution to the relaxed MO SOS-program (16)-(20) and consequently the results of Theorem 3 are further extended to Theorem 4. This completes the proof. □

7. SIMULATION

Consider the linearized double inverted pendulum in a cart, with dynamics used in Lopez and Lewis (2019). The quadratic cost functions chosen as $Q_1 = I_6$, $Q_2 = 200 * I_6$, and $R_1 = R_2 = 1$. After the implementation of Algorithm 2 with three different aspiration levels as $\delta^i = \delta^r(0.2x_1^2x_3 + 0.1x_2^2x_5 + 0.25x_4^2 + 0.2x_2x_4x_6 + 0.5x_5x_6 + 0.7x_1^2x_4 + 0.2x_5^2 + 0.1x_6x_2^2 + 0.5x_4x_5x_6 + 0.2x_1x_2x_3)$, $i = 1, 2, 3$ with $\delta^r \in \{0.001, 0.14, 2\}$, three suboptimal control policies are obtained. To save space, the obtained control policies will not be shown here. Fig. 1 shows the evolution of the system states after applying the obtained policies. It can be seen in Fig. 1 that by changing the aspiration level on second objective the obtained control policies and corresponding system states are changed. That is, the trade-off between regulation error and control effort are changed by changing the aspiration level on second objective.

8. CONCLUSION

This paper has developed an iterative data-driven adaptive dynamic programming (ADP) algorithm for dynamic multi-objective (MO) optimal control problem for nonlinear continues-time polynomial systems. The MO optimal control problem was, first, formulated as a dynamic ε-constraint MO problem with relaxed Hamilton-Jacobi-bellman (HJB) equations as constraints. To deal with this problem, then, a Sum-of-Square (SOS)-based iterative algorithm was presented to find some Pareto optimal solutions of MO optimal control problem with relaxed HJB Equations. This SOS-based iterative algorithm required the knowledge of the system dynamic. To obviate the requirement of complete knowledge of the system dynamics, an online data-driven reinforcement learning method was proposed for online implementation of the proposed SOS-based algorithm. Finally, a simulation example was provided to show the effectiveness of the proposed algorithm.

REFERENCES

Ahmadi, A. (2018). Sum of Squares (SOS) Techniques: An Introduction. 1–9.

Carmichael, D. G. (1980). Computation of Pareto optima in structural design. Int. J. Numer. Methods Eng., 15 (6), 925-929.

Caramia, M., and Dell'Olmo, P. (2008). Multi-objective optimization. in Multi-Objective Management in Freight Logistics. Increasing Capacity, Service Level and Safety with Optimization Algorithms. London, U.K. Springer.

Das, I., and Dennis, J. E. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. Struct. Optim., 14 (1), 63-69.

Gambier, A. and Jipp, M. (2011). Multi-objective optimal control: An introduction. ASCC 2011 - 8th Asian Control Conf. - Final Progr. Proc., 1084-1089.

Jiang, Y. and Jiang, Z. (2015). Global Adaptive Dynamic Programming for Continuous-Time Nonlinear Systems. in IEEE Transactions on Automatic Control, 60(11). 2917-2929.

Kamalapurkar, R., Walters, P., Rosenfeld, J., and Dixon, W. (2018). Reinforcement Learning for Optimal Feedback Control. Cham: Springer International Publishing.

Kang, D. O., and Bien, Z. (2004). Multi-objective control problems by reinforcement learning. in Handbook of Learning and Approximate Dynamic Programming, 433-461.

Lewis, Frank L., and Draguna Vrabie. (2009). Adaptive Dynamic Programming for Feedback Control. Proceedings of 2009 7th Asian Control Conference, ASCC 2009, 1402-9.

Lewis, F. L., Vrabie, D. L., and Syrmos, V. L. (2012). Optimal Control: Third Edition. Hoboken, NJ, USA: John Wiley and Sons, Inc.

Logist, F., Sager, S., Kirches, C., and Van Impe, J. F. (2010). Efficient multiple objective optimal control of dynamic systems with integer controls. J. Process Control, 20 (7), 810-822.

Lopez, V. G., and Lewis, F. L. (2019). Dynamic Multi-objective Control for Continuous-Time Systems Using Reinforcement Learning. IEEE Trans. Automat. Contr., 64(7), 2869-2874.

Marler, R. T., and Arora, J. S. (2004). Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization, 26 (6), 369-395.

Modares, H. Lewis, F. L. and Jiang, Z. P. (2016). Optimal Output-Feedback Control of Unknown Continuous-Time Linear Systems Using Off-policy Reinforcement Learning. IEEE Trans. Cybern., 46(11), 2401-2410.

Ober-Blobaum, S., Ringkamp, M., and Zum Felde, G. (2012). Solving multiobjective Optimal Control problems in space mission design using Discrete Mechanics and reference point techniques. Proc. IEEE Conf. Decis. Control, 5711-5716.

Peitz, S. and Dellnitz, M. (2018). A Survey of Recent Trends in Multi-objective Optimal ControlSurrogate Models, Feedback Control and Objective Reduction. Math. Comput. Appl., 23(2), 30.

Roijers, D. M., Vamplew, P., Dazeley, R., Whiteson, S. and Dazeley, R. (2013). A survey of multi-objective sequential decision-making. J. Artif. Intell. Res., 48, 67-113.

Toivonen, H. T. (1986). A primaldual method for linearquadratic gaussian control problems with quadratic constraints. Optim. Control Appl. Methods, 7 (3), 305-314.

Toivonen, H. T., and Makila, P. M. (1989). Computer-aided design procedure for multiobjective LQG control problems. Int. J. Control, 49 (2), 655-666.