



Data-Driven Optimal Control of Affine Systems: A Linear Programming Perspective

Journal Article**Author(s):**

[Martinelli, Andrea](#) ; [Gargiani, Matilde](#); [Draskovic, Marina](#); [Lygeros, John](#) 

Publication date:

2022

Permanent link:

<https://doi.org/10.3929/ethz-b-000557360>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Originally published in:

IEEE Control Systems Letters 6, <https://doi.org/10.1109/lcsys.2022.3180898>

Funding acknowledgement:

787845 - Optimal control at large (EC)

Data-Driven Optimal Control of Affine Systems: A Linear Programming Perspective

Andrea Martinelli, *Graduate Student Member, IEEE*, Matilde Gargiani, Marina Draskovic, and John Lygeros, *Fellow, IEEE*

Abstract—In this letter, we discuss the problem of optimal control for affine systems in the context of data-driven linear programming. First, we introduce a unified framework for the fixed point characterization of the value function, Q-function and relaxed Bellman operators. Then, in a model-free setting, we show how to synthesize and estimate Bellman inequalities from a small but sufficiently rich dataset. To guarantee exploration richness, we complete the extension of Willems’ fundamental lemma to affine systems.

Index Terms—Approximate dynamic programming, data-driven control, affine dynamical systems

I. INTRODUCTION

THE linear programming (LP) approach to optimal control problems was initially developed by A.S. Manne in the 1960s [18], following the well-known studies conducted by R. Bellman in the 1950s [3]. The problem of deriving the fixed point of the Bellman operator can be cast as an LP by exploiting monotonicity and contractivity properties [5]. An advantage of the LP formulation is that there exist efficient and fast algorithms to tackle such programs [8]. On the other hand, similarly to the classic dynamic programming approach introduced by Bellman, the LP approach suffers from poor scalability properties often referred to as *curse of dimensionality* [6]. The sources of intractability for systems with continuous state and action spaces include the optimization variables in infinite dimensional spaces and infinite number of constraints. For this reason, the infinite dimensional LPs are usually approximated by tractable finite dimensional ones [7], [12], [16], [17], [21].

In recent years, the LP approach has experienced an increasing interest, especially in combination with model-free control techniques [1], [20], [22], [23]. In such a setting, one assumes the dynamical system to be unknown but observable via state-input trajectories, and builds one constraint (here called Bellman inequality) of the LP for each observed transition. In this way, one can bypass the classic system identification step and at the same time tackle a source of intractability by solving an LP with finite constraints. Empirical evidence suggests that

the solution quality may dramatically depend on the number of sampled constraints [20]. To avoid massive exploration, one can attempt to generate additional constraints offline from a small but sufficiently rich dataset. A preliminary analysis is conducted in [19] for linear systems in the value function formulation. Another fundamental problem is to estimate the expectation in the Bellman inequalities. A typical approach, *e.g.* in [22] and [20], is to reinitialize the dynamics in the same state-input pairs and compute a Monte-Carlo estimate, even though this procedure could be unrealistic in stochastic settings.

Motivated by the poor scalability often affecting the LP approach and inspired by the recent literature revolving around Willems’ fundamental lemma and data-driven control of affine systems, in the present work we discuss a unified framework to study data-driven control problems arising from this problem class. The authors in [2] show that an augmented state-space formulation allows one to tackle different affine stochastic control problems. The fundamental lemma [28] states that, for controllable *linear* systems, persistency of excitation is a sufficient condition on the control signal to generate a trajectory that contains enough information to express any other trajectory of appropriate length as a linear combination of the input-output data. This result lies at the heart of many recent works on data-driven control of linear systems [10], [14], [26]. Extending the fundamental lemma to affine systems is not trivial, since the collected trajectories no longer form a linear subspace. To do so, we complement the initial result in [4] by a persistency of excitation argument, inspired by the state-space proof of the fundamental lemma described in [25]. Our main contributions can be summarised as follows:

- In Section II, we introduce the stochastic optimal control framework for affine systems including the characterization of the fixed points corresponding to the value function, Q-function and relaxed Bellman operators;
- In Section III, we show how the Bellman inequalities can be reconstructed starting from a sufficiently rich dataset. Moreover, we provide estimators for the corresponding expectations that do not need reinitialization and show how to build LPs that preserve the optimal policy;
- To ensure the dataset is sufficiently rich, in Section IV we extend Willems’ fundamental lemma to affine systems by showing that controllability and persistency of excitation are still sufficient conditions to generate trajectories containing enough information.

Research supported by the European Research Council under the Horizon 2020 Advanced Grant No. 787845 (OCAL).

The authors are with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH) Zurich, 8092 Zurich, Switzerland. (e-mail: andremar@ethz.ch, gmatilde@ethz.ch, mdraskovic@ethz.ch, lygeros@ethz.ch).

Notation. We denote with $\text{TR}(\cdot)$, $\text{SPEC}(\cdot)$, $\text{ROWSP}(\cdot)$, $\text{VEC}(\cdot)$ and $\mathbf{1}$ the trace, spectrum, row space and vectorization of a real matrix and a vector of ones of appropriate dimension. For a subset \mathbb{Y} of a finite dimensional vector space, we denote with $\mathcal{S}(\mathbb{Y})$ the vector space of all real-valued measurable functions $g : \mathbb{Y} \rightarrow \mathbb{R}$ that have a finite weighted sup-norm [5, §2], that is, $\|g(y)\|_{\infty, z} = \sup_{y \in \mathbb{Y}} \frac{|g(y)|}{z(y)} < \infty$ with $z : \mathbb{Y} \rightarrow \mathbb{R}_{>0}$.

II. OPTIMAL CONTROL OF AFFINE SYSTEMS WITH GENERALIZED QUADRATIC STAGE-COST

In this section, we introduce the stochastic optimal control problem for affine systems. We characterize the fixed point of the value function, Q -function and relaxed Bellman operators, discussing how their contraction properties allow one to express the fixed points via infinite-dimensional LPs.

A. Stochastic Optimal Control

Consider the following discrete-time affine dynamics,

$$x^+ = f(x, u, \psi) = Ax + Bu + c + \psi, \quad (1)$$

with possibly infinite state and action spaces $x \in \mathbb{X} \subseteq \mathbb{R}^n$ and $u \in \mathbb{U} \subseteq \mathbb{R}^m$. Here $\psi \in \mathbb{D} \subseteq \mathbb{R}^n$ denotes a random vector with possibly non-zero mean μ and covariance $\Sigma \succeq 0$. Moreover, $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ are such that $(\sqrt{\gamma}A, \sqrt{\gamma}B)$ is stabilizable, $c \in \mathbb{R}^n$ is a constant term and $\gamma \in (0, 1)$ is a discount factor. We define a generalized quadratic stage-cost $\ell : \mathbb{X} \times \mathbb{U} \rightarrow \mathbb{R}_{\geq 0}$ as

$$\ell(x, u) = \begin{bmatrix} x \\ u \end{bmatrix}^\top \underbrace{\begin{bmatrix} L_{xx} & L_{xu} \\ \star & L_{uu} \end{bmatrix}}_L \begin{bmatrix} x \\ u \end{bmatrix} + 2 \begin{bmatrix} x \\ u \end{bmatrix}^\top \underbrace{\begin{bmatrix} L_x \\ L_u \end{bmatrix}}_{L_\ell} + L_c, \quad (2)$$

where the symbol \star is used to denote symmetry. Consider the γ -discounted infinite-horizon cost associated to a deterministic stationary feedback policy $\pi : \mathbb{X} \rightarrow \mathbb{U}$,

$$v_\pi(x) = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k \ell(x_k, \pi(x_k)) \mid x_0 = x \right]. \quad (3)$$

The objective of the optimal control problem is to find an optimal policy π^* such that $v_{\pi^*} = \inf_{\pi} v_\pi = v^*$. Throughout the paper, to ensure that $v^* \in \mathcal{S}(\mathbb{X})$, π^* is measurable and the infimum of v_π is attained, we assume the stage-cost to be lower semi-continuous, nonnegative and inf-compact on $\mathbb{X} \times \mathbb{U}$, and that there exists a policy π such that $v_\pi(x) < \infty$ for each $x \in \mathbb{X}$ [16, Assumptions 4.2.1 and 4.2.2].

B. Value Function Formulation

The optimal value function v^* is generally difficult to compute since, among other issues, it involves the minimization of an infinite sum of costs. However, it allows for the following recursive definition [3],

$$\begin{aligned} v^*(x) &= \inf_{u \in \mathbb{U}} \{ \ell(x, u) + \gamma \mathbb{E}[v^*(f(x, u, \psi))] \} \\ &= (\mathcal{T}v^*)(x), \end{aligned} \quad (4)$$

where $\mathcal{T} : \mathcal{S}(\mathbb{X}) \rightarrow \mathcal{S}(\mathbb{X})$ is known as the *Bellman operator*. \mathcal{T} is a monotone, γ -contractive operator whose unique fixed point is v^* [5], [15].

When the dynamics is linear and the stage-cost quadratic, the resulting optimal control problem (LQR) enjoys well-known closed-form solutions [11] based on the algebraic Riccati equations (ARE) $P^* = q^* - q_\ell^* q_c^{*-1} q_\ell^{*\top}$, where

$$Q^* = \begin{bmatrix} q^* & q_\ell^* \\ \star & q_c^* \end{bmatrix} = \begin{bmatrix} L_{xx} + \gamma A^\top P^* A & L_{xu} + \gamma A^\top P^* B \\ \star & L_{uu} + \gamma B^\top P^* B \end{bmatrix}.$$

In case of affine systems, by suitably augmenting the system's coordinates, it is possible to show that the optimal policy also has an (affine) closed-form [2]. The next result characterizes v^* and π^* for affine systems and generalized quadratic stage-cost, by introducing notation and methods that will be reused in the extensions to Q -function and relaxed Bellman operator.

Proposition 1 *The fixed point of (4) under dynamics (1), stage-cost (2) and $(\mathbb{X}, \mathbb{U}, \mathbb{D}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n)$ is*

$$v^*(x) = x^\top P^* x + 2x^\top P_\ell^* + P_c^* + \frac{\gamma \text{TR}(P^* \Sigma)}{1-\gamma}. \quad (5)$$

$\tilde{P}^* = \begin{bmatrix} P^* & P_\ell^* \\ \star & P_c^* \end{bmatrix}$ is the unique positive definite solution to the following augmented ARE

$$\tilde{P}^* = \tilde{q}^* - \tilde{q}_\ell^* \tilde{q}_c^{*-1} \tilde{q}_\ell^{*\top}, \quad (6)$$

$$\begin{bmatrix} \tilde{q}^* & \tilde{q}_\ell^* \\ \star & \tilde{q}_c^* \end{bmatrix} = \begin{bmatrix} \tilde{L}_{xx} + \gamma \tilde{A}^\top \tilde{P} \tilde{A} & \tilde{L}_{xu} + \gamma \tilde{A}^\top \tilde{P} \tilde{B} \\ \star & L_{uu} + \gamma \tilde{B}^\top \tilde{P} \tilde{B} \end{bmatrix}, \quad (7)$$

$$\tilde{A} = \begin{bmatrix} A & c + \mu \\ 0 & 1 \end{bmatrix}, \tilde{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \tilde{L}_{xx} = \begin{bmatrix} L_{xx} & L_x \\ \star & L_c \end{bmatrix}, \tilde{L}_{xu} = \begin{bmatrix} L_{xu} \\ L_u \end{bmatrix}.$$

Finally, the associated optimal policy is

$$\pi^*(x) = -q_c^{*-1} (q_\ell^{*\top} x + q), \quad (8)$$

where $q = L_u + \gamma B^\top (P_\ell^* + P^*(c + \mu))$.

Proof: Let us consider the constant update $y^+ = y$ initialized at $y_0 = 1$ and the augmented dynamics

$$\tilde{x}^+ = \tilde{A} \tilde{x} + \tilde{B} u + \tilde{\psi}, \quad (9)$$

where $\tilde{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $\tilde{\psi} = \begin{bmatrix} \psi - \mu \\ 0 \end{bmatrix}$. Then, $\mathbb{E}[\tilde{\psi}] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\mathbb{E}[\tilde{\psi} \tilde{\psi}^\top] = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} = \tilde{\Sigma} \succeq 0$. Since $y = 1 \forall t$, the stage-cost can be represented in augmented coordinates as

$$\tilde{\ell}(\tilde{x}, u) = \begin{bmatrix} \tilde{x} \\ u \end{bmatrix}^\top \begin{bmatrix} \tilde{L}_{xx} & \tilde{L}_{xu} \\ \star & L_{uu} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ u \end{bmatrix} = \ell(x, u). \quad (10)$$

Stabilizability of $(\sqrt{\gamma} \tilde{A}, \sqrt{\gamma} \tilde{B})$ follows from stabilizability of the original pair. Indeed, $\text{SPEC}(\tilde{A}) = \text{SPEC}(A) \cup \{1\}$, and asymptotic stability of the uncontrollable mode $y^+ = y$ is guaranteed by the discount factor. Hence, we can formulate a classic LQR problem in augmented coordinates whose unique solution is given by

$$\tilde{v}^*(\tilde{x}) = \tilde{x}^\top \tilde{P}^* \tilde{x} + \frac{\gamma \text{TR}(\tilde{P}^* \tilde{\Sigma})}{1-\gamma} = v^*(x), \quad (11)$$

where \tilde{P}^* is the solution to the augmented ARE (6), and the associated optimal policy is

$$\begin{aligned} \tilde{\pi}^*(\tilde{x}) &= \arg \min_{u \in \mathbb{R}^m} \{ \tilde{\ell}(\tilde{x}, u) + \gamma \mathbb{E}[\tilde{v}^*(\tilde{A} \tilde{x} + \tilde{B} u + \tilde{\psi})] \} \\ &= -\tilde{q}_c^{*-1} \tilde{q}_\ell^{*\top} \tilde{x}. \end{aligned}$$

The claim then follows by writing the solution in the original coordinates. \blacksquare

Note that if the system is linear ($c = 0$), the noise zero-mean ($\mu = 0$) and the stage-cost pure quadratic ($L_\ell = 0$), then $q = 0$ and we recover the linear policy $\pi^*(x) = -q_c^{*-1}q_\ell^{*\top}x$.

The fixed point of \mathcal{T} can be computed via linear programming [12]. By observing that the Bellman inequality $v \leq \mathcal{T}v$ implies $v \leq v^*$, one can characterize the fixed point of \mathcal{T} by looking for the greatest $v \in \mathcal{S}(\mathbb{X})$ that satisfies $v \leq \mathcal{T}v$,

$$\begin{aligned} \sup_{v \in \mathcal{S}(\mathbb{X})} \int_{\mathbb{X}} v(x)c(dx) \\ \text{s.t. } v(x) \leq (\mathcal{T}v)(x) \quad \forall x \in \mathbb{X}, \end{aligned} \quad (12)$$

where $c(\cdot)$ is a positive measure with finite moments. In the LP literature, $c(\cdot)$ is typically selected to be a probability measure [7], [12]. For example, if the state-space is unbounded one can use a Gaussian distribution, if it is compact a uniform distribution. Moreover, the measure $c(\cdot)$ can be used to give different weight in the quality of the approximation of the value function in different parts of the state-space. Now, notice that \mathcal{T} is a nonlinear operator. However, it is possible to reformulate (12) as an equivalent linear program [12] by dropping the infimum in \mathcal{T} and substituting the nonlinear constraint set with the following linear one

$$v(x) \leq \underbrace{\ell(x, u) + \gamma \mathbb{E}[v(f(x, u, \psi))]}_{(\mathcal{T}_\ell v)(x, u)} \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U}.$$

The associated optimal policy can then be computed by

$$\pi^*(x) = \arg \min_{u \in \mathbb{U}} \{ \ell(x, u) + \gamma \mathbb{E}[v^*(f(x, u, \psi))] \}. \quad (13)$$

C. Q-function Formulation

Policy extraction (13) is in general not possible if the dynamics f or the stage-cost ℓ are not known. This difficulty can be overcome by introducing the Bellman operator associated to Q -functions [27], $\mathcal{F} : \mathcal{S}(\mathbb{X} \times \mathbb{U}) \rightarrow \mathcal{S}(\mathbb{X} \times \mathbb{U})$,

$$\begin{aligned} q^*(x, u) &= \ell(x, u) + \gamma \mathbb{E} \left[\inf_{w \in \mathbb{U}} q^*(f(x, u, \psi), w) \right] \\ &= (\mathcal{F}q^*)(x, u). \end{aligned} \quad (14)$$

The advantage of the Q -function reformulation is that policy extraction is model-free:

$$\pi^*(x) = \arg \min_{u \in \mathbb{U}} q^*(x, u). \quad (15)$$

In the following, we characterize the fixed point of \mathcal{F} under affine dynamics and generalized quadratic stage-cost.

Proposition 2 *The fixed point of (14) under dynamics (1), stage-cost (2) and $(\mathbb{X}, \mathbb{U}, \mathbb{D}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n)$ is*

$$q^*(x, u) = \begin{bmatrix} x \\ u \end{bmatrix}^\top Q^* \begin{bmatrix} x \\ u \end{bmatrix} + 2 \begin{bmatrix} x \\ u \end{bmatrix}^\top Q_\ell^* + Q_c^* + \frac{\gamma \text{Tr}(P^* \Sigma)}{1-\gamma}. \quad (16)$$

By considering $Q^* = \begin{bmatrix} q_x^* & q_\ell^* \\ \star & q_c^* \end{bmatrix}$, $Q_\ell^* = \begin{bmatrix} q_x^* \\ q_u^* \end{bmatrix}$ and (7), it holds

$$\left[\begin{array}{c|c} q_x^* & q_\ell^* \\ \star & Q_c^* \\ \star & \star \end{array} \middle| \begin{array}{c} q_\ell^{*\top} \\ q_c^* \end{array} \right] = \left[\begin{array}{c|c} \tilde{q}^* & \tilde{q}_\ell^* \\ \star & \tilde{q}_c^* \end{array} \right]. \quad (17)$$

The optimal policy is again given by (8).

Proof: Similarly to Proposition 1, we can consider the augmented dynamics (9) and augmented stage-cost (10) and verify that (16) satisfies the fixed point equation (14). \blacksquare Since the operator \mathcal{F} shares the same monotonicity and contractivity properties of \mathcal{T} [6], we can write again a (nonlinear) exact program for the Q -function

$$\begin{aligned} \sup_{q \in \mathcal{S}(\mathbb{X} \times \mathbb{U})} \int_{\mathbb{X} \times \mathbb{U}} q(x, u)c(dx, du) \\ \text{s.t. } q(x, u) \leq (\mathcal{F}q)(x, u) \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U}, \end{aligned} \quad (18)$$

where $c(\cdot, \cdot)$ takes the same role as in (12). Unlike (12), it is not straightforward to replace the nonlinear constraints in (18) with linear ones due to the nesting of the \mathbb{E} and \inf operators in (14). A linear reformulation of (18) can be obtained, as shown in [9] and [7], by introducing additional decision variables,

$$\begin{aligned} \sup_{\substack{v \in \mathcal{S}(\mathbb{X}), \\ q \in \mathcal{S}(\mathbb{X} \times \mathbb{U})}} \int_{\mathbb{X} \times \mathbb{U}} q(x, u)c(dx, du) \\ \text{s.t. } q(x, u) \leq (\mathcal{T}_\ell v)(x, u) \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U} \\ v(x) \leq q(x, u) \quad \forall (x, u) \in \mathbb{X} \times \mathbb{U}. \end{aligned} \quad (19)$$

In the next section, we show how to reduce the number of decision variables by introducing a modified operator.

D. Relaxed Bellman Operator Formulation

The authors in [20] introduce the relaxed Bellman operator $\hat{\mathcal{F}} : \mathcal{S}(\mathbb{X} \times \mathbb{U}) \rightarrow \mathcal{S}(\mathbb{X} \times \mathbb{U})$,

$$(\hat{\mathcal{F}}\hat{q})(x, u) = \ell(x, u) + \gamma \inf_{w \in \mathbb{U}} \mathbb{E}[\hat{q}(f(x, u, \psi), w)], \quad (20)$$

which retains the same structure of (14), but with the infimum and expectation operators exchanged. The next result extends [20, Theorem 3] to affine dynamics and generalized quadratic stage-cost, showing that the fixed point of $\hat{\mathcal{F}}$ is again an upper estimator of the fixed point of \mathcal{F} that preserves the minimizer with respect to u , *i.e.* the optimal policy.

Proposition 3 *The fixed point of (20) under dynamics (1), stage-cost (2) and $(\mathbb{X}, \mathbb{U}, \mathbb{D}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n)$ is*

$$\hat{q}(x, u) = q^*(x, u) + \frac{\gamma \text{Tr}(q_\ell^* q_c^{*-1} q_\ell^{*\top} \Sigma)}{1-\gamma}, \quad (21)$$

and the optimal policy is again given by (8).

Proof: Following the main steps of the proof of Theorem 3 in [20], one can verify that (21) satisfies the fixed point equation $\hat{q} = \hat{\mathcal{F}}\hat{q}$ and, by uniqueness of the fixed point of $\hat{\mathcal{F}}$, conclude the proof. \blacksquare

The relaxed operator $\hat{\mathcal{F}}$ shows significant computational advantage with respect to \mathcal{F} when used in the LP formulation [20]. In fact, since $\hat{\mathcal{F}}$ is also a monotone contraction mapping, its unique fixed point can be computed via a relaxation of (19) with reduced decision variables and constraints,

$$\begin{aligned} \sup_{q \in \mathcal{S}(\mathbb{X} \times \mathbb{U})} \int_{\mathbb{X} \times \mathbb{U}} q(x, u)c(dx, du) \\ \text{s.t. } q(x, u) \leq (\hat{\mathcal{F}}_\ell q)(x, u, w) \quad \forall (x, u, w) \in \mathbb{X} \times \mathbb{U}^2, \end{aligned} \quad (22)$$

where $(\hat{\mathcal{F}}_\ell q)(x, u, w) = \ell(x, u) + \gamma \mathbb{E}[q(f(x, u, \psi), w)]$. The relaxed LP (22) and the fixed point characterization (21)

constitute the starting point for the next discussion on how to synthesize Bellman inequalities from data.

III. SYNTHESIS OF BELLMAN INEQUALITIES FROM DATA

In the data-driven context, two fundamental problems in the LP formulation are the synthesis of Bellman inequalities from data (to avoid massive exploration of the state-space) and the estimation of expected values. In [19], a preliminary study on linear systems in the value function formulation is conducted. Here, we generalize the analysis to affine systems in the relaxed Bellman operator formulation. Moreover, we provide novel estimators for the expectation in the constraints that do not require reinitialization and discuss how to employ them to build LPs that preserve the optimal policy.

Consider the family of generalized quadratic functions

$$\mathcal{Q} = \{q(x, u) = [x \ u]^T Q [x \ u] + 2[x \ u]^T Q_\ell + Q_c, Q = Q^T\},$$

and note that $Q \in \mathcal{S}(\mathbb{X} \times \mathbb{U})$. Then, define $m_c \in \mathbb{R}_{\geq 0}$, $\mu_c \in \mathbb{R}^{n+m}$ and $\Sigma_c \succeq 0$ as the zeroth, first and second raw moment (i.e. centered about zero) of the measure $c(\cdot, \cdot)$.

Proposition 4 When $q(x, u) \in \mathcal{Q}$, the LP (22) takes the form

$$\begin{aligned} \max_{\varphi} \quad & [(\text{VEC } \Sigma_c)^T \ 2\mu_c^T \ m_c] \varphi \\ \text{s.t.} \quad & \mathbb{E}[\theta(x, u, \psi, w)]^T \varphi \leq \ell(x, u), \end{aligned} \quad (23)$$

for all $(x, u, w) \in \mathbb{X} \times \mathbb{U}^2$, where

$$\theta = \begin{bmatrix} \text{VEC}([x \ u][x \ u]^T - \gamma[x^+ \ w^+][x^+ \ w^+]^T) \\ 2[x \ u] - 2\gamma[x^+ \ w^+] \\ 1 - \gamma \end{bmatrix}, \quad \varphi = \begin{bmatrix} \text{VEC } Q \\ Q_\ell \\ Q_c \end{bmatrix}.$$

Proof: Any quadratic form can be decomposed as

$$[x \ u]^T Q [x \ u] = \text{VEC}([x \ u][x \ u]^T)^T \text{VEC } Q. \quad (24)$$

We obtain (23) by rearranging the constraints in (22) as $\mathbb{E}[q(x, u) - \gamma q(x^+, w)] \leq \ell(x, u)$, imposing $q(x, u) \in \mathcal{Q}$ and (24) and performing the integration in the objective. ■

Definition 1 (Dataset) When $X \in \mathbb{R}^{n \times d}$, $U \in \mathbb{R}^{m \times d}$ and $\Psi \in \mathbb{R}^{n \times d}$ are a collection of states, inputs and noise realizations and $X^+ = AX + BU + c\mathbf{1}^T + \Psi$, we say that (X, U, X^+) is a dataset of length d . A dataset corresponds to a single trajectory when $X_{i+1} = X_i^+$ for all $i = 1, \dots, d-1$, where X_i denotes the i -th column of X . To specify a length different from d we use the notation $X_{1:h} = [X_1 \ \dots \ X_h]$.

In order to estimate the expected values in the Bellman inequalities (23), as discussed e.g. in [20] and [22], one could reinitialize the dynamics at a fixed state-input pair (x, u) a number of times d , observe the corresponding transition $f(x, u, \Psi_i)$ and compute the unbiased estimator $\hat{\theta} = \frac{1}{d} \sum_{i=1}^d \theta(x, u, \Psi_i, w)$, such that $\mathbb{E}[\hat{\theta}] = \mathbb{E}[\theta(x, u, \psi, w)]$ and $\text{VAR}(\hat{\theta}) = \frac{1}{d} \text{VAR}(\theta(x, u, \psi, w))$. On the other hand, such an estimation can only be performed if one can reinitialize the dynamics at the same state x and play the same input u multiple times. Since this might be unrealistic in a stochastic framework, we show the effect of removing the reinitialization assumption by averaging the observations over the data directly instead of over the vectors $\theta(x, u, \Psi_i, w)$.

Lemma 1 Consider a dataset (X, U, X^+) of length d and a matrix $W \in \mathbb{R}^{m \times d}$ such that

$$\text{RANK} \begin{bmatrix} X \\ U \\ \mathbf{1}^T \\ W \end{bmatrix} = n + 2m + 1. \quad (25)$$

Then, $\forall (x, u, w) \in \mathbb{X} \times \mathbb{U}^2$, there exists $\alpha \in \mathbb{R}^d$ satisfying

$$\begin{bmatrix} X \\ U \\ \mathbf{1}^T \\ W \end{bmatrix} \alpha = \begin{bmatrix} x \\ u \\ 1 \\ w \end{bmatrix}, \quad (26)$$

and an estimator $\bar{\theta} = \theta(X\alpha, U\alpha, \Psi\alpha, W\alpha)$ such that

- (i) $\bar{\theta} = \theta(x, u, \Psi\alpha, w)$,
- (ii) $\bar{\theta}$ has mean $\mathbb{E}[\bar{\theta}] = \mathbb{E}[\theta(x, u, \bar{\psi}, w)]$ and covariance $\text{VAR}(\bar{\theta}) = \|\alpha\|_2^2 \text{VAR}(\theta(x, u, \psi, w))$, where $\bar{\psi}$ is a random vector with mean $\mathbb{E}[\bar{\psi}] = \mu$ and covariance $\bar{\Sigma} = \|\alpha\|_2^2 \Sigma$.

Proof: (i) First, note that the rank-condition (25) implies that (26) always has a solution. Then, we have $X^+ \alpha = (AX + BU + c\mathbf{1}^T + \Psi)\alpha = Ax + Bu + c + \Psi\alpha$. Finally, the result holds by substituting (26) and $X^+ \alpha$ into the definition of $\theta(x, u, \psi, w)$.

(ii) Let us define ψ_i , $i = 1, \dots, d$ as independent random vectors with mean μ and covariance Σ . Then, $\bar{\psi} = [\psi_1 \ \dots \ \psi_d] \alpha$ is also a random vector. In particular, its mean is $\mathbb{E}[\bar{\psi}] = \mathbf{1}^T \alpha \mu = \mu$ and, since $\text{COV}(\psi_i, \psi_j) = 0$ for all $i \neq j$ due to independence, its covariance is $\text{VAR}(\bar{\psi}) = \|\alpha\|_2^2 \Sigma = \bar{\Sigma}$. The claim then follows by considering (i). ■

Note that if the underlying dynamics is deterministic (i.e. $\Psi = 0$), the estimator reduces to $\bar{\theta} = \theta(x, u, 0, w)$. Then, if (25) holds, due to the lack of expectation in (23) one can potentially reconstruct all infinite constraints by taking linear combination of the data, similarly to the discussion in [19] for linear systems in the value function formulation.

In general, $\bar{\theta}$ is an unbiased estimator of $\mathbb{E}[\theta(x, u, \bar{\psi}, w)]$, instead of $\mathbb{E}[\theta(x, u, \psi, w)]$. The former is the coefficient associated to the Bellman inequalities of the dynamical system $x^+ = Ax + Bu + c + \bar{\psi}$. By inspecting (8), we note that the optimal policy $\pi^*(x)$ depends on $\mathbb{E}[\psi]$ but not on Σ ; the latter appears only in the constant terms of $v^*(x)$, $q^*(x, u)$ and $\hat{q}(x, u)$ (see (5), (16) and (21)). Hence, for $(\mathbb{X}, \mathbb{U}, \mathbb{D}) = (\mathbb{R}^n, \mathbb{R}^m, \mathbb{R}^n)$, the solution to (23) associated to $\bar{\psi}$ is

$$\bar{q}(x, u) = \hat{q}(x, u) + \frac{\gamma \text{TR}(\|\alpha\|_2^2 - 1) q^* \Sigma}{1 - \gamma},$$

and the associated optimal policy remains (8).

In summary, the estimator $\bar{\theta}$ can be computed from system's trajectories, it does not require reinitialization and can be used to construct LPs with biased constraints that preserve the optimal policy. Note that, to implement the approximation described above, one has to rely on estimators with the same covariance, i.e. same $\|\alpha\|_2^2$. The study of statistical bounds due to constraint approximation is deferred to future research, while the interested reader is referred to [13] and [21] for a discussion on error bounds due to constraint sampling and randomized optimization in the LP framework. Finally, please refer to [20] for a description on how to implement the LPs described above in a model-free fashion and on the observed control performance.

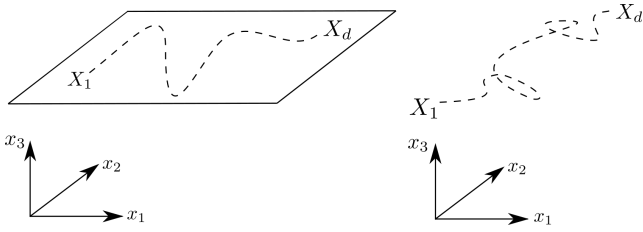


Fig. 1. Two example trajectories in \mathbb{R}^3 for a controllable system. The one on the left is contained in an affine subspace orthogonal to a unit vector: the associated input sequence is not persistently exciting of order $n + 2 = 5$ or higher.

It follows from Theorem 1 that, similarly to the linear setting and even though the augmented pair (\tilde{A}, \tilde{B}) is uncontrollable, persistency of excitation and controllability of (A, B) are sufficient conditions to obtain a dataset with enough information so that (Theorem 1(i)) the rank condition on the data matrix is satisfied, and (Theorem 1(ii)) L -long trajectories are representable as linear combinations of the input-output data. The notable difference is that in case of affine systems we need excitability of one order higher to guarantee that $\mathbf{1}^\top \alpha = 1$ is satisfied.

Theorem 1(i) provides an additional insight: the trajectory of a controllable affine system with a persistently exciting input of order $n + 2$ can not be contained in any affine subspace of \mathbb{R}^n orthogonal to a unit vector (see Fig. 1). The same consideration is valid for linear systems by considering Theorem 1(i) with $c = 0$.

Lastly, we comment on the use of W , which originates from the relaxation of the constraints in (22). Its design is independent from the collected data and, as mentioned in [1] in deterministic settings, it can be used offline to generate new constraints associated with the same (x, u) pairs but different w . For the first time, in the present paper, we provide a design condition on W via (25) and establish that it must be selected to be independent from the observed state-input trajectories.

V. CONCLUSION

The present letter focuses on optimal control for affine systems via data-driven linear programming. After introducing the fixed point characterization of three fundamental operators, we show how to synthesize the Bellman inequalities in the LP formulations from data and provide estimators for the associated expected values that preserve the optimal policy. To provide sufficient conditions for the mentioned results, we complete the proof of Willems' fundamental lemma for affine systems. Future research directions will include relaxation of the sufficient conditions in the spirit of [29] and online experiment design [24].

REFERENCES

[1] G. Banjac and J. Lygeros. A data-driven policy iteration scheme based on linear programming. In *58th IEEE Conference on Decision and Control*, pages 816–821, 2019.

[2] S. Barratt and S. Boyd. Stochastic control with affine dynamics and extended quadratic costs. *IEEE Transactions on Automatic Control*, 67(1):320–335, 2022.

[3] R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719, 1952.

[4] J. Berberich, J. Koehler, M.A. Muller, and F. Allgower. Linear tracking MPC for nonlinear systems part II: The data-driven case. *IEEE Transactions on Automatic Control*, Early Access, 2022.

[5] D.P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, 2013.

[6] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

[7] P.N. Beuchat, A. Georghiou, and J. Lygeros. Performance guarantees for model-based approximate dynamic programming in continuous spaces. *IEEE Transactions on Automatic Control*, 65(1):143–158, 2020.

[8] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[9] R. Cogill, M. Rotkowitz, B. Van Roy, and S. Lall. An approximate dynamic programming approach to decentralized control of stochastic systems. In *Control of Uncertain Systems: Modelling, Approximation, and Design*, pages 243–256. Springer-Verlag Berlin Heidelberg, 2006.

[10] J. Coulson, J. Lygeros, and F. Dörfler. Data-enabled predictive control: In the shallows of the DeePC. In *18th European Control Conference (ECC)*, pages 307–312, 2019.

[11] M.H.A. Davis and R.B. Vinter. *Stochastic Modelling and Control*. Chapman and Hall, 1985.

[12] D.P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[13] D.P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.

[14] C. De Persis and P. Tesi. Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65(3):909–924, 2020.

[15] E.V. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9(2):165–177, 1967.

[16] O. Hernandez-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag NY, 1996.

[17] J.B. Lasserre. A sum of squares approximation of nonnegative polynomials. *SIAM Review*, 49(4):651–669, 2007.

[18] A.S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[19] A. Martinelli, M. Gargiani, and J. Lygeros. On the synthesis of Bellman inequalities for data-driven optimal control. In *60th IEEE Conference on Decision and Control*, pages 4352–4357, 2021.

[20] A. Martinelli, M. Gargiani, and J. Lygeros. Data-driven optimal control with a relaxed linear program. *Automatica*, 136:110052, 2022.

[21] P. Mohajerin Esfahani, T. Sutter, D. Kuhn, and J. Lygeros. From infinite to finite programs: Explicit error bounds with applications to approximate dynamic programming. *SIAM Journal on Optimization*, 28(3):1968–1998, 2018.

[22] T. Sutter, A. Kamoutsis, P. Mohajerin Esfahani, and J. Lygeros. Data-driven approximate dynamic programming: A linear programming approach. In *56th IEEE Conference on Decision and Control*, pages 5174–5179, 2017.

[23] A. Tzafanakis and J. Lygeros. Data-driven control of unknown systems: A linear programming approach. *IFAC-PapersOnLine*, 53(2):7–13, 2020. 21th IFAC World Congress.

[24] H.J. van Waarde. Beyond persistent excitation: Online experiment design for data-driven modeling and control. *IEEE Control Systems Letters*, 6:319–324, 2021.

[25] H.J. van Waarde, C. De Persis, M.K. Camlibel, and P. Tesi. Willems' fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4(3):602–607, 2020.

[26] H.J. van Waarde, J. Eising, H.L. Trentelman, and M.K. Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020.

[27] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992.

[28] J.C. Willems, P. Rapisarda, I. Markovskiy, and B.L.M. De Moor. A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329, 2005.

[29] Y. Yu, S. Talebi, H.J. van Waarde, U. Topcu, M. Mesbahi, and B. Açıkmeşe. On controllability and persistency of excitation in data-driven control: Extensions of Willems' fundamental lemma. In *60th IEEE Conference on Decision and Control*, pages 6485–6490, 2021.