

Learning continuous Q -functions using generalized Benders cuts

Joseph Warrington, *Member, IEEE*

Abstract— Q -functions are widely used in discrete-time learning and control to model future costs arising from a given control policy, when the initial state and input are given. Although some of their properties are understood, Q -functions generating optimal policies for continuous problems are usually hard to compute. Even when a system model is available, optimal control is generally difficult to achieve except in rare cases where an analytical solution happens to exist, or an explicit exact solution can be computed. It is typically necessary to discretize the state and action spaces, or parameterize the Q -function with a basis that can be hard to select *a priori*. This paper describes a model-based algorithm based on generalized Benders theory that yields ever-tighter outer-approximations of the optimal Q -function. Under a strong duality assumption, we prove that the algorithm yields an arbitrarily small Bellman optimality error at any finite number of arbitrary points in the state-input space, in finite iterations. Under additional assumptions, the same guarantee holds when the inputs are determined online by the algorithm’s updating Q -function. We demonstrate these properties numerically on scalar and 8-dimensional systems.

I. INTRODUCTION

Reinforcement learning (RL) and approximate dynamic programming (ADP) commonly employ so-called Q -functions to model the costs incurred in the future evolution of a discrete-time system under a given control policy. The Q -function associated with control policy $u = \pi(x)$ takes a state \hat{x} and input \hat{u} as parameters, and is equal to the stage costs incurred immediately for (\hat{x}, \hat{u}) plus the costs (typically infinite-horizon with a discount factor) of following policy π thereafter.

Q -functions are widely associated with RL (i.e., model-free learning), thanks to work stemming from Watkins’ Q -learning algorithm [17]. However, model-free Q -learning suffers from slow convergence, even despite new insights into optimizing the rate [7]. New work such as [11] is an example of interest in cases where model data, known or itself learned, can improve learning performance for difficult control problems. The present paper is motivated by a desire to learn an approximate Q -function to control a system with a known model.

Mathematically, Q -functions have much in common with *value* (or V -) functions, the chief difference being that they are defined on state-input space rather than on the state alone. Although they are generally more expensive to store, their higher-dimensional domain often makes approximate, finitely-parameterized Q -functions more expressive than V -functions [4, Ch. 2].

It is common to discretize continuous problems in order to obtain a finite parameterization of the V - or Q -function

[5]. However, performing even one iteration of the canonical algorithms, such as value iteration, then has an undesirable exponential cost. ADP methods have arisen to find more tractable parameterizations of the continuous V -function. Several are based on continuous extensions of the “linear programming approach” to ADP [6], in which a valid lower bound on the optimal value function is maximized. Examples include the quadratic lower bound in [14], and the polynomial derived using sum-of-squares techniques in [13]. Approximate V -functions represented as the pointwise maximum of multiple lower-bounding functions have been used in [1], [9], [10], [15]. Recent work utilizing a point-wise maximum representation [16] has extended the Benders decomposition argument used for linear multi-stage decision problems in Dual DP (DDP, [12]), to a general nonlinear, infinite-horizon setting.

In this paper we adapt the Benders approach from [16] to learn Q -functions. We define an algorithm that successively produces tighter approximations of a problem’s optimal Q -function from below, and prove convergence results for off-policy and policy-driven learning of the Q -function in this manner. In the former case, (x, u) pairs are pre-selected at the start of the algorithm, whereas in the latter case, only the x points are pre-selected, and the u decisions are made according to a policy from the update Q -function estimate. We then demonstrate the method’s efficacy for test systems.

Section II describes the infinite horizon problem, Section III describes the Benders decomposition approach, and Section IV proposes an algorithm and proves its key properties. Section V presents numerical examples, and Section VI concludes.

II. PROBLEM STATEMENT

A. Infinite-horizon control problem

The scope considered is the class of infinite-horizon, discrete-time, deterministic optimal control problems with time-invariant stage cost functions, dynamics, and constraints:

$$V^*(x) := \inf_{u_0, u_1, \dots} \sum_{t=0}^{\infty} \gamma^t \ell(x_t, u_t) \quad (1a)$$

$$\text{s. t. } x_{t+1} = f(x_t, u_t), \quad t = 0, 1, \dots, \quad (1b)$$

$$h(x_t, u_t) \leq 0, \quad t = 0, 1, \dots, \quad (1c)$$

$$x_0 = x. \quad (1d)$$

For each time step t we denote the state $x_t \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$, and the action, or input, $u_t \in \mathcal{U} \subseteq \mathbb{R}^{n_u}$. Sets \mathcal{X} and \mathcal{U} are the state and action spaces, and are continuous. Future costs are discounted according to a discount factor $\gamma \in (0, 1]$, the

(non-negative) stage cost function is $\ell : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}_+$, and the dynamics are governed by the mapping $f : \mathcal{X} \times \mathcal{U} \rightarrow \mathcal{X}$. There are n_c state-input constraints (1c), parameterized by a vector-valued mapping $h : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}^{n_c}$. The parametric infimum $V^*(x)$ of problem (1) is referred to as the *optimal value function* (or optimal V -function) of the problem.

B. Q -functions

We now define Q -functions and briefly state some of their well-known properties for later use. For more detail, see for example [4, Chapter 2]. Given a policy $\pi : \mathcal{X} \rightarrow \mathcal{U}$, its associated Q -function, $Q^\pi : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R} \cup \{+\infty\}$, is

$$Q^\pi(x, u) = \ell(x, u) + \sum_{t=1}^{\infty} \gamma^t \ell(x_t, \pi(x_t)), \quad (2)$$

in which the relation $x_{t+1} = f(x_t, \pi(x_t))$ holds for $t \geq 1$, and $x_1 = f(x, u)$. The Q -function is the sum of the stage cost incurred for some initial state and input x and u , and the infinite sum of (discounted) costs under policy π thereafter.

The optimal Q -function, which we denote Q^* , minimizes (2) over policies π , and satisfies

$$Q^*(x, u) = \ell(x, u) + \inf_{u' \in \mathcal{U}(f(x, u))} Q^*(f(x, u), u') \quad (3)$$

for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. In (3) we use the notation $\mathcal{U}(x) := \{u \in \mathcal{U} : h(x, u) \leq 0\}$. An associated Bellman operator for Q -functions, \mathcal{T}_Q , can be defined as

$$\mathcal{T}_Q Q(x, u) := \ell(x, u) + \inf_{u' \in \mathcal{U}(f(x, u))} Q(f(x, u), u'). \quad (4)$$

On the left-hand side, $\mathcal{T}_Q Q$ is to be interpreted as a new function with the same domain as Q , and evaluated at (x, u) . Thus, condition (3) can be written $\mathcal{T}_Q Q^*(x, u) = Q^*(x, u)$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. If an optimal Q - and V -function exist for problem (1), they are related by $V^*(x) = \inf_{u \in \mathcal{U}(x)} Q^*(x, u)$, and thus from (3), $Q^*(x, u) = \ell(x, u) + \gamma V^*(f(x, u))$.

For any approximate Q -function for which the infimum in (4) is attained, one can define an associated control policy consistent with definition (2):

$$\pi(x; Q) \in \arg \min_{u \in \mathcal{U}(x)} Q(x, u). \quad (5)$$

The attraction of a Q -function is that in a wide range of cases it is simpler to solve (5) than it would be to solve, for the same x , the full infinite-horizon problem (1), or a finite-horizon truncation thereof, as in Model Predictive Control (MPC) [2].

Lastly, for the benefit of developments in Section III, we note it is easy to show that the operator \mathcal{T}_Q is monotonic:

$$\begin{aligned} Q_a(x, u) &\leq Q_b(x, u) \quad \forall (x, u) \in \mathcal{X} \times \mathcal{U} \\ \Rightarrow \mathcal{T}_Q Q_a(x, u) &\leq \mathcal{T}_Q Q_b(x, u) \quad \forall (x, u) \in \mathcal{X} \times \mathcal{U}. \end{aligned} \quad (6)$$

III. BENDERS CUTS

A. Pointwise maximum representation

Let $Q_I : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ be a function of the following “pointwise maximum” form,

$$Q_I(x, u) = \max_{i=0, \dots, I} \{q_i(x, u)\}, \quad (7)$$

where I is a non-negative integer, and each function $q_i : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is known to satisfy

$$q_i(x, u) \leq Q^*(x, u), \quad \forall (x, u) \in \mathcal{X} \times \mathcal{U}.$$

Thus $Q_I(x, u) \leq Q^*(x, u)$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. From (5) the control policy associated with Q_I is simply $\pi(x; Q_I) \in \arg \min_{u \in \mathcal{U}(x)} \max_{i=0, \dots, I} \{q_i(x, u)\}$.

In Section IV we will propose an algorithm that uses Q_I to construct an additional function, or “cut” q_{I+1} . Under certain assumptions, the new cut satisfies

$$q_{I+1}(x, u) \leq Q^*(x, u) \quad \forall (x, u) \in \mathcal{X} \times \mathcal{U}, \quad (8a)$$

$$\text{and } q_{I+1}(\hat{x}, \hat{u}) > Q_I(\hat{x}, \hat{u}) \text{ for some } (\hat{x}, \hat{u}) \in \mathcal{X} \times \mathcal{U}. \quad (8b)$$

Thus the new function, $Q_{I+1}(x, u) := \max\{Q_I(x, u), q_{I+1}(x, u)\}$, will be a tighter under-approximation of Q^* than Q_I . We now derive a Benders-type procedure to achieve this, which is related to that in [16] for V -functions.

B. Duality in operator \mathcal{T}_Q

We start by taking the dual of the minimization problem solved inside the operator \mathcal{T}_Q at some point (\hat{x}, \hat{u}) in the state-action space. For a function Q_I taking the form (7), the right-hand side of (4) can be written equivalently as

$$\begin{aligned} \mathcal{T}_Q Q_I(\hat{x}, \hat{u}) &= \inf_{x', u'} \ell(\hat{x}, \hat{u}) + \gamma \max_{i=0, \dots, I} \{q_i(x', u')\} \\ \text{s. t. } &x' = f(\hat{x}, \hat{u}), \\ &h(x', u') \leq 0, \end{aligned}$$

where the extra variable $x' \in \mathbb{R}^{n_x}$ is introduced to model the successor state explicitly. An epigraph variable $\alpha \in \mathbb{R}$ can be introduced to replace the inconvenient maximum operator in the objective with $I + 1$ separate constraints. This leads to an equivalent problem:

$$\mathcal{T}_Q Q_I(\hat{x}, \hat{u}) = \inf_{x', u', \alpha} \ell(\hat{x}, \hat{u}) + \gamma \alpha \quad (9a)$$

$$\text{s. t. } x' = f(\hat{x}, \hat{u}), \quad (9b)$$

$$h(x', u') \leq 0, \quad (9c)$$

$$q_i(x', u') \leq \alpha, \quad i = 0, \dots, I. \quad (9d)$$

Assigning the Lagrange multipliers $\nu \in \mathbb{R}^{n_x}$, $\lambda_c \in \mathbb{R}_+^{n_c}$, and $\lambda_\alpha \in \mathbb{R}_+^{I+1}$ to constraints (9b), (9c), and (9d) respectively, one can form the Lagrangian,

$$\begin{aligned} \mathcal{L}(x', u', \alpha, \nu, \lambda_c, \lambda_\alpha) &:= \ell(\hat{x}, \hat{u}) + \gamma \alpha + \nu^\top (f(\hat{x}, \hat{u}) - x') \\ &\quad + \lambda_c^\top h(x', u') + \sum_{i=0}^I \lambda_{\alpha, i} (q_i(x', u') - \alpha). \end{aligned}$$

Following standard procedure, the dual of (9) is then

$$J_D(\hat{x}, \hat{u}) := \sup_{\nu, \lambda_c, \lambda_\alpha} \ell(\hat{x}, \hat{u}) + \nu^\top f(\hat{x}, \hat{u}) + \xi(\nu, \lambda_c, \lambda_\alpha) \quad (10a)$$

$$\text{s. t. } \mathbf{1}^\top \lambda_\alpha = \gamma, \quad (10b)$$

$$\lambda_c \geq 0, \quad \lambda_\alpha \geq 0, \quad (10c)$$

where the function

$$\xi(\nu, \lambda_c, \lambda_\alpha) := \inf_{x', u'} \left\{ -\nu^\top x' + \lambda_c^\top h(x', u') + \sum_{i=0}^I \lambda_{\alpha, i} q_i(x', u') \right\}$$

depends only on the multipliers. Although problem (9) may not be convex, the objective of (10) is always concave [3, §5.2], and weak duality implies $J_D(\hat{x}, \hat{u}) \leq \mathcal{T}_Q Q_I(\hat{x}, \hat{u})$ for any choice of parameter $(\hat{x}, \hat{u}) \in \mathcal{X} \times \mathcal{U}$.

C. Generalized Benders cut

Given a function Q_I of the form (7) such that $Q_I \leq Q^*$, suppose that optimal multipliers $(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*)$ are attained when (10) is solved with parameter (\hat{x}, \hat{u}) . These can be used to form a new cut $q_{I+1}(\cdot, \cdot)$ with the following attractive properties.

Lemma III.1. *The function*

$$q_{I+1}(x, u) := \ell(x, u) + \hat{\nu}^{*\top} f(x, u) + \xi(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*) \quad (11)$$

satisfies $q_{I+1}(x, u) \leq Q^*(x, u)$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$.

Proof. An optimal dual solution $(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*)$ for parameter (\hat{x}, \hat{u}) must in general be a suboptimal solution to problem (10) when any other parameter $(\bar{x}, \bar{u}) \in \mathcal{X} \times \mathcal{U}$ is used, i.e.,

$$\ell(\bar{x}, \bar{u}) + \hat{\nu}^{*\top} f(\bar{x}, \bar{u}) + \xi(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*) \leq J_D(\bar{x}, \bar{u}).$$

Note that $(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*)$ is feasible in (10) for all parameters (\bar{x}, \bar{u}) , as the feasible set is independent of the parameter. From weak duality, $J_D(\bar{x}, \bar{u}) \leq \mathcal{T}_Q Q_I(\bar{x}, \bar{u})$. As we start with $Q_I \leq Q^*$ on its domain, we have from the mononicity property (6) that $\mathcal{T}_Q Q_I(\bar{x}, \bar{u}) \leq \mathcal{T}_Q Q^*(\bar{x}, \bar{u})$, and the Bellman optimality condition states that $\mathcal{T}_Q Q^*(\bar{x}, \bar{u}) = Q^*(\bar{x}, \bar{u})$. Combining these relationships we obtain

$$\ell(\bar{x}, \bar{u}) + \hat{\nu}^{*\top} f(\bar{x}, \bar{u}) + \xi(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*) \leq Q^*(\bar{x}, \bar{u}),$$

and the result follows simply by noting that (\bar{x}, \bar{u}) can refer to any $(x, u) \in \mathcal{X} \times \mathcal{U}$ in the argument above. \square

This proof leverages the (generalized) Benders decomposition argument, which was first developed in [8] to partition a two-stage problem into two subproblems linked by an approximate value function. Here we have used the properties of Q -functions to accommodate the infinite number of stages in problem (1). A similar result was derived for V -functions in [16].

The following properties concern the violation of the Bellman optimality condition (3), or the Q -Bellman error:

$$\varepsilon(x, u; Q_I) := \mathcal{T}_Q Q_I(x, u) - Q_I(x, u). \quad (12)$$

Lemma III.2. *If $\varepsilon(x, u; Q_I) \geq 0$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$, then $\varepsilon(x, u; Q_{I+1}) \geq 0$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$, where*

$$Q_{I+1}(\cdot, \cdot) = \max\{q_{I+1}(\cdot, \cdot), Q_I(\cdot, \cdot)\}.$$

Proof. A simple adaptation of [16, Lemma III.3]. \square

Lemma III.3. *Suppose strong duality holds between problems (9) and (10) and that $\varepsilon(x, u; Q_I) = \mathcal{T}_Q Q_I(x, u) - Q_I(x, u) \geq 0$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$. Then if at some (\hat{x}, \hat{u}) we have $\mathcal{T}_Q Q_I(\hat{x}, \hat{u}) > Q_I(\hat{x}, \hat{u})$, a cut there is strictly improving:*

$$Q_{I+1}(\hat{x}, \hat{u}) > Q_I(\hat{x}, \hat{u}),$$

and the increase is equal to $\varepsilon(\hat{x}, \hat{u}; Q_I)$.

Proof. If strong duality holds, we have $J_D(\hat{x}, \hat{u}) = \mathcal{T}_Q Q_I(x, u)$, and the new function q_{I+1} satisfies

Algorithm 1 Q -Benders algorithm

```

1: Variant A: Choose  $\mathcal{Z}_{\text{Alg}} := \{(x_1, u_1), \dots, (x_M, u_M)\}$ 
2: Variant B: Choose  $\mathcal{X}_{\text{Alg}} := \{x_1, \dots, x_M\}$ 
3: Set  $I = 0$  and  $q_0(x, u) = \ell(x, u)$ 
4: while TRUE do
5:    $Q_I(\cdot, \cdot) \leftarrow \max_{i=0, \dots, I} q_i(\cdot, \cdot)$ 
6:   for  $m = 1, \dots, M$  do
7:     Variant A:  $(x_m, u_m)$  taken from  $\mathcal{Z}_{\text{Alg}}$ 
8:     Variant B:  $x_m$  taken from  $\mathcal{X}_{\text{Alg}}$ ,
9:        $u_m \leftarrow \arg \min_{u \in \mathcal{U}(x_m)} Q_I(x_m, u)$ 
10:     $\varepsilon(x_m, u_m; Q_I) \leftarrow \mathcal{T}_Q Q_I(x_m, u_m) - Q_I(x_m, u_m)$ 
11:  end for
12:  if  $\max_{m=1, \dots, M} \{\varepsilon(x_m, u_m; Q_I)\} \leq \varepsilon_{\text{tol}}$  then
13:    break
14:  end if
15:   $m \leftarrow \text{UniformRandom}(1, 2, \dots, M)$ 
16:  Pick  $(x_m, u_m)$  as in lines 7-9
17:  if problem (9) feasible with parameter  $(x_m, u_m)$  and
    dual optimal solution  $(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*)$  available then
18:    Add  $q_{I+1}(\cdot, \cdot)$  param'd by  $(\hat{\nu}^*, \hat{\lambda}_c^*, \hat{\lambda}_\alpha^*)$  as in (11)
19:  end if
20:   $I \leftarrow I + 1$ 
21: end while
22: Return  $Q_I(\cdot, \cdot) = \max_{i=0, \dots, I} q_i(\cdot, \cdot)$ 

```

$q_{I+1}(\hat{x}, \hat{u}) = \mathcal{T}_Q Q_I(\hat{x}, \hat{u})$. Since $Q_{I+1}(x, u) := \max\{Q_I(x, u), q_{I+1}(x, u)\}$ the result follows. \square

Lastly, the following property facilitates a “greedy” cut $q_{I+1}(\cdot, \cdot)$ with respect to some particular (\hat{x}, \hat{u}) location.

Lemma III.4. *The Benders cut that yields the greatest increase at (\hat{x}, \hat{u}) , i.e., for which $Q_{I+1}(\hat{x}, \hat{u}) - Q_I(\hat{x}, \hat{u})$ is maximized, is that obtained by solving problem (10) at (\hat{x}, \hat{u}) .*

Proof. The result follows by reversing the roles of (\hat{x}, \hat{u}) and (\bar{x}, \bar{u}) in the proof of Lemma III.1. \square

IV. BENDERS ALGORITHM FOR Q -FUNCTIONS

We propose Algorithm 1 as a means of approximating Q^* by generating Benders cuts of the form (11). It starts with $q_0 = \ell$, which from (2) trivially lower-bounds Q^* , and by Lemmas III.2 and III.3 guarantees $\varepsilon(x, u; Q_I) \geq 0$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$ and for all $I \geq 0$. New cuts are created at certain points (x_m, u_m) , and Variants A and B differ in how these are chosen:

- Select a list of state-input pairs $\mathcal{Z}_{\text{Alg}} := \{(x_1, u_1), \dots, (x_M, u_M)\}$ *a priori*, and choose a random (x_m, u_m) at each algorithm iteration.
- Select a list of state space points $\mathcal{X}_{\text{Alg}} := \{x_1, \dots, x_M\}$ *a priori*, and within the algorithm pick a random x_m , letting u_m follow from policy (5) parameterized by Q_I .

We now state convergence results for both variants.

A. Fixed (x, u) pairs

The following results hold for Variant A. We omit the proofs of both, because they carry across with little modification from the V -function results in [16, Thms. III.5 and III.6]:

Theorem IV.1 (Pointwise convergence of $\{Q_I(x, u)\}_{I=0}^\infty$). *For each $(x, u) \in \mathcal{X} \times \mathcal{U}$ for which $Q^*(x, u)$ is finite, there exists a limiting value $Q_{\lim}(x, u) \leq Q^*(x, u)$ such that $\lim_{I \rightarrow \infty} Q_I(x, u) = Q_{\lim}(x, u)$.*

Theorem IV.2 (Finite termination of Variant A). *Suppose the following conditions are met:*

- (i) *Strong duality holds for the one-stage problem (9) with parameter (x_m, u_m) each time it is solved, for each $(x_m, u_m) \in \mathcal{Z}_{\text{Alg}}$.*
- (ii) *$Q^*(x_m, u_m)$ is finite for each pair $(x_m, u_m) \in \mathcal{Z}_{\text{Alg}}$.*

Then Variant A of Algorithm 1 terminates in finite iterations with probability 1 for any tolerance $\varepsilon_{\text{tol}} > 0$.

B. Fixed x , policy-driven u

Although Variant A has attractive convergence properties, it learns a Q -function based only on performance at pre-selected pairs (x, u) , in the sense of minimizing the Q -Bellman error there. Variant B instead learns a Q -function based on performance at (x, u) pairs in which the u is consistent with the policy derived from the learnt Q -function. One expects this criterion to be more relevant to performance of the final policy, as state-input trajectories will pass closer to these points.

Finite termination of Variant B is our main result, which we now state precisely along with the required assumptions.

Assumption 1. *For each $x_m \in \mathcal{X}_{\text{Alg}}$, the set of feasible inputs $\mathcal{U}(x_m)$ contains an element \hat{u} such that $Q^*(x_m, \hat{u}) < \infty$.*

This assumption implies $V^*(x_m)$ is finite for each x_m . Introducing the notation $\underline{Q}(x_m) := \inf_{u \in \mathcal{U}(x_m)} Q(x_m, u)$, the following holds:

Theorem IV.3 (Monotone convergence of $\{\underline{Q}_I(x_m)\}_{I=0}^\infty$). *Under Assumption 1, the limit $\underline{Q}_{\lim}(x_m) := \lim_{I \rightarrow \infty} \underline{Q}_I(x_m)$ exists for each $x_m \in \mathcal{X}_{\text{Alg}}$.*

Proof. It follows from Assumption 1 that $Q^*(x_m) < \infty$, and from Lemma III.1, $Q_I(x_m, u) \leq Q^*(x_m, u)$ for all $u \in \mathcal{U}(x_m)$. Thus, $\underline{Q}_I(x_m) < \infty$ at each iteration I . As the sequence of functions $\{Q_I\}_{I=0}^\infty$ increases monotonically, the sequence $\{\underline{Q}_I(x_m)\}_{I=0}^\infty$ must also increase monotonically. This latter sequence is bounded from above, thus the limit $\lim_{I \rightarrow \infty} \underline{Q}_I(x_m) = \underline{Q}_{\lim}(x_m)$ exists from the Monotone Convergence Theorem. \square

An additional performance guarantee for Variant B is available when the following additional assumptions hold.

Assumption 2. *For each $x_m \in \mathcal{X}_{\text{Alg}}$, set $\mathcal{U}(x_m)$ is compact, and each entry of $f(x_m, u)$ is Lipschitz-continuous on $\mathcal{U}(x_m)$.*

Assumption 3. *The problem data in (1) is such that the lower-bounding functions q_0, q_1, \dots generated in Variant B:*

- (i) *Maintain strong duality between problems (9) and (10) with parameter (x_m, u) at each iteration of the algorithm, with $u = \pi(x_m; Q_I)$, for all $x_m \in \mathcal{X}_{\text{Alg}}$.*
- (ii) *Are Lipschitz continuous in u with some constant L_m common to all functions q_i , for each $x_m \in \mathcal{X}_{\text{Alg}}$.*

Assumptions 2 and 3 must be verified for a given problem. A widespread setting where these hold is the constrained, stable linear-quadratic regulator (LQR); see the Appendix.

Theorem IV.4 (Finite termination of Variant B). *Suppose that in addition to Assumption 1, Assumptions 2 and 3 hold. Then Variant B of Algorithm 1 terminates in finite iterations with probability 1 for any tolerance $\varepsilon_{\text{tol}} > 0$.*

Proof. Let the sequence of iterations I where a given m is chosen in line 15 of the algorithm be indexed by I_m . With probability 1, this sequence is infinitely long for each m . We now show that the sequence of Q -Bellman errors

$$\{\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m})\}_{I_m=0}^\infty$$

is a Cauchy sequence converging to zero for each $x_m \in \mathcal{X}_{\text{Alg}}$. As $\mathcal{U}(x_m)$ is compact for all $x_m \in \mathcal{X}_{\text{Alg}}$, the policy $\pi(x_m; Q_I)$ defined in (5) can always be evaluated.

Recall that Lemma III.2 implies $\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m}) \geq 0$ for all x_m and I_m . Suppose for the sake of contradiction that the sequence $\{\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m})\}_{I_m=0}^\infty$ is not a Cauchy sequence converging to 0. Then there must exist some $\delta > 0$ for which there is no iteration number beyond which $\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m}) < \delta$. Whenever point x_m is picked in line 15 of the algorithm, the strong duality condition in Assumption 3 and Lemma III.3 together imply that

$$\begin{aligned} Q_{I_m+1}(x_m, \pi(x_m; Q_{I_m})) - Q_{I_m}(x_m, \pi(x_m; Q_{I_m})) \\ = \varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m}). \end{aligned}$$

If $\{\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m})\}_{I_m=0}^\infty$ is not a Cauchy sequence, there will be an infinite number of occasions on which

$$Q_{I_m+1}(x_m, \pi(x_m; Q_{I_m})) - Q_{I_m}(x_m, \pi(x_m; Q_{I_m})) \geq \delta.$$

Furthermore, Assumption 1 and part (ii) of Assumption 3 together imply that

$$Q_I(x_m, u) \leq Q^*(x_m, \hat{u}) + L_m \|u - \hat{u}\|, \quad \forall u \in \mathcal{U}(x_m), \quad \forall I \in \mathbb{N},$$

as Q_I is always a lower bound on Q^* . Compactness of $\mathcal{U}(x_m)$ implies that the volume of the truncated hypograph

$$\mathcal{H}_m := \left\{ u, s \mid \begin{array}{l} Q_I(x_m, u) \leq s \leq Q^*(x_m, \hat{u}) + L_m \|u - \hat{u}\|, \\ u \in \mathcal{U}(x_m) \end{array} \right\}$$

is finite for each x_m ; recall that $q_0(\cdot, \cdot) \equiv \ell(\cdot, \cdot) \geq 0$.

Due to Lipschitz continuity, cut q_{I_m+1} decreases the volume of \mathcal{H}_m by an amount that is lower bounded by a function of δ , L_m , and the input dimension n_u . Thus, this volume cannot be removed infinitely many times from \mathcal{H}_m , and we have a contradiction. Cuts made at iterations I where some other index $m' \neq m$ is picked in line 15 may also remove some volume from \mathcal{H}_m , but this does not affect the argument. Thus $\{\varepsilon(x_m, \pi(x_m; Q_{I_m}); Q_{I_m})\}_{I_m=0}^\infty$ is a Cauchy sequence converging to zero, and Algorithm 1 terminates in finite iterations for any $\varepsilon_{\text{tol}} > 0$. \square

Therefore, under certain assumptions one need only specify $\mathcal{X}_{\text{Alg}} = \{x_1, \dots, x_M\}$, and Variant B minimizes the Q -Bellman error at a $u \in \mathcal{U}(x_m)$ associated with each $x_m \in \mathcal{X}_{\text{Alg}}$ that is consistent with policy (5). One then expects the optimal Q -function to be learnt more accurately around the policy surface than elsewhere in the state-action space.

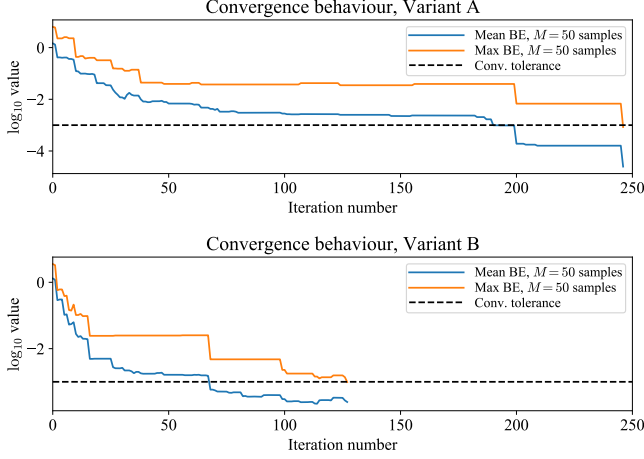


Fig. 1. Convergence behaviour under Variants A (top) and B (bottom) of Algorithm 1 for a 1-dimensional system. Variant A terminated in 246 iterations, while Variant B terminated in 127 iterations.

V. NUMERICAL EXAMPLES

We now report two numerical tests of Algorithm 1. In both cases, systems were of the class **C-LQR** described in the Appendix, for which finite termination of Variants A and B is guaranteed by Theorems IV.2 and IV.4 respectively, and lower bounding functions $q_i(\cdot, \cdot)$ are quadratic. All tests used the stage cost $\ell(x_t, u_t) = \frac{1}{2}x_t^\top x_t + \frac{1}{2}u_t^\top u_t$, discount rate $\gamma = 1$, and termination tolerance $\varepsilon_{\text{tol}} = 10^{-3}$, with $h(x_t, u_t)$ encoding an input constraint $\|u_t\|_\infty \leq 1$. Tests were implemented in Python with subproblems solved using Gurobi 7.0.2, on a computer with an Intel i7 CPU at 2.60 GHz and 16 GB RAM.

Scalar system: For ease of visualization, we used the simple system $x_{t+1} = 0.9x_t + u_t$, with $x_t, u_t \in \mathbb{R}$, and ran both algorithm variants. In Variant A, \mathcal{X}_{Alg} contained 50 random states x sampled uniformly from the interval $[0, 3]$, and a random $u \in [-1, 1]$ associated with each x . In Variant B the associated inputs were dropped to form \mathcal{X}_{Alg} . Fig. 1 shows convergence of the maximum Q -Bellman error $\max_{x_m \in \mathcal{X}_{\text{Alg}}} \varepsilon(x_m, \pi(x; Q_I); Q_I)$, with the mean error $\frac{1}{M} \sum_{m=1}^M \varepsilon(x_m, \pi(x_m; Q_I); Q_I)$ shown for comparison. Total time spent generating lower-bounding functions was 243 ms for Variant A and 121 ms for Variant B. For this simple system the optimal policy can be computed as $\pi^*(x) = \min\{1, \max\{-0.5377x, -1\}\}$. Fig. 2 shows the evolution of Q_I at selected iterations; after termination, the policy (5) from Variant B was closer to the optimal policy than that from Variant A. The average closed-loop cost starting from points $x \in \mathcal{X}_{\text{Alg}}$ was also lower at 2.47991, compared to 2.48716 for Variant A, and 2.47968 for the optimal policy.

Higher-dimensional systems: We tested Variant B for systems too large for the optimal policy to be computed exactly. 20 random 8-state, 3-input linear systems were created with $\|A\| \leq 0.99$. For each system, M points in \mathcal{X}_{Alg} were sampled from a normal distribution with zero mean and variance 25 times the identity matrix, for $M \in \{10, 20, 50, 100, 200, 500\}$. Table I reports statistics upon termination. The number of iterations is roughly linear in M , while computation time is roughly quadratic. The latter excludes the Bellman error measurement in line 11, on the basis that in practice, the

TABLE I
PERFORMANCE STATISTICS (MEAN \pm STANDARD DEVIATION) FOR VARIANT B OF ALGORITHM 1; 20 RANDOM 8-STATE, 3-INPUT SYSTEMS.

M	Iterations to termination	Computation time (s)	Number of cuts added ¹
10	217.3 \pm 57.9	0.313 \pm 0.120	160.8 \pm 31.7
20	485.7 \pm 74.5	1.054 \pm 0.296	356.0 \pm 71.0
50	1293 \pm 193	6.083 \pm 1.724	907.0 \pm 163.4
100	2736 \pm 491	25.09 \pm 8.60	1828 \pm 367
200	6043 \pm 1017	114.7 \pm 34.5	3696 \pm 626
500	16245 \pm 2527	777.1 \pm 215.3	9654 \pm 1644

¹Cut was not added if $\varepsilon(x_m, \pi(x_m; Q_I); Q_I) < 10^{-5}$ when x_m selected

convergence check for which it is used need not be carried out at every iteration.

It is likely that other ways of choosing m in line 15, e.g. largest Bellman error, would reduce the number of iterations required, although the assumptions under which finite convergence can be guaranteed may differ. Nevertheless, total times are already modest, and we note that alternative “exact” DP approaches such as value iteration [4, Ch. 2] and explicit MPC [2] are impractically expensive for problems of this size.

VI. CONCLUSION

This paper presented a general algorithm able to learn Q -functions, in the sense of minimizing Bellman error at arbitrary state space locations, for infinite-horizon problems. Convergence results were provided, both for fixed pairs (x, u) and for “policy-driven” pairs $(x, \pi(x))$. A further variant of Algorithm 1 could augment \mathcal{X}_{Alg} with sequences of states $\{x_\tau\}$ following the policy at each iteration, i.e., $x_{\tau+1} = f(x_\tau, \pi(x_\tau; Q_I))$. This would potentially learn a Q -function that approaches Q^* around entire trajectories, which is stronger than minimizing $\varepsilon(x_m, \pi(x_m; Q_I); Q_I)$ in individual locations $x_m \in \mathcal{X}_{\text{Alg}}$.

An added attraction of our formulation is that in many cases problem (5) remains convex even when the Q -function is not convex in the state. Future work will investigate such situations, and consider an extension to stochastic systems.

ACKNOWLEDGEMENT

The author thanks Rahul Jain of the University of Southern California for valuable discussions on the topic of this paper.

REFERENCES

- [1] P. N. Beuchat, J. C. Warrington, and J. Lygeros. Point-wise Maximum Approach to Approximate Dynamic Programming. In *IEEE Conference on Decision and Control*, Melbourne, Australia, 2017.
- [2] F. Borrelli, A. Bemporad, and M. Morari. *Predictive Control for Linear and Hybrid Systems*. Cambridge Univ. Press, 2017.
- [3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.
- [4] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst. *Reinforcement learning and dynamic programming using function approximators*, volume 39. CRC press, 2010.
- [5] C. S. Chow and J. N. Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control*, 36(8):898–914, 1991.
- [6] D. P. de Farias and B. Van Roy. The Linear Programming Approach to Approximate Dynamic Programming. *Operations Research*, 51(6):850–865, 2003.
- [7] A. M. Devraj and S. Meyn. Zap Q-Learning. *Advances in Neural Information Processing Systems (NIPS)* 30, pages 2235–2244, 2017.

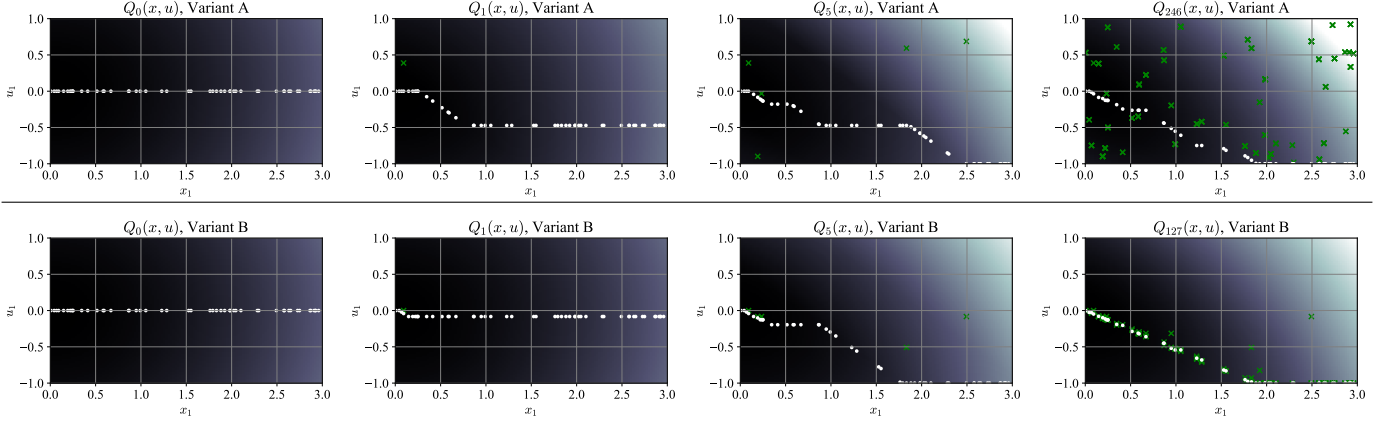


Fig. 2. Evolution of $Q_I(x, u)$ under Variant A (top) and Variant B (bottom) of Algorithm 1, for a 1-dimensional system. Lighter shading indicates a higher function value. Green markers show points (x_m, u_m) cumulatively visited during iterations. White points show the policy $u = \pi(x_m; Q_I)$ for each x_m in \mathcal{X}_{Alg} (Variant A) or \mathcal{X}_{Alg} (Variant B). The last plot in each row shows the last iteration, on which the termination criterion was satisfied. The policy returned by Variant B is closer to the optimal piecewise affine policy, and the (x, u) points visited concentrate around this policy in the limit.

- [8] A. M. Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10(4):237–260, 1972.
- [9] M. Hohmann, J. Warrington, and J. Lygeros. A Moment and Sum-of-Squares Extension of Dual Dynamic Programming with Application to Nonlinear Energy Storage Problems. *arXiv:1807.05947*, 2018.
- [10] B. Lincoln and A. Rantzer. Relaxing dynamic programming. *IEEE Transactions on Automatic Control*, 51(8):1249–1260, 2006.
- [11] H. Mania, A. Guy, and B. Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv:1803.07055*, 2018.
- [12] M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1-3):359–375, 1991.
- [13] T. H. Summers, K. Kunz, N. Kariotoglou, M. Kamgarpour, S. Summers, and J. Lygeros. Approximate dynamic programming via sum of squares programming. In *European Control Conf. (ECC)*, pp. 191–197, 2013.
- [14] Y. Wang and S. Boyd. Performance bounds for linear stochastic control. *Systems & Control Letters*, 58(3):178–182, 2009.
- [15] Y. Wang, B. O’Donoghue, and S. Boyd. Approximate dynamic programming via iterated Bellman inequalities. *International Journal of Robust and Nonlinear Control*, 25(10):1472–1496, 2015.
- [16] J. Warrington, P. N. Beuchat, and J. Lygeros. Generalized Dual Dynamic Programming for Infinite Horizon Problems in Continuous State and Action Spaces, rev. Sep 2018. *arXiv: 1711.07222*.
- [17] C. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge, UK, 1989.

APPENDIX

An example of a class of problems where Assumptions 2 and 3 hold is that which we refer to as **C-LQR**, for which:

- $f(x, u) = Ax + Bu$;
- $\ell(x, u) = \frac{1}{2}x^\top Qx + \frac{1}{2}u^\top Ru$, with $Q, R \succeq 0$;
- $h(x, u) = Dx + Eu - \bar{h}$, defining decoupled state and input constraints, where the latter are compact;
- $\gamma\|A\| < 1$, meaning “discounted-asymptotically” stable.

Proposition A.1. *Any problem of class C-LQR satisfies Assumptions 2 and 3.*

Proof. Assumption 2 is satisfied trivially. The lower-bounding functions in constraint (9d) have the quadratic form

$$q_i(x, u) = \frac{1}{2}x^\top Qx + \frac{1}{2}u^\top Ru + \nu_i^\top (Ax + Bu) + \xi_i, \quad (13)$$

and thus the problem remains convex at each iteration I . A Slater point exists, namely any feasible (x', u') together with

any $\alpha > q_i(x', u') \forall i$. Thus the strong duality condition in Assumption 3 holds.

To prove Lipschitz continuity in Assumption 3, one must bound the gradient in u -space of the functions $q_i(x_m, \cdot)$ for any given x_m . Inspection of problem (9) shows that each new function q_{I+1} depends on the existing functions q_0, \dots, q_I , and (13) shows that, due to u -compactness, a Lipschitz constant exists if the sequence $\{\|\nu_I\|\}_{I=0}^\infty$ is bounded. The KKT optimality conditions of (9) include the stationarity equations

$$\begin{aligned} \nu_{I+1} &= D^\top \lambda_{c,I+1} + \sum_{i=0}^I \lambda_{\alpha,i} (Qx' + A^\top \nu_i), \\ 0 &= E^\top \lambda_{c,I+1} + \sum_{i=0}^I \lambda_{\alpha,i} (Ru' + B^\top \nu_i), \\ \gamma &= \sum_{i=0}^I \lambda_{\alpha,i}. \end{aligned}$$

Without loss of generality, one can redefine the system with a linearly scaled input, $B \rightarrow \tilde{B}$ and $R \rightarrow \tilde{R}$, such that $\|\tilde{B}\| = 1$, and then linearly scale the input constraints in $h(x, u)$ such that $\|E\| = 1$. Triangle inequalities then yield

$$\begin{aligned} \|\nu_{I+1}\| &\leq \|D\| \cdot \|\lambda_{c,I+1}\| + \gamma \|Q\| \cdot \|x'\| \\ &\quad + \|A\| \cdot \left\| \sum_{i=0}^I \lambda_{\alpha,i} \nu_i \right\|, \\ \|\lambda_{c,I+1}\| &\leq \gamma \|\tilde{R}\| \cdot \|u'\| + \left\| \sum_{i=0}^I \lambda_{\alpha,i} \nu_i \right\|. \end{aligned}$$

As $\mathcal{U}(x_m)$ is compact and x_m is fixed, the norms of $x' = Ax_m + Bu$ and u' are both bounded by some constants X and U respectively. Eliminating $\|\lambda_{c,I+1}\|$, one obtains

$$\begin{aligned} \|\nu_{I+1}\| &\leq \|D\| \left(\gamma U \|\tilde{R}\| + \left\| \sum_{i=0}^I \lambda_{\alpha,i} \nu_i \right\| \right) \\ &\quad + \gamma X \|Q\| + \|A\| \cdot \left\| \sum_{i=0}^I \lambda_{\alpha,i} \nu_i \right\| \\ &= \gamma (U \|D\| \cdot \|\tilde{R}\| + X \|Q\|) \\ &\quad + (\|D\| + \|A\|) \left\| \sum_{i=0}^I \lambda_{\alpha,i} \nu_i \right\| \\ &\leq \gamma (U \|D\| \cdot \|\tilde{R}\| + X \|Q\|) \\ &\quad + \gamma (\|D\| + \|A\|) \max_{i=0, \dots, I} \|\nu_i\|. \end{aligned}$$

Thus, $\{\|\nu_I\|\}_{I=0}^\infty$ can grow no larger than $\frac{\gamma(U\|D\|\cdot\|\tilde{R}\|+X\|Q\|)}{1-\gamma(\|D\|+\|A\|)}$. As the state and input constraints are decoupled, $\|D\|$ can be made arbitrarily small by scaling the relevant rows of $h(x, u)$. Thus the denominator can be made strictly positive, and functions q_i of the form (13) are Lipschitz continuous. \square