A Convex Optimization Approach to Dynamic Programming in Continuous State and Action Spaces

Insoon Yang

Received: 15 January 2020 / Accepted: 3 September 2020 / Published online: 14 September 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In this paper, a convex optimization-based method is proposed for numerically solving dynamic programs in continuous state and action spaces. The key idea is to approximate the output of the Bellman operator at a particular state by the optimal value of a convex program. The approximate Bellman operator has a computational advantage because it involves a convex optimization problem in the case of control-affine systems and convex costs. Using this feature, we propose a simple dynamic programming algorithm to evaluate the approximate value function at pre-specified grid points by solving convex optimization problems in each iteration. We show that the proposed method approximates the optimal value function with a uniform convergence property in the case of convex optimal value functions. We also propose an interpolation-free design method for a control policy, of which performance converges uniformly to the optimum as the grid resolution becomes finer. When a nonlinear control-affine system is considered, the convex optimization approach provides an approximate policy with a provable suboptimality bound. For general cases, the proposed convex formulation of dynamic programming operators can be modified as a nonconvex bilevel program, in which the inner problem is a linear program, without losing the uniform convergence properties.

Keywords Dynamic programming · Convex optimization · Optimal control · Stochastic control

Mathematics Subject Classification 90C39 · 49L20 · 90C25

Communicated by Lars Grüne.

Insoon Yang
insoonyang@snu.ac.kr

1 Department of Electrical and Computer Engineering, Automation Systems Research Institute, Seoul National University, Seoul, South Korea
1 Introduction

Dynamic programming (DP) has been one of the most important methods for solving and analyzing sequential decision-making problems in optimal control, dynamic games, and reinforcement learning, among others. Using DP, we can decompose a complicated sequential decision-making problem into multiple tractable subproblems, of which optimal solutions are used to construct an optimal policy of the original problem. Numerical methods for DP are well studied for discrete-time Markov decision processes (MDPs) with discrete state and action spaces (e.g., [1]) and continuous-time deterministic and stochastic optimal control problems in continuous state spaces (e.g., [2]). This paper focuses on the discrete-time case with continuous state and action spaces in the finite-horizon setting. Unlike infinite-dimensional linear programming (LP) methods (e.g., [3–5]), which require a finite-dimensional approximation of the LP problems, we develop a finite-dimensional convex optimization-based method, which uses a discretization of the state space, while not discretizing the action space. Moreover, we assume that a system model and a cost function are explicitly known unlike the literature on reinforcement learning (RL) (e.g., see [6–8] and the references therein). Note that our focus is not to resolve the scalability issue in DP. As opposed to the RL algorithms that seek approximate solutions to possibly high-dimensional problems, our method is useful, when a provable convergence guarantee is needed for problems with relatively low-dimensional state spaces.1

Several discretization methods have been developed for discrete-time DP problems in continuous (Borel) state and action spaces. These methods can be assigned to two categories. The first category discretizes both state and action spaces. Bertsekas [9] proposed two discretization methods and proved their convergence under a set of assumptions, including Lipschitz-type continuity conditions. Langen [10] studied the weak convergence of an approximation procedure, although no explicit error bound was provided. However, the discretization method, proposed by Whitt [11] and Hinderer [12], is shown to be convergent and to have error bounds. Unfortunately, these error bounds are sensitive to the choice of partitions, and additional compactness and continuity assumptions are needed to reduce the sensitivity. Chow and Tsitsiklis [13] developed a multi-grid algorithm, which could be more efficient than its single-grid counterpart in achieving a desired level of accuracy. The discretization procedure, proposed by Dufour and Prieto-Rumeau [14], can handle unbounded cost functions and locally compact state spaces. However, it still requires the Lipschitz continuity of some components of dynamic programs. A measure concentration result for the Wasserstein metric has also been used to measure the accuracy of the approximation method in [15] for Markov control problems in the average cost setting. Unlike the aforementioned approaches, the finite-state and finite-action approximation method for MDPs with \( \sigma \)-compact state spaces, proposed by Saldi et al., does not rely on any Lipschitz-type continuity conditions [16].

The second category of discretization methods uses computational grids only for the state space, i.e., this class of methods does not discretize the action space. These approaches have a computational advantage over the methods in the first category,

---

1 However, our method is suitable for problems with high-dimensional action spaces.
particularly when the action space dimension is large. The state space discretization procedures, proposed by Hernández-Lerma [17], are shown to have a convergence guarantee with an error bound under Lipschitz continuity conditions on elements of control models. However, they are subject to the issue of local optimality in solving the nonconvex optimization problem over the action space involved in the Bellman equation. Johnson et al. [18] suggested spline interpolation methods, which are computationally efficient in high-dimensional state and action spaces. Unfortunately, these spline-based approaches do not have a convergence guarantee or an explicit suboptimality bound. Furthermore, the Bellman equation approximated by these methods involves nonconvex optimization problems.

The method proposed in this paper is classified into the second category of discretization procedures. Specifically, our approach only discretizes the state space, and thus, it can handle high-dimensional action spaces. The key idea is to use auxiliary optimization variables that assign the contribution of each grid point when evaluating the value function at a particular state. By doing so, we can avoid an explicit interpolation of the optimal value function and control policies evaluated at the pre-specified grid points in the state space, unlike most existing methods. The contributions of this work are threefold. First, we propose an approximate version of the Bellman operator and show that the corresponding approximate value function converges uniformly to the optimal value function \( v^* \) when \( v^* \) is convex. The proposed approximate Bellman operator has a computational advantage because it involves a convex optimization problem in the case of linear systems and convex costs. Thus, we can construct a control policy, of which performance converges uniformly to the optimum, by solving \( M \) convex programs in each iteration of the DP algorithm, where \( M \) is the number of grid points required for the desired accuracy. Second, we show that the proposed convex optimization approach provides an approximate policy with a provable suboptimality bound in the case of control-affine systems. This error bound is useful when gauging the performance of the approximate policy relative to an optimal policy. Third, we propose a modified version of our approximation method for general cases by localizing the effect of the auxiliary variables. The modified Bellman operator involves a nonconvex bilevel optimization problem wherein the inner problem is a linear program. We show that both the approximate value function and the cost-to-go function of a policy obtained by this method converge uniformly to the optimal value function if a globally optimal solution to the nonconvex bilevel problem can be evaluated; related error bounds are also characterized. The convergence property and the performance of our methods are demonstrated through three DP problems.

2 The Setup

2.1 Notation

Let \( \mathcal{B}_b(\mathcal{X}) \) denote the set of bounded measurable functions on \( \mathcal{X} \), equipped with the sup norm \( \| v \|_\infty := \sup_{x \in \mathcal{X}} |v(x)| < +\infty \). Given a measurable function \( w : \mathcal{X} \to \mathbb{R} \), let \( \mathcal{B}_w(\mathcal{X}) \) denote the set of measurable functions \( v \) on \( \mathcal{X} \) such that \( \| v \|_w := \sup_{x \in \mathcal{X}} (|v(x)|/w(x)) < +\infty \). Let \( \mathcal{L}(\mathcal{X}) \) denote the set of lower semicontinuous
functions on $\mathcal{X}$. Finally, we let $\mathbb{L}_b(\mathcal{X}) := \mathbb{L}(\mathcal{X}) \cap \mathbb{B}_b(\mathcal{X})$ and $\mathbb{L}_w(\mathcal{X}) := \mathbb{L}(\mathcal{X}) \cap \mathbb{B}_w(\mathcal{X})$.

### 2.2 Dynamic Programming Problems

Consider a discrete-time Markov control system of the form

$$x_{t+1} = f(x_t, u_t, \xi_t),$$

where $x_t \in \mathcal{X} \subseteq \mathbb{R}^n$ is the state, $u_t \in \mathcal{U}(x_t) \subseteq \mathcal{U} \subseteq \mathbb{R}^m$ is the control input. The stochastic disturbance process $\{\xi_t\}_{t \geq 0}$, $\xi_t \in \mathcal{E} \subseteq \mathbb{R}^l$, is i.i.d. and defined on a standard filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$. The sets $\mathcal{X}, \mathcal{U}(x_t), \mathcal{U}$ and $\mathcal{E}$ are assumed to be Borel sets, and the function $f : \mathcal{X} \times \mathcal{U} \times \mathcal{E} \rightarrow \mathcal{X}$ is assumed to be measurable.

A history up to stage $t$ is defined by $h_t := (x_0, u_0, \ldots, x_{t-1}, u_{t-1}, x_t)$. Let $H_t$ be the set of histories up to stage $t$ and $\pi_t$ be a stochastic kernel from $H_t$ to $\mathcal{U}$. The set of admissible policies is chosen as $\Pi := \{\pi = (\pi_0, \ldots, \pi_{K-1}) : \pi_t(\mathcal{U}(x_t)|h_t) = 1 \ \forall h_t \in H_t\}$. Our goal is to solve the following finite-horizon stochastic optimal control problem:

$$\inf_{\pi \in \Pi} \mathbb{E}^\pi \left[ \sum_{t=0}^{K-1} r(x_t, u_t) + q(x_K) \mid x_0 = x \right],$$

where $r : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ is a measurable stage-wise cost function, $q : \mathcal{X} \rightarrow \mathbb{R}$ is a measurable terminal cost function, and $\mathbb{E}^\pi$ represents the expectation taken with respect to the probability measure induced by a policy $\pi$. Under the following standard assumption for the measurable selection condition, there exists a deterministic Markov policy, which is optimal (e.g., [3, Condition 3.3.3, Theorem 3.3.5]) and [19, Assumptions 8.5.1–8.5.3, Lemma 8.5.5]). In other words, we can find an optimal policy $\pi^*$, which is deterministic and Markov, i.e., $\pi^* \in \Pi^{DM} := \{\pi = (\pi_0, \ldots, \pi_{K-1}) : \pi_t(x) = u \in \mathcal{U}(x), \pi_t \text{ is measurable}\}$.

**Assumption 1** Let $\mathcal{K} := \{(x, u) : x \in \mathcal{X}, u \in \mathcal{U}(x)\}$.

(i) The control set $\mathcal{U}(x)$ is compact for each $x \in \mathcal{X}$, and the multifunction $x \mapsto \mathcal{U}(x)$ is upper semicontinuous;

(ii) The real-valued functions $r$ and $q$ are lower semicontinuous on $\mathcal{K}$ and $\mathcal{X}$, respectively.

(iii) There exist nonnegative functions $\tilde{r}$, $\tilde{q}$ and $\beta$, with $\beta \geq 1$, and a continuous weight function $w : \mathcal{X} \rightarrow [1, \infty]$ such that $|r(x, u)| \leq \tilde{r}w(x)$, $|q(x)| \leq \tilde{q}w(x)$, and $\mathbb{E}[w(f(x, u, \xi))] \leq \beta w(x)$ for all $(x, u) \in \mathcal{K}$;

(iv) The function $(x, u) \mapsto f(x, u, \xi)$ is continuous on $\mathcal{K}$ for each $\xi \in \mathcal{E}$.

(v) The support $\mathcal{E}$ is finite, i.e., $\mathcal{E} := \{\xi^{[1]}, \ldots, \xi^{[N]}\}$ for some $\xi^{[s]}$'s in $\mathbb{R}^l$. 

\( \diamond \) Springer
Note that by (iii), (iv) and (v), $(x, u) \mapsto \mathbb{E}[w(f(x, u, \xi))]$ is continuous on $\mathcal{K}$. For any $v \in \mathbb{B}_w(\mathcal{X})$, let

$$(Tv)(x) := \inf_{u \in \mathcal{U}(x)} r(x, u) + \mathbb{E}[v(f(x, u, \xi))].$$  \hfill(2)$$

Under Assumption 1, the dynamic programming (DP) operator $T$ maps $\mathbb{L}_w(\mathcal{X})$ into itself by [19, Lemma 8.5.5]. Let $v_{opt}^t : \mathcal{X} \to \mathbb{R}$, $t = 0, \ldots, K$, be defined by the following Bellman equation: $v_{opt}^t := T v_{opt}^{t+1}$ for $t = 0, \ldots, K - 1$ with $v_{opt}^K \equiv q$. Then, $v_{opt}^t \in \mathbb{L}_w(\mathcal{X})$, and $v_{opt}^t$ is the optimal value function, i.e.,

$$v_{opt}^t(x) = \inf_{\pi \in \Pi} \mathbb{E}^\pi\left[ \sum_{k=t}^{K-1} r(x_k, u_k) + q(x_K) | x_t = x \right]$$

under Assumption 1. Given a measurable function $\varphi : \mathcal{X} \to \mathcal{U}$, let

$$(Tv^\varphi)(x) := r(x, \varphi(x)) + \mathbb{E}[v(f(x, \varphi(x), \xi))].$$

Under Assumption 1, there exists a measurable function $\pi_{opt}^t : \mathcal{X} \to \mathcal{U}$ such that $\pi_{opt}^t(x) \in \mathcal{U}(x)$, and $v_{opt}^t(x) = T^{\pi_{opt}^t} v_{opt}^{t+1}(x)$ for all $x \in \mathcal{X}$ and all $t$. Then, the policy $\pi_{opt} := (\pi_{opt}^0, \ldots, \pi_{opt}^{K-1}) \in \Pi^{DM}$ is an optimal solution to (1) by the dynamic programming principle [3, Theorem 3.2.1]. In practice, Assumption 1-(v) can be relaxed using sampling-based methods to a probability distribution with a continuous support. Specifically, the proposed convex optimization approach can be used in conjunction with the sampling-based (or ‘empirical’) method in [20–22]. The convergence of the sampling-based method has been shown under restrictive conditions such as compact state spaces or finite action spaces. The validity of such a combination is numerically tested in Sect. 5.1. Its extension using stochastic subgradient methods can be found in [23].

### 2.3 State Space Discretization

Our method aims to approximate the optimal value function $v_{opt}^t$ at pre-specified nodes in a convex compact set $\mathcal{Z}_t \subseteq \mathcal{X}$ for all $t$. We select a sequence $\{\mathcal{Z}_t\}_{t=0}^K$ of convex compact sets such that $\mathcal{Z}_0 \subseteq \mathcal{Z}_1 \subseteq \cdots \subseteq \mathcal{Z}_K \subseteq \mathcal{X}$, and

$$\{ f(x, u, \xi) : x \in \mathcal{Z}_t, u \in \mathcal{U}(x), \xi \in \mathcal{E} \} \subseteq \mathcal{Z}_{t+1}. \hfill(3)$$

The existence of such a sequence is guaranteed under Assumption 1 because $f$ is continuous in $(x, u)$, $\mathcal{U}(x)$ is compact, and $\mathcal{E}$ is finite. We note that $\mathcal{Z}_K \neq \mathcal{X}$ in general, and the state space $\mathcal{X}$ does not have to be compact. Several examples satisfying these two conditions can be found in Sect. 5. This choice of computational domains is also used in [9]. With such a sequence, we can evaluate $v_{opt}^t$ on $\mathcal{Z}_t$ by using only information about $v_{opt}^{t+1}$ on $\mathcal{Z}_{t+1}$.
We choose $N_C$ convex polytopes $C_1, \ldots, C_{N_C} \subseteq Z_K$ satisfying the following properties:

1. $\bigcup_{i=1}^{N_C} C_i = Z_K$;
2. $C_i^o \cap C_j^o = \emptyset$ for $i \neq j$, where $C_i^o$ denotes the interior of $C_i$;
3. For each $t$, there exists a subsequence $\{C_{i_j}\}_{j=1}^{N_C,t}$ of $\{C_i\}_{i=1}^{N_C}$ such that $\bigcup_{j=1}^{N_C,t} C_{i_j} = Z_t$.

We relabel $C_i$, if necessary, so that $\bigcup_{i=1}^{N_C,t} C_i = Z_t$ for all $t$. We note that $N_{C,K} = N_C$, and $N_{C,0} \leq N_{C,1} \leq \cdots \leq N_{C,K}$.

Let $M_t$ denote the number of nodes (or grid points) in $Z_t$. The nodes $x^{[1]}, \ldots, x^{[M_K]} \in Z_K$ are chosen such that each $C_i$ is the convex hull of a subset of the nodes. A concrete way to construct $Z_t$’s and $C_i$’s using a rectilinear grid is described in Appendix. For example, in Fig. 1, $C_1$ is the convex hull of $\{x^{[1]}, x^{[2]}, x^{[4]}, x^{[5]}\}$. The maximum of the diameter of the sets $\{C_i\}_{i=1}^{N_C}$ is given by $\delta := \max_{k=1,\ldots,N_C} \max_{x^{[i]},x^{[j]} \in C_k} \|x^{[i]} - x^{[j]}\|$.

For example, in Fig. 1, $\delta$ is equal to the length of $C_i$’s diagonal. We also relabel $x^{[i]}$, if necessary, so that $x^{[1]}, \ldots, x^{[M_t]} \in Z_t$ for all $t$.

### 2.4 Value Functions on Restricted Spaces

Let $v_t^* : Z_t \rightarrow \mathbb{R}$ be the optimal value function restricted on $Z_t$ for each $t$. In other words, $v_t^* (x) = v_t^{opt} (x)$ for all $x \in Z_t$. Given any $v \in \mathbb{B} (Z_{t+1})$, let
\[(T_t v)(x) := \inf_{u \in \mathcal{U}(x)} \left[ r(x, u) + \mathbb{E}[v(f(x, u, \xi))] \right] \]

\[
= \inf_{u \in \mathcal{U}(x)} \left[ r(x, u) + \sum_{s=1}^{N} p_s v(f(x, u, \hat{\xi}^{[s]})) \right] \tag{4}
\]

for all \(x \in \mathcal{Z}_t\), where \(p_s := \text{Prob}(\xi = \hat{\xi}^{[s]}) \in [0, 1]\) for each \(s\), and thus \(\sum_{s=1}^{N} p_s = 1\) by definition. Under Assumption 1, the restricted optimal value functions satisfy \(v_t^*(x) = (T_t v_{t+1}^*)(x)\) for all \(x \in \mathcal{Z}_t\). For any \(x \in \mathcal{Z}_t\), \(v_t^*(x)\) can be computed by using \(v_{t+1}^*\), given our choice of \(\mathcal{Z}_t\)'s in Sect. 2.3. This computation does not require information about \(v_{t+1}^{opt}\) outside \(\mathcal{Z}_{t+1}\). Our goal is to develop a convex optimization-based approach to approximating the restricted optimal value function \(v_t^*\) in a convergent manner.

3 Convex Value Functions

3.1 Approximation of the Bellman Operator

When computing the optimal value functions via the dynamic programming algorithm \(v_t^* := T_t v_{t+1}^*\), a closed-form expression of \(v_t^*\) is unavailable in general. Unfortunately, it is challenging to solve the optimization problem in the definition (4) of \(T_t\) without a closed-form expression of \(v_t^*\). To resolve this issue, we propose a method that uses the value of \(v_t^*\) only at the nodes \(x^{[1]}, \ldots, x^{[M_t+1]}\) in \(\mathcal{Z}_{t+1}\). Given any \(v \in \mathbb{B}_b(\mathcal{Z}_{t+1})\), let

\[
(\hat{T}_t v)(x) := \inf_{u, \gamma} \left[ r(x, u) + \sum_{s=1}^{N} \sum_{i=1}^{M_t+1} p_s \gamma_{s,i} v(x^{[i]}) \right] \tag{5}
\]

\[
\text{s.t. } f(x, u, \hat{\xi}^{[s]}) = \sum_{i=1}^{M_t+1} \gamma_{s,i} x^{[i]} \quad \forall s \in \mathcal{S}
\]

\[
u \in \mathcal{U}(x), \quad \gamma_s \in \Delta \quad \forall s \in \mathcal{S}
\]

for each \(x \in \mathcal{Z}_t\), where \(\Delta\) is the \((M_t+1)\)-dimensional probability simplex, i.e., \(\Delta := \{\gamma \in \mathbb{R}^{M_t+1}: \sum_{i=1}^{M_t+1} \gamma_i = 1, \gamma_i \geq 0 \forall i = 1, \ldots, M_t+1\}\), and \(\mathcal{S} := \{1, \ldots, N\}\). Under Assumption 1, the objective function of (5) is lower semicontinuous in \((u, \gamma) \in \mathcal{U}(x) \times \Delta^N\), and the feasible set is compact. Therefore, by [3, Proposition D.5], (5) admits an optimal solution, and \(\hat{T}_t\) maps \(\mathbb{B}_b(\mathcal{Z}_{t+1})\) to \(\mathbb{B}_b(\mathcal{Z}_t)\). Moreover, \(\hat{T}_t\) is monotone, i.e., for any \(v, v' \in \mathbb{B}_b(\mathcal{Z}_{t+1})\) such that \(v \leq v'\), we have \(\hat{T}_t v \leq \hat{T}_t v'\).

The constrained-optimization problem in (5) for a fixed \(x \in \mathcal{Z}_t\) can be numerically solved to obtain a globally optimal solution using several existing convex optimization
algorithms (e.g., [24–26]) if the problem is convex, i.e., \( u \mapsto r(x, u) \) is a convex function and \( u \mapsto f(x, u, \xi[i]) \) is an affine function.\(^2\)

Note that for each sample index \( s \) the term \( v(f(x, u, \xi[x])) \) in the original Bellman operator is approximated by \( \sum_{i=1}^{M_{t+1}} \gamma_{s,i} v(x[i]) \), where the auxiliary weight variable \( \gamma_{s,i} \) can be interpreted as the degree to which \( f(x, u, \xi[x]) \) is represented by \( x[i] \). This idea of representing the next state as a convex combination of grid points or nodes has also been adopted in the analysis and design of semi-Lagrangian schemes for Hamilton–Jacobi–Bellman partial differential equations arising in continuous-time optimal control problems [27]. Although this paper focuses on discrete-time DP problems, it would be an interesting future research to extend our method to the continuous-time setting, possibly using advanced techniques developed in the recent literature on semi-Lagrangian methods (e.g., [28,29]).

When \( v \) is convex, we immediately observe that \( \hat{T}_t v \) is upper bounded by \( \hat{T}_t v \) on \( \mathcal{Z}_t \) because \( v(f(x, u, \xi[x])) = v(\sum_{i=1}^{M_{t+1}} \gamma_{s,i} x[i]) \leq \sum_{i=1}^{M_{t+1}} \gamma_{s,i} v(x[i]) \).

Further showing that \( T_t v \) is lower bounded by \( \hat{T}_t v - C \delta \) for some positive constant \( C \), we will prove that \( \hat{T}_t v \) converges uniformly to \( T_t v \) on \( \mathcal{Z}_t \) as the maximum distance \( \delta \) between neighboring nodes tends to zero. By definition, \( \hat{T}_t \) depends on the discretization parameter \( \delta \) as well as \( N \). For notational simplicity, however, we suppress the dependence on \( \delta \) and \( N \).

### 3.2 Error Bound and Convergence

In this section, we assume the convexity of \( v \) and show the uniform convergence property of our method. This assumption will be relaxed in Sect. 4.

**Proposition 3.1** Suppose that Assumption 1 holds, and that \( v \in \mathbb{L}_b(\mathcal{Z}_{t+1}) \) is convex. Then, we have

\[
(\hat{T}_t v)(x) - L_v \delta \leq (T_t v)(x) \leq (\hat{T}_t v)(x) \quad \forall x \in \mathcal{Z}_t,
\]

where \( L_v := \max_{j=1,\ldots,N_{t+1}} \sup_{x,x' \in \mathcal{C}_j} \| x - x' \| \). Therefore, \( \hat{T}_t v \) converges uniformly to \( T_t v \) on \( \mathcal{Z}_t \) as \( \delta \to 0 \).

Note that \( L_v < \infty \) because \( v \) is continuous on the compact set \( \mathcal{Z}_{t+1} \) under the assumption of convexity and lower semicontinuity.

**Proof** Fix an arbitrary \( x \in \mathcal{Z}_t \). We first show that \( (T_t v)(x) \leq (\hat{T}_t v)(x) \). Under Assumption 1, the objective function of (5) is lower semicontinuous, and the feasible set is compact. Let \( (\hat{u}, \hat{\gamma}) \in \mathcal{U}(x) \times \Delta^N \) be an optimal solution to (5), i.e., it satisfies \( (\hat{T}_t v)(x) = r(x, \hat{u}) + \sum_{s=1}^{N} \sum_{i=1}^{M_{t+1}} p_s \gamma_{s,i} v(x[i]) \) and \( f(x, \hat{u}, \xi[x]) \leq \sum_{i=1}^{M_{t+1}} \gamma_{s,i} x[i] \in \mathcal{Z}_{t+1} \) for all \( s \in S \). The convexity of \( v \) on \( \mathcal{Z}_{t+1} \) implies that

\[2\] More precisely, the set \( \mathcal{U}(x) \) needs to be represented by convex inequalities, i.e., there exist functions \( a_k : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R} \) and \( b_k : \mathcal{X} \to \mathbb{R} \) such that

\[
\mathcal{U}(x) := \{ u \in \mathbb{R}^m : a_k(x, u) \leq b_k(x), k = 1, \ldots, N_{ineq} \},
\]

where \( u \mapsto a_k(x, u) \) is a convex function for each fixed \( x \in \mathcal{X} \) and each \( k \).
\[ v(\sum_{i=1}^{M_{t+1}} \hat{y}_{s,i}x^{[i]}) \leq \sum_{i=1}^{M_{t+1}} \hat{y}_{s,i}v(x^{[i]}) \text{ for all } s \in S. \] Therefore, by the definition of \((\hat{u}, \hat{y})\), we have

\[
(\hat{T}_t v)(x) \geq r(x, \hat{u}) + \sum_{s=1}^{N} p_s v(\sum_{i=1}^{M_{t+1}} \hat{y}_{s,i}x^{[i]}) = r(x, \hat{u}) + \sum_{s=1}^{N} p_s v(f(x, \hat{u}, \xi^{[s]})) \geq \inf_{u \in U(x)} r(x, u) + \sum_{s=1}^{N} p_s v(f(x, u, \xi^{[s]})) = (T_t v)(x). \]

We now show that \((\hat{T}_t v)(x) - L_v \delta \leq (T_t v)(x)\). Under Assumption 1, the objective function of (4) is lower semicontinuous [19, Lemma 8.5.5], and the feasible set is compact. Thus, by [3, Proposition D.5], there exists \(u^* \in \mathcal{U}(x)\) such that \((T_t v)(x) = r(x, u^*) + \sum_{s=1}^{N} p_s v(f(x, u^*, \xi^{[s]}))\). Let \(y_s := f(x, u^*, \xi^{[s]})\) for all \(s \in S\). Then, \(y_s \in \mathcal{Z}_{t+1}\) because \(x \in \mathcal{Z}_t\). Thus, there exists a unique \(j \in \{1, \ldots, N_{C_{t+1}}\}\) such that \(y_s \in C_j\). Let \(N(y_s)\) denote the set of grid points on the cell \(C_j\) that contains \(y_s\). For each \(s\), choose \(\gamma^{*, s}_s \in \Delta\) such that \(f(x, u^*, \xi^{[s]}) = \sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]}\) and \(\sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]} = 1\). Note that \(\gamma^{*, s}_s = 0\) for all \(i \notin N(y_s)\). We then have

\[
(T_t v)(x) = r(x, u^*) + \sum_{s=1}^{N} p_s v(\sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]})
\]

By the definition of \(L_v\), \(v(\sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]}) \geq v(x^{[i]}) - L_v \|x^{[i]} - \sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]}\| \) for all \(i' \in N(y_s)\). This implies that

\[
v(\sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]}) \geq \max_{i' \in N(y_s)} \left[ v(x^{[i']}) - L_v \|x^{[i']} - \sum_{i \in N(y_s)} \gamma^{*, s}_s x^{[i]}\| \right]
\]

\[
\geq \max_{i' \in N(y_s)} \left[ v(x^{[i']}) - L_v \sum_{i \in N(y_s)} \gamma^{*, s}_s \|x^{[i']} - x^{[i]}\| \right]
\]

\[
\geq \max_{i' \in N(y_s)} \left[ v(x^{[i']}) - L_v \sum_{i \in N(y_s)} \gamma^{*, s}_s \delta \right] \geq \sum_{i' \in N(y_s)} \gamma^{*, s}_{s,i'} (v(x^{[i']}) - L_v \delta)
\]

\[
= \sum_{i' = 1}^{M_{t+1}} \gamma^{*, s}_{s,i'} v(x^{[i']}) - L_v \delta,
\]
where the third inequality holds by the definition of $\delta$, and the last equality holds because $\sum_{i' \in N(y, x)} y_{s, i'} = 1$. Combining (6) and (7) yields

$$(T_t v)(x) \geq r(x, u^*) + \sum_{s=1}^{N} \sum_{i=1}^{M_{t+1}} p_s y_{s, i}^* v(x^{[i]}) - L_t \delta \geq (\hat{T}_t v)(x) - L_v \delta,$$

where the second inequality holds because $(u^*, \gamma^*)$ is a feasible solution to (5). Therefore, the result follows.

Let $\hat{v}_t : Z_t \to \mathbb{R}$ be recursively defined by $\hat{v}_t(x) := (\hat{T}_t \hat{v}_{t+1})(x)$ for all $x \in Z_t$ with $\hat{v}_K \equiv q$ on $Z_K$. We now show that $\hat{v}_t$ converges uniformly to the optimal value function $v_t^*$ on $Z_t$ as $\delta$ tends to zero when all $v_t^*$’s are convex. A sufficient condition for the convexity of $v_t^*$’s is provided in Sect. 3.3.

**Theorem 3.1** (Uniform Convergence and Error Bound I) **Suppose that Assumption 1 holds, and that the optimal value function $v_t^*$ is convex on $Z_t$ for all $t = 0, \ldots, K$.** Then, we have

$$0 \leq \hat{v}_t(x) - v_t^*(x) \leq \sum_{k=t+1}^{K} L_k \delta \quad \forall x \in Z_t \quad \forall t = 0, \ldots, K,$$

where $L_k := \sup_{j=1, \ldots, N_{C, k}} \sup_{x, x' \in C_j : x \neq x'} \frac{\|v_t^*(x) - v_t^*(x')\|}{\|x - x'\|}$. Therefore, $\hat{v}_t$ converges uniformly to $v_t^*$ on $Z_t$ as $\delta \to 0$.

**Proof** We use mathematical induction to prove (8). For $t = K$, $\hat{v}_K = v_K^* \equiv q$ on $Z_K$, and thus (8) holds. Suppose that the induction hypothesis holds for some $t$. Applying the operator $\hat{T}_{t-1}$ to all the sides of (8) yields

$$(\hat{T}_{t-1} \hat{v}_t)(x) - \sum_{k=t+1}^{K} L_k \delta \leq (\hat{T}_{t-1} v_t^*)(x) \leq (\hat{T}_{t-1} \hat{v}_t)(x) \quad \forall x \in Z_{t-1}$$

by the monotonicity of $\hat{T}_{t-1}$. On the other hand, by Proposition 3.1,

$$(\hat{T}_{t-1} v_t^*)(x) - L_t \delta \leq (\hat{T}_{t-1} v_t^*)(x) \leq (\hat{T}_{t-1} v_t^*)(x) \quad \forall x \in Z_{t-1}$$

because $v_t^* \in \mathbb{L}_b(Z_t)$ [19, Lemma 8.5.5] and it is convex. Therefore, we have

$$(\hat{T}_{t-1} \hat{v}_t)(x) - \sum_{k=t}^{K} L_k \delta \leq (\hat{T}_{t-1} v_t^*)(x) \leq (\hat{T}_{t-1} \hat{v}_t)(x) \quad \forall x \in Z_{t-1}.$$ 

Note that $\hat{v}_{t-1} = \hat{T}_{t-1} v_t$ and $v_{t-1}^* = T_{t-1} v_t^*$ on $Z_{t-1}$ by definition. This implies that $\hat{v}_{t-1}(x) - \sum_{k=t}^{K} L_k \delta \leq v_{t-1}^*(x) \leq \hat{v}_{t-1}(x)$ for all $x \in Z_{t-1}$ as desired.  \[ \square \]
The approximate value function $\hat{v}_{t+1}$ evaluated at the nodes $x^{[1]}$, $\ldots$, $x^{[M_{t+1}]}$ for $t = 0, \ldots, K$ can be used to construct a deterministic stationary policy, $\hat{\pi} := (\hat{\pi}_0, \ldots, \hat{\pi}_{K-1})$, by setting

$$\hat{\pi}_t(x) = \arg\min_{u \in \mathcal{U}(x)} \min_{v \in \Delta^N} \left[ r(x, u) + \sum_{s=1}^{N} \sum_{i=1}^{M_{t+1}} \gamma_{s,i} \hat{v}_{t+1}(x^{[i]}) \right]$$

s.t. $f(x, u, \hat{\xi}[s]) = \sum_{i=1}^{M_{t+1}} \gamma_{s,i}x^{[i]} \quad \forall s \in S$

for all $x \in Z_t$. Let $v^K_t : Z_t \to \mathbb{R}$ be defined by

$$v^K_t(x) := (T^{\hat{\pi}_t} v^K_{t+1})(x) := r(x, \hat{\pi}_t(x)) + \sum_{s=1}^{N} \sum_{i=1}^{M_{t+1}} p_s v^K_{t+1} \left( f(x, \hat{\pi}_t(x), \hat{\xi}[s]) \right) \quad \forall x \in Z_t$$

with $v^K_0 = q$ on $Z_K$. It is straightforward to check that $T^{\hat{\pi}_t}$ is monotone. To show that the cost-to-go function $v^K_t$ of $\hat{\pi}$ converges uniformly to the optimal value function $v^*_t$, we first observe the following property:

**Lemma 3.1** Suppose that Assumption 1 holds, and that $\hat{v}_t$ is convex on $Z_t$ for all $t = 0, \ldots, K$. Then, $v^K_t(x) \leq \hat{v}_t(x)$ for all $x \in Z_t$ and $t = 0, \ldots, K$.

**Proof** We use mathematical induction. For $t = K$, $v^K_K = \hat{v}_K \equiv q$ on $Z_K$, and thus, the induction hypothesis holds. Suppose that $v^K_{t+1} \leq \hat{v}_{t+1}$ on $Z_{t+1}$ for some $t + 1$. By the monotonicity of $T^{\hat{\pi}_t}$, we have

$$(T^{\hat{\pi}_t} v^K_{t+1})(x) \leq (T^{\hat{\pi}_t} \hat{v}_{t+1})(x) \quad \forall x \in Z_t. \quad (9)$$

Fix an arbitrary $x \in Z_t$. By the definition of $\hat{T}_t$ and $\hat{\pi}_t$, under Assumption 1, there exists $\hat{\gamma} \in \Delta^N$ such that

$$(\hat{T}_t \hat{v}_{t+1})(x) = r(x, \hat{\pi}_t(x)) + \sum_{s=1}^{N} \sum_{i=1}^{M_{t+1}} \gamma_{s,i} \hat{v}_{t+1}(x^{[i]}) \quad (10)$$

and $f(x, \hat{\pi}_t(x), \hat{\xi}[s]) = \sum_{i=1}^{M_{t+1}} \gamma_{s,i}x^{[i]}$ for all $s \in S$. By the convexity of $\hat{v}_{t+1}$,

$$\sum_{i=1}^{M_{t+1}} \gamma_{s,i} \hat{v}_{t+1}(x^{[i]}) \geq \hat{v}_{t+1} \left( \sum_{i=1}^{M_{t+1}} \gamma_{s,i}x^{[i]} \right) = \hat{v}_{t+1} \left( f(x, \hat{\pi}_t(x), \hat{\xi}[s]) \right) \quad (11)$$

for each $s \in S$. Combining the inequalities (10) and (11) yields

$$(\hat{T}_t \hat{v}_{t+1})(x) \geq r(x, \hat{\pi}_t(x)) + \sum_{s=1}^{N} p_s \hat{v}_{t+1} \left( f(x, \hat{\pi}_t(x), \hat{\xi}[s]) \right) = (T^{\hat{\pi}_t} \hat{v}_{t+1})(x).$$
Recall (9), we conclude that for all $x \in Z_t$,

$$v_t^\delta(x) = (T^\pi_t v_{t+1}^\delta)(x) \leq (T^\pi_t \hat{v}_{t+1})(x) \leq (\hat{T}_t \hat{v}_{t+1})(x) = \hat{v}_t(x).$$

This completes mathematical induction, and the result follows. \qed

Using this lemma and Theorem 3.1, we obtain the following uniform convergence result:

**Theorem 3.2** (Uniform Convergence and Error Bound II) Suppose that Assumption 1 holds, and that $v_t^*$ and $\hat{v}_t$ are convex on $Z_t$ for all $t$. Then,

$$0 \leq v_t^\delta(x) - v_t^*(x) \leq \sum_{k=t+1}^K L_k \delta \quad \forall x \in Z_t \quad \forall t = 0, \ldots, K,$$

where $L_k := \sup_{j=1, \ldots, N_{c,k}} \sup_{x, x' \in C_j : x \neq x'} \frac{||v_t^*(x) - v_t^*(x')||}{||x - x'||}$. Therefore, $v_t^\delta$ converges uniformly to $v_t^*$ on $Z_t$ as $\delta \to 0$.

**Proof** Fix an arbitrary $t \in \{0, \ldots, K\}$. By Theorem 3.1 and Lemma 3.1, we first observe that $v_t^\delta(x) - v_t^*(x) \leq \hat{v}_t(x) - v_t^*(x) \leq \sum_{k=t+1}^K L_k \delta$ for all $x \in Z_t$. In addition, we have $v_t^* \leq v_t^\delta$ since $v_t^*$ is the minimal cost-to-go function under Assumption 1. Therefore, the result follows. \qed

### 3.3 Interpolation-Free DP Algorithm

The error bounds and convergence results established in the previous subsection are valid if $v_t^*$ and $\hat{v}_t$ are convex. We now introduce a sufficient condition for the convexity $v_t^*$ and $\hat{v}_t$.

**Assumption 2** (Linear-Convex Control)

1. The function $(x, u) \mapsto f(x, u, \xi)$ is affine on $K := \{(x, u) : x \in X, u \in U(x)\}$ for each $\xi \in \Xi$. In addition, the stage-wise cost function $r$ is convex on $K$, and the terminal cost function $q$ is convex on $X$.
2. If $u(k) \in U(x(k))$ for $k = 1, 2$, then $\lambda u(1) + (1 - \lambda) u(2)$ is an element of $U(\lambda x(1) + (1 - \lambda) x(2))$ for any $x(1), x(2) \in X$ and any $\lambda \in (0, 1)$.

**Lemma 3.2** Suppose that Assumptions 1 and 2 hold, and that $v : Z_{t+1} \to \mathbb{R}$ is convex. Then, $T_t v, \hat{T}_t v : Z_t \to \mathbb{R}$ are convex.

Its proof can be found in [30]. An immediate observation obtained from Lemma 3.2 is the convexity of $v_t^*$ and $\hat{v}_t$ for all $t = 0, \ldots, K$ because $v_K^*$ and $\hat{v}_K$ are convex on $Z_K$.

**Proposition 3.2** Under Assumptions 1 and 2, the functions $v_t^*$ and $\hat{v}_t$ are convex for all $t = 0, \ldots, K$. \qed
Algorithm 1: Interpolation-Free DP algorithm

1. Initialize $\hat{v}_K \equiv q$ on $Z_K$;
2. for $t = K - 1 : -1 : 0$ do
   3. Set $\hat{v}_t(x[i]) := (\hat{T}_t \hat{v}_{t+1})(x[i])$ by solving the problem in (5) with $v \equiv \hat{v}_{t+1}$ for each $i = 1, \ldots, M_t$;
   4. Set $\hat{\pi}_t(x[i])$ as an optimal $u$ of the problem in (5) with $v \equiv \hat{v}_{t+1}$ for each $i = 1, \ldots, M_t$;
5. end

By Proposition 3.2, the error bounds and convergence results in Theorems 3.1 and 3.2 are valid under Assumptions 1–2. Note, however, that Assumption 2 is merely a sufficient condition for convexity.

Algorithm 1, which is based on the proposed method, can be used to evaluate the approximate value function $\hat{v}_t$ and to construct the corresponding policy $\hat{\pi}_t$ at all grid point $x[i]$'s in $Z_t$. Given $\hat{v}_t(x[i])$ for all $i = 1, \ldots, M_t$, the value of $\hat{v}_t$ at the other points in $Z_t$, the approximate value function can be computed using (5) with $v := \hat{v}_{t+1}$. We also observe that the optimization problem in (5) used to evaluate $(\hat{T}_t v)(x)$ is convex regardless of the convexity of $v$ under Assumption 2. Thus, in each iteration of our DP algorithm, it suffices to solve $M_t$ convex optimization problems, each of which is for $x := x[i]$.

Proposition 3.3 Under Assumption 2, the optimization problem in (5) for any fixed $x \in Z_t$ is a convex optimization problem if $v \in B_b(Z_{t+1})$.

Proof Fix an arbitrary state $x \in Z_t$. The objective function is convex in $u$ and linear in $\gamma$ (even when $v$ is nonconvex). Furthermore, the equality constraints are linear in $(u, \gamma)$. Assumption 2 implies that $U(x)$ is convex. Also, the probability simplex $\Delta$ is convex. Therefore, this optimization problem for any fixed $x$ is convex. □

Remark 3.1 (Interpolation-free property) Note that we can evaluate $\hat{v}_t$ at an arbitrary $x \in Z_t$ by using the definition (5) of $\hat{T}_t$ without any explicit interpolation that may introduce additional numerical errors. This feature is also useful when the output of $\hat{T}_t$ needs to be specified at a particular state that is different from the grid points $\{x[i]\}_{i=1}^{M_t}$. Unlike many existing discretization-based methods, our approach does not require a separate interpolation stage in constructing both the optimal value function and control policies.

4 Nonconvex Value Functions

The convexity of the optimal value function plays a critical role in obtaining the convergence results in the previous section. To relax the convexity condition (e.g., Assumption 2), we first show that the proposed approximation method is useful when constructing a suboptimal policy with a provable error bound in the case of nonlinear control-affine systems with convex cost functions. For further general cases, we propose a modified method based on a nonconvex bi-level optimization problem, where the inner problem is a linear program.

 Springer
4.1 Control-Affine Systems

Consider a control-affine system of the form
\[ x_{t+1} = f(x_t, u_t, \xi_t) := g(x_t, \xi_t) + h(x_t, \xi_t)u_t. \]  
(12)

More precisely, we assume the following:

**Assumption 3** The function \( f : \mathcal{X} \times \mathcal{U} \times \mathcal{E} \rightarrow \mathcal{X} \) can be expressed as (12), where \( g : \mathcal{X} \times \mathcal{E} \rightarrow \mathbb{R}^n \) and \( h : \mathcal{X} \times \mathcal{E} \rightarrow \mathbb{R}^{n \times m} \) are (possibly nonlinear) measurable functions such that \( g(\cdot, \xi) \) and \( h(\cdot, \xi) \) are continuous for each \( \xi \in \mathcal{E} \). In addition, \( u \mapsto r(x, u) \) is convex on \( \mathcal{U}(x) \) for each \( x \in \mathcal{X} \).

Note that the condition on \( r \) imposed by this assumption is weaker than Assumption 2 which requires the joint convexity of \( r \). In this setting, each iteration of the DP algorithm in Sect. 3.3 still involves \( M_t \) convex optimization problems. This can be shown by using the same argument as the proof of Proposition 3.3.

**Proposition 4.1** Under Assumption 3, the optimization problem in (5) is a convex program.

Due to the nonconvexity of the optimal value function, \( \hat{\nu}_t \) obtained by value iteration in the previous section is no longer guaranteed to converge to the optimal value function as \( \delta \) tends to zero. However, we are still able to characterize an error bound for the approximate policy \( \hat{\pi} \) as follows:

**Proposition 4.2** (Error Bound) Suppose that Assumptions 1 and 3 hold, and that \( v^*_t \) is locally Lipschitz continuous on \( \mathcal{Z}_t \) for each \( t \). Then, for all \( t = 0, \ldots, K \), we have
\[
0 \leq v^*_t(x) - v^*_t(x) \leq v^*_t(x) - v^*_t(x) + \sum_{k=t+1}^{K} L_k \delta \quad \forall x \in \mathcal{Z}_t, \]
where \( L_k := \sup_{j=1,\ldots,N_{C,j}} \sup_{x, x' \in C_j : x \neq x'} \frac{\|u^*_k(x) - u^*_k(x')\|}{\|x - x'\|} \).

**Proof** Fix an arbitrary \( t \in \{0, \ldots, K\} \). By the optimality of \( v^*_t \), we have \( v^*_t(x) \leq v^*_t(x) \) for all \( x \in \mathcal{Z}_t \). Note that in the second part of the proof of Proposition 3.1, the convexity of \( v \) is unused. It only requires the continuity of \( v \) on \( \mathcal{Z}_t \). Therefore, the second inequality of (8) in Theorem 3.1 holds when \( v^*_t \) is continuous, i.e., \( \hat{\nu}_t(x) - v^*_t(x) \leq \sum_{k=t+1}^{K} L_k \delta \) for all \( x \in \mathcal{Z}_t \). Thus, the result follows.

This proposition implies that the performance of \( \hat{\pi} \) converges to the optimum as \( \delta \) tends to zero if \( v^*_t \) converges to the approximate value function \( \hat{\nu}_t \) for all \( t \). Otherwise, it is possible that \( \hat{\nu}_t \) converges to some function which is different from the optimal value function. Or it may oscillate within the error bound. However, this \textit{a posteriori} error bound is useful when we need to design a controller with a provable performance guarantee, for example, in safety-critical systems where the objective is to maximize the probability of safety (e.g., [31]), as this problem is subject to nonconvexity issues [32].


4.2 General Case

In the case of general nonlinear systems with nonlinear stage-wise cost functions, we modify the operator $\hat{T}_i$ as follows. For any $v \in \mathbb{B}_b(Z_{t+1})$, let

$$
(\hat{T}_i v)(x) := \inf_{u \in \mathcal{U}(x)} \left[ r(x, u) + (H_i v)(x, u) \right] \quad \forall x \in Z_t,
$$

where

$$
(H_i v)(x, u) := \inf_{y} \sum_{s=1}^{N} \sum_{i \in \mathcal{N}(y_s)} p_s y_{s,i} u(x^{[i]})
$$

s.t. $y_s = \sum_{i \in \mathcal{N}(y_s)} y_{s,i} x^{[i]} \quad \forall s \in \mathcal{S}$

$$
y_s \in \Delta \quad \forall s \in \mathcal{S}
$$

$$
y_{s,i} = 0 \quad \forall i \notin \mathcal{N}(y_s) \quad \forall s \in \mathcal{S}
$$

for each $(x, u) \in Z_t \times \mathcal{U}$ such that $u \in \mathcal{U}(x)$, and $y_s = f(x, u, \xi^{[s]})$ for all $s \in \mathcal{S}$. Here, $\mathcal{N}(y_s)$ denotes the set of grid points on the cell $\mathcal{C}_j$ that contains $y_s$. The inner optimization problem (14) for $H_i v$ is a linear program for any fixed $(x, u)$. However, the outer problem (13) is nonconvex in general. The nonconvexity originates from the local representation of $y_s = f(x, u, \xi^{[s]})$ by only using the grid points in the cell that contains $y_s$. However, such a local representation reduces the problem size and allows us to show that $\hat{T}_i v$ converges uniformly to $T_i v$ on $Z_t$ as $\delta$ tends to zero if $v$ is locally Lipschitz continuous on $Z_{t+1}$.

Proposition 4.3 Suppose that Assumptions 1 holds, and that $v \in \mathbb{B}_b(Z_{t+1})$ is locally Lipschitz continuous on $Z_t$. Then, we have

$$
| (\hat{T}_i v)(x) - (T_i v)(x) | \leq L_v \delta \quad \forall x \in Z_t,
$$

where $L_v := \max_{j=1,\ldots,N_{C,t+1}} \sup_{x,x' \in \mathcal{C}_j, x \neq x'} \frac{\|v(x) - v(x')\|}{\|x - x'\|}$. Therefore, $\hat{T}_i v$ converges uniformly to $T_i v$ on $Z_t$ as $\delta \to 0$.

Proof Fix an arbitrary $x \in Z_t$. Under Assumption 1, both (13) and (14) have optimal solutions. Let $\bar{u}$ be an optimal solution to (13), and let $\bar{y}$ be a corresponding optimal solution to (14) given $u := \bar{u}$. We also let $\bar{y}_s := f(x, \bar{u}, \xi^{[s]})$ for each $s$. Then, we have $\bar{y}_s = \sum_{i \in \mathcal{N}(\bar{y}_s)} \bar{y}_{s,i} x^{[i]}$. By the definition of $L_v$, we have $|v(\bar{y}_s) - v(x^{[i]})| \leq L_v \|\bar{y}_s - x^{[i]}\| \leq L_v \delta$ for all $i \in \mathcal{N}(\bar{y}_s)$. Since $\sum_{i \in \mathcal{N}(\bar{y}_s)} \bar{y}_{s,i} = 1$, we have

$$
\left| v(\bar{y}_s) - \sum_{i \in \mathcal{N}(\bar{y}_s)} \bar{y}_{s,i} u(x^{[i]}) \right| \leq \sum_{i \in \mathcal{N}(\bar{y}_s)} \left| \bar{y}_{s,i} v(\bar{y}_s) - \bar{y}_{s,i} u(x^{[i]}) \right| \leq \sum_{i \in \mathcal{N}(\bar{y}_s)} \bar{y}_{s,i} L_v \delta = L_v \delta
$$
for each $s$. By using this bound, we obtain that

$$
(\tilde{T}_t)v(x) = r(x, \tilde{u}) + \sum_{s=1}^{N} \sum_{i \in N_1(s)} p_s \tilde{y}_{s,i} v(x^{[s]}) \geq r(x, \tilde{u}) + \sum_{s=1}^{N} p_s v(\tilde{y}_s) - L_v \delta
$$

$$
= r(x, \tilde{u}) + \sum_{s=1}^{N} p_s v(f(x, \tilde{u}, \tilde{\xi}[s])) - L_v \delta \geq (T_t)v(x) - L_v \delta,
$$

where the last inequality holds because $\tilde{u}$ is a feasible solution to the minimization problem in the definition (4) of $T_t$.

The other inequality, $(\tilde{T}_t)v(x) - L_v \delta \leq (T_t)v(x)$, can be shown by using the second part of the proof of Proposition 3.1. Since $x$ was arbitrarily chosen in $Z_t$, the result follows.

Let $\tilde{v}_t : Z_t \rightarrow \mathbb{R}, t = 0, \ldots, K$, be recursively defined by $\tilde{v}_t(x) = (\tilde{T}_t \tilde{v}_{t+1})(x)$ for all $x \in Z_t$ with $\tilde{v}_K = q$ on $Z_K$. By using Proposition 4.3 and the inductive argument in the proof of Theorem 3.1, we can show that $\tilde{v}_t$ converges uniformly to the optimal value function on $Z_t$ as $\delta$ tends to zero.

**Theorem 4.1** Suppose that Assumption 1 holds, and that $v^*_t$ is locally Lipschitz continuous on $Z_t$ for each $t = 0, \ldots, K$. Then, we have

$$
|v^*_t(x) - \tilde{v}_t(x)| \leq \sum_{k=t+1}^{K} L_k \delta \quad \forall x \in Z_t \forall t = 0, \ldots, K,
$$

where $L_k := \sup_{j=1, \ldots, N_{C,k}} \sup_{x, x' \in C_j : x \neq x'} \frac{\|v^*_t(x) - v^*_t(x')\|}{\|x - x'\|}$. Therefore, $\tilde{v}_t$ converges uniformly to $v^*_t$ on $Z_t$ as $\delta \rightarrow 0$.

As before, we construct a deterministic Markov policy $\tilde{\pi} := (\tilde{\pi}_0, \ldots, \tilde{\pi}_{K-1})$ by setting $\tilde{\pi}_t(x)$ to be an optimal solution to (13) with $v := \tilde{v}_{t+1}$. Then, the cost-to-go function $v^*_t : Z_t \rightarrow \mathbb{R}$ under this policy can be evaluated by solving the Bellman equation $v^*_t := T^* v^*_{t+1}$ for all $t = 0, \ldots, K - 1$, where $v^*_{K} \equiv q$ on $Z_K$. This cost $v^*_t$ incurred by the approximate policy $\tilde{\pi}$ converges uniformly to the optimal value function $v^*_t$ as the grid resolution becomes finer.

**Corollary 4.1** Suppose that Assumption 1 holds, and that $v^*_t : Z_t \rightarrow \mathbb{R}$ is locally Lipschitz continuous for all $t = 0, \ldots, K$. Then, we have

$$
0 \leq v^*_t(x) - v^*_t(x) \leq \sum_{k=t+1}^{K} 2L_k \delta \quad \forall x \in Z_t \forall t = 0, \ldots, K,
$$

where $L_k := \sup_{j=1, \ldots, N_{C,k}} \sup_{x, x' \in C_j : x \neq x'} \frac{\|v^*_t(x) - v^*_t(x')\|}{\|x - x'\|}$. Therefore, $v^*_{t}$ converges uniformly to $v^*_{t}$ on $Z_t$ as $\delta \rightarrow 0$.

---

3 Note that the convexity of $v$ is unused in the second part of the proof of Proposition 3.1. Thus, it is valid in the nonconvex case.
Proof By the optimality of $v_i^*$ under Assumption 1, we have $v_i^* \leq \tilde{v}_i^*$ on $Z_t$ for all $t$. We now show that $\tilde{v}_i^* \leq v_i^* + \sum_{k=t+2}^{K} 2L_k \delta$ on $Z_t$ by using mathematical induction. For $t = K$, $\tilde{v}_K^* = v_K^*$ equals $q$ on $Z_K$, and thus the induction hypothesis holds. Suppose that the induction hypothesis is valid for some $t + 1$, that is $0 \leq \tilde{v}_{t+1}^* (x) - v_{t+1}^* (x) \leq \sum_{k=t+2}^{K} 2L_k \delta$ for all $x \in Z_{t+1}$. Then, by the monotonicity of $T_i^\tilde{v}$, we have

$$(T_i^\tilde{v} v_{t+1}^*)(x) \leq (T_i^\tilde{v} v_{t+1}^*)(x) + \sum_{k=t+2}^{K} 2L_k \delta \quad \forall x \in Z_t. \quad (16)$$

Fix an arbitrary state $x \in Z_t$. Under Assumption 1, by the definition of $\tilde{T}_t$, there exists $\tilde{\gamma} \in \Delta^N$ such that

$$(\tilde{T}_t v_{t+1}^*)(x) = r(x, \tilde{\pi}_t(x)) + \sum_{s=1}^{N} \sum_{i \in \mathcal{N}(\tilde{s}_s)} p_{s, i} v_{t+1}^*(x^{[i]}), \quad (17)$$

where $\tilde{s}_s := f(x, \tilde{\pi}_t(x), \tilde{\xi}^{[s]})$, and $f(x, \tilde{\pi}_t(x), \tilde{\xi}^{[s]}) = \sum_{i \in \mathcal{N}(\tilde{s}_s)} \tilde{s}_s, i x^{[i]}$. Since $v_{t+1}^*$ is continuous on $Z_{t+1}$, we have

$$|v_{t+1}^*(\tilde{s}_s) - v_{t+1}^*(x^{[i]})| \leq L_{t+1} \|\tilde{s}_s - x^{[i]}\| \leq L_{t+1} \delta \quad \forall i \in \mathcal{N}(\tilde{s}_s).$$

Note that $\sum_{i \in \mathcal{N}(\tilde{s}_s)} \tilde{s}_s, i = 1$. Thus, the following inequalities hold:

$$\left| v_{t+1}^*(\tilde{s}_s) - \sum_{i \in \mathcal{N}(\tilde{s}_s)} \tilde{s}_s, i v_{t+1}^*(x^{[i]}) \right| \leq \sum_{i \in \mathcal{N}(\tilde{s}_s)} \tilde{s}_s, i |v_{t+1}^*(\tilde{s}_s) - v_{t+1}^*(x^{[i]})| \leq L_{t+1} \delta. \quad (18)$$

By (17) and (18), we have

$$\begin{align*}
(\tilde{T}_t v_{t+1}^*)(x) &\geq r(x, \tilde{\pi}_t(x)) + \sum_{s=1}^{N} p_s v_{t+1}^*(\tilde{s}_s) - L_{t+1} \delta \\
&= r(x, \tilde{\pi}_t(x)) + \sum_{s=1}^{N} p_s v_{t+1}^* (f(x, \tilde{\pi}_t(x), \tilde{\xi}^{[s]})) - L_{t+1} \delta \\
&\geq (T_i^\tilde{v} v_{t+1}^*)(x) - L_{t+1} \delta. \quad (19)
\end{align*}$$

On the other hand, by Proposition 4.3, we have

$$(\tilde{T}_t v_{t+1}^*)(x) \geq (T_i^v v_{t+1}^*)(x) + L_{t+1} \delta. \quad (20)$$
Since $\mathbf{x}$ was arbitrarily chosen in $\mathcal{Z}_t$, combining inequalities (16), (19) and (20) yields

$$v_t^* = T_t^* v_{t+1}^* \leq T_t v_{t+1}^* + \sum_{k=t+1}^{K} 2L_k \delta = v_t^* + \sum_{k=t+1}^{K} 2L_k \delta$$

on $\mathcal{Z}_t$, which implies that the induction hypothesis is valid for $t$ as desired.

From the computational perspective, it is challenging to obtain a globally optimal solution to (13) due to nonconvexity. Thus, over-approximating $\tilde{T}_t v$ is inevitable in practice, and the quality of an approximate policy $\tilde{\pi}$ from an over-approximation of $\tilde{v}_t$’s depends on the quality of locally optimal solutions to (13) evaluated in the process of value iteration. Another classical and possibly practical method for solving the minimization problem in (13) is to discretize the action space. After discretization, we evaluate the optimal value of the program (14) for each action, and then compare the optimal values to approximately solve (13). This approach is used in solving the nonconvex problem in Sect. 5.2.

On the other hand, the inner optimization problem (14) can be efficiently solved using existing linear programming algorithms such as interior point and simplex methods (e.g., [33,34]). The linear optimization problem has an equivalent form with the reduced optimization variables $\tilde{\gamma} := (\gamma_i)_{i \in \mathcal{N}_t}$ instead of using the full vector $\gamma$. Using this reduced optimization problem significantly gains computational efficiency since the size of the reduced optimization variable is independent of grid resolution. For example, in the case of regular Cartesian grid with size $N^n_x$, the number of optimization variables decreases from $N^n_x \times N$ to $2^n \times N$.

5 Numerical Experiments

5.1 Linear-Convex Control

We first demonstrate the performance of the proposed method through a linear-convex stochastic control problem. Consider the linear stochastic system $x_{t+1} = Ax_t + Bu_t + C \xi_t$, where $x_t \in \mathbb{R}^2$ and $\xi_t \in \mathbb{R}$. We set $A = \begin{bmatrix} 0.85 & 0.1 \\ 0.1 & 0.85 \end{bmatrix}$, and $C = (1, 1)$. To demonstrate that our method can handle high-dimensional action spaces, the control variable $u_t$ is chosen to be 1000 dimensional. The matrix $B$ is an 2 by 1000 matrix of all entries being sampled independently from the uniform distribution over $[0, 1]$ and then normalized so that the 1-norm of each row is 1. The stage-wise cost function is given by

$$r(x_t, u_t) = \|x_t\|_1 + \|u_t\|_2^2,$$

which is convex but non-quadratic. The set of admissible control actions is chosen as $\mathcal{U} := [-0.15, 0.15]^{1000}$. The support elements of $\xi_t$ are sampled according to

\footnote{The matrix $B$ used in our experiments can be downloaded from the following link: \url{http://coregroup.snu.ac.kr/DB/B1000.mat}.}
the uniform distribution over \([-0.1, 0.1]\). The computational domains are chosen as follows: \(Z_t := [-1 - 0.2t, 1 + 0.2t]^2\) for \(t = 0, \ldots, 5\) so that \(Z_0 = [-1, 1]^2\) and \(Z_5 = [-2, 2]^2\). This choice of computational domains satisfies \(\{Ax + Bu + C\xi : x \in Z_t, u \in [-0.15, 0.15]^2, \xi \in [-0.1, 0.1]\} \subseteq Z_{t+1},\) and \(Z_0 \subseteq Z_1 \subseteq \cdots \subseteq Z_5\). Each domain \(Z_t\) is discretized as a two-dimensional regular Cartesian grid using the approach in Appendix with grid points \((i^\frac{-\delta x}{2}, j^\frac{-\delta x}{2})\) for \(i, j = 1, \ldots, (2 + 0.4t)/\delta_x + 1\), where \(\delta_x\) is the grid spacing. The terminal cost is set to be \(q \equiv 0\). The numerical experiments were conducted on a PC with 4.2 GHz Intel Core i7 and 64GB RAM. The optimization problems were solved using the interior point method in CPLEX.

Let \(N_x\) denote the number of grid points on one axis of \(Z_5\). Thus, the total number of grid points is \(N_x^2\). To demonstrate the convergence of the proposed method as \(\delta = \sqrt{2}\delta_x \rightarrow 0\) or equivalently as \(N_x \rightarrow \infty\), we fix the number of sample data as \(N = 10\), and compute \(\hat{v}_{(N_x),t}\) for \(N_x = 21, 41, 81, 161\) or equivalently for \(\delta_x = 0.2, 0.1, 0.05, 0.025\), where \(\hat{v}_{(N_x),t}\) denotes the approximate value function at \(t\) with the number of grid points being \(N_x^2\). Setting \(\tilde{v}_0 := \hat{v}_{(N_x=321),0}\), the errors \(\|\hat{v}_{(N_x),0} - \tilde{v}_0\|\) over \(Z_0 = [-1, 1]^2\), measured by the 1- and \(\infty\)-norm, are shown in Fig. 2. The CPU time for computation is also reported in Table 1.\(^5\) In this example, the empirical convergence rate of our method is beyond the second order.\(^6\) Furthermore, the relative error \(\|\hat{v}_{(N_x),0} - \tilde{v}_0\|/\tilde{v}_0\|\) over \(Z_0 = [-1, 1]^2\) is 0.02% in the case of \(N_x = 161\).

\(^5\) The CPU time increases superlinearly with the number of grid points. This is because the size of the optimization problem (5) also increases with the grid size. Note that the problem size is invariant when using the bi-level method in Sect. 4.2. Thus, in that case the CPU time scales linearly as shown in Table 3.

\(^6\) The observation of the second-order empirical convergence rate is consistent with our theoretical result since Theorem 3.1 only suggests that the suboptimality gap decreases with the first-order rate. Thus, the actual convergence rate can be higher than the convergence rate for the suboptimality gap.
Table 1 Computation time (in seconds) and errors for the linear $L^1$-control problem with different grid sizes (evaluated over $Z_0 = [-1, 1]^2$)

| # of grid points | 21$^2$ | 41$^2$ | 81$^2$ | 161$^2$ |
|------------------|--------|--------|--------|--------|
| Time (s)         | 18.22  | 139.93 | 1877.25| 43736.92|
| $\ell_1$-error   | 0.0339 | 0.0090 | 0.0020 | 0.0004 |
| $\ell_\infty$-error | 0.0527 | 0.0157 | 0.0036 | 0.0009 |

Figure 3 shows the errors in $1$- and $\infty$-norm, evaluated over $Z_0 = [-1, 1]^2$ depending on the number of DP iterations or equivalently the length of time horizon when $\delta x = 0.2$. The errors grow sublinearly with respect to the number of DP iterations. This result is consistent with the error bound in Theorem 3.1, assuming that the constant $L_k$ does not change much over time.

To test the empirical convergence of the proposed method as the sample size increases, we also compute $\hat{v}(N), 0$ for $N = 20, 40, 80, 160$ with $\delta x$ fixed as 0.2, where $\hat{v}(N), 0$ denotes the approximate value function at $t = 0$ with sample size $N$. The errors $\|\hat{v}(N), 0 - \tilde{v}_0\|$ over $Z_0 = [-1, 1]^2$ with $N = 20, 40, 80, 160$ are shown in Fig. 4, where $\tilde{v}_0 := \hat{v}(N=320), 0$. The CPU time is reported in Table 2. In this example, the empirical convergence rate is below the first order. The relative error $\|(\hat{v}(N), 0 - \tilde{v}_0)/\tilde{v}_0\|_1$ is 0.72% in the case of $N = 160$.

5.2 Control-Affine Nonlinear System

We consider the following nonlinear, control-affine system:

$$x_{t+1} = x_t + [w_t(1 - x_t)x_t - x_t u_t] \Delta t,$$

which models the outbreak of an infectious disease [35]. Here, the system state $x_t \in [0, 1]$ represents the ratio of infected people in a given population, the control input $u_t \in [0, u_{\text{max}}]$ represents a public treatment action, and $w_t > 0$ denotes infectivity.
To examine the suboptimality of our method in this problem with a control-affine system, we compare the value function $v^\pi_t$ of our approximate policy and the optimal value function $v^*_t$. As shown in Fig. 5a, b, $v^\pi_t$ and $v^*_t$ look very similar to each other. Figure 5c shows the relative error $\frac{v^\pi_t(x) - v^*_t(x)}{v^*_t(x)} \times 100\%$. The relative error is less than 11\% for all $x \in [0, 1]$ and $t = 0, \ldots, 20$. The $\ell_1$-norm and the $\ell_2$-norm of the relative error are 0.45\% and 1.44\%, respectively. This result confirms the utility of our method even in nonconvex problems with control-affine systems.

7 To compute the optimal value function, we used the method in Sect. 4.2 discretizing the action space with 1001 equally spacing grid points.
5.3 Fully Nonlinear System

Consider the following Dubins car model [36], which is fully nonlinear:

\[
\begin{align*}
    x_{t+1} &= x_t + v_t \cos \theta_t, \\
    y_{t+1} &= y_t + v_t \sin \theta_t, \\
    \theta_{t+1} &= \theta_t + \frac{1}{L} v_t \tan s_t,
\end{align*}
\]

where \( x_t := (x_t, y_t, \theta_t) \in \mathbb{R}^3 \) is the system state, and \( u_t := (v_t, s_t) \in \mathbb{R}^2 \) is the control action. Specifically, \((x_t, y_t)\) represents the position of the vehicle in \( \mathbb{R}^2 \), and \( \theta_t \) denotes the heading angle. Moreover, \( v_t \) and \( s_t \) denote a velocity input and a steering input, respectively. The third state equation specifies the turning rate. The Dubins vehicle model is commonly used in robotics and controls as a model for describing the planar motion of wheeled vehicles. We set \( v_t \in \{0, -0.1\}, s_t \in \{-1, 0, 1\}, \) and \( L = 0.5 \). The control objective is to steer the vehicle to the y-axis with heading angle \( \pi \). Accordingly, the terminal cost function is set to be \( q(x) = \| (x_2, x_3) - (0, \pi) \|^2 \), where \( x = (x_1, x_2, x_3) \in \mathbb{R}^3 \) with no running costs. The computational domains are chosen as

\[
\mathcal{Z}_t := [-0.5 - 0.1t, 0.5 + 0.1t] \times [-0.5 - 0.1t, 0.5 + 0.1t] \times [0, 2\pi]
\]

for \( t = 0, 1, \ldots, 20 \), where \( K = 20 \). It is straightforward to check that this choice satisfies the condition (3). Each \( \mathcal{Z}_t \) is discretized as a three-dimensional Cartesian grid using the approach in Appendix with grid points

\[
\left\{ (-0.5 - 0.1t + \delta_x(i - 1), -0.5 - 0.1t + \delta_x(j - 1), 0 + \delta_\theta(k - 1)) : i, j = 1, \ldots, (1 + 0.2t)/\delta_x + 1, k = 1, \ldots, 2\pi/\delta_\theta + 1 \right\}.
\]

In our numerical experiment, we set \( \delta_x = 0.1 \) and \( \delta_\theta = \pi/20 \), and thus \( \mathcal{Z}_{20} \) is discretized with a grid of size \( 51 \times 51 \times 41 \). The CPU time for running our method in Sect. 4.2 is 4630.93 seconds. Figure 6 shows the resulting vehicle trajectories starting from \((x_0, y_0) = (0, -0.5)\) with four different initial heading angles \( \theta_0 = \ldots \)
In all four cases, the vehicle moves to the y-axis with heading $\pi$ as desired.

As in Sect. 5.1, let $N_x$ denote the number of grid points on one axis of $Z_5$. Thus, the total number of grid points is $N^3_x$. To demonstrate the convergence of our method as $\delta_x, \delta_0 \to 0$ or equivalently as $N_x \to \infty$, we evaluate the errors $\|\hat{v}(N_x),0 - \bar{v}_0\|$ in the 1- and $\infty$-norm over $Z_0$, where $\hat{v}(N_x),t$ denotes the approximate value function at $t$ with the number of grid points being $N^3_x$, and $\bar{v}_0 := \hat{v}(N_x=161),0$. Table 3 shows the errors and the CPU time with different grid sizes and $K = 5$. This result indicates that the empirical convergence rate in this example is approximately the second order. Another important observation is that the computation time scales linearly with the number of grid points. This is because the size of the linear optimization problem (14) is invariant with the grid size as discussed in Sect. 4.2.

**Acknowledgements**  This work was supported in part by the Creative-Pioneering Researchers Program through SNU, the National Research Foundation of Korea funded by the MSIT (2020R1C1C1009766), and Samsung Electronics.
Appendix: State Space Discretization Using a Rectilinear Grid

In this appendix, we provide a concrete way to discretize the state space using a rectilinear grid. The construction below satisfies all the conditions in Sect. 2.3.

1. Choose a convex compact set $Z_0 := [x_{0,1}, x_{0,1}] \times [x_{0,2}, x_{0,2}] \times \cdots \times [x_{0,n}, x_{0,n}]$, and discretize it using an $n$-dimensional rectilinear grid. Set $t \leftarrow 0$.

2. Compute (or over-approximate) the forward reachable set $R_t := \{ f(x, u, \xi) : x \in Z_t, u \in \mathcal{U}(x), \xi \in \mathcal{E} \}$.

3. Choose a convex compact set $Z_{t+1} := [x_{t+1,1}, x_{t+1,1}] \times [x_{t+1,2}, x_{t+1,2}] \times \cdots \times [x_{t+1,n}, x_{t+1,n}]$ such that $R_t \subseteq Z_{t+1}$.

4. Expand the rectilinear grid to fit $Z_{t+1}$.

5. Stop if $t + 1 = K$; otherwise, set $t \leftarrow t + 1$ and go to Step 2.

We can then choose $C_i$ as each grid cell. We label $C_i$ so that $\bigcup_{i=1}^{N_C} C_i = Z_t$ for all $t$. A two-dimensional example is shown in Fig. 1. This construction approach was used in Sects. 5.1 and 5.3.

References

1. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, Hoboken (2014)
2. Kushner, H., Dupuis, P.G.: Numerical Methods for Stochastic Control Problems in Continuous Time. Springer, New York (2013)
3. Hernández-Lerma, O., Lasserre, J.B.: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York (2012)
4. Savorgnan, C., Lasserre, J.B., Diehl, M.: Discrete-time stochastic optimal control via occupation measures and moment relaxations. In: Proceedings of Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference (2009)
5. Dufour, F., Prieto-Rumeau, T.: Finite linear programming approximations of constrained discounted Markov decision processes. SIAM J. Control Optim. 51(2), 1298–1324 (2013)
6. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, 2nd edn. MIT Press, Cambridge (2018)
7. Bertsekas, D.P.: Reinforcement Learning and Optimal Control. Athena Scientific, Belmont (2019)
8. Szepesvari, C.: Algorithms for Reinforcement Learning. Morgan and Claypool Publishers, San Rafael (2010)
9. Bertsekas, D.P.: Convergence of discretization procedures in dynamic programming. IEEE Trans. Autom. Control 20(3), 415–419 (1975)
10. Langen, H.-J.: Convergence of dynamic programming models. Math. Oper. Res. 6(4), 493–512 (1981)
11. Whitt, W.: Approximations of dynamic programs, I. Math. Oper. Res. 3(3), 231–243 (1978)
12. Hinderer, K.: On approximate solutions of finite-stage dynamic programs. In: Puterman, M.L. (ed.) Dynamic Programming and Its Applications, pp. 289–317. Academic Press, New York (1978)
13. Chow, C.-S., Tsitsiklis, J.N.: An optimal one-way multigrid algorithm for discrete-time stochastic control. IEEE Trans. Autom. Control 36(8), 898–914 (1991)
14. Dufour, F., Prieto-Rumeau, T.: Approximation of Markov decision processes with general state space. J. Math. Anal. Appl. 388, 1254–1267 (2012)

---

8 The forward reachable set can be over-approximated in an analytical way, particularly when a loose approximation is allowed. For a high quality of approximation, one may use advanced computational techniques with semidefinite approximation [37] and ellipsoidal approximation [38], among others.
15. Dufour, F., Prieto-Rumeau, T.: Approximation of average cost Markov decision processes using empirical distributions and concentration inequalities. Stochast. Int. J. Probab. Stochast. Process. 87(2), 273–307 (2015)
16. Saldi, N., Yüksel, S., Linder, T.: On the asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. Math. Oper. Res. 42(4), 945–978 (2017)
17. Hernández-Lerma, O.: Discretization procedures for adaptive Markov control processes. J. Math. Anal. Appl. 137, 485–514 (1989)
18. Johnson, S.A., Stedinger, J.R., Shoemaker, C.A., Li, Y., Tejada-Guibert, J.A.: Numerical solution of continuous-state dynamic programs using linear and spline interpolation. Oper. Res. 41(3), 484–500 (1993)
19. Hernández-Lerma, O., Lasserre, J.B.: Further Topics on Discrete-Time Markov Control Processes. Springer, New York (2012)
20. Rust, J.: Using randomization to break the curse of dimensionality. Econometrica 65(3), 487–516 (1997)
21. Munos, R., Szepesvári, C.: Finite-time bounds for fitted value iteration. J. Mach. Learn. Res. 1, 815–857 (2008)
22. Haskell, W.B., Jain, R., Sharma, H., Yu, P.: A universal empirical dynamic programming algorithm for continuous state MDPs. IEEE Trans. Autom. Control 65(1), 115–129 (2020)
23. Jang, S., Yang, I.: Stochastic subgradient methods for dynamic programming in continuous state and action spaces. In: Proceedings of the 58th IEEE Conference on Decision and Control, pp. 7287–7293 (2019)
24. Nesterov, Y.: Lectures on Convex Optimization, 2nd edn. Springer, Cham (2018)
25. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
26. Bertsekas, D.P.: Convex Optimization Algorithms. Athena Scientific, Belmont (2015)
27. Falcone, M., Ferretti, R.: Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations. SIAM, Philadelphia (2013)
28. Alla, A., Falcone, M., Saluzzi, L.: An efficient DP algorithm on a tree-structure for finite horizon optimal control problems. SIAM J. Sci. Comput. 41(4), A2384–A2406 (2019)
29. Picarelli, A., Reisinger, C.: Probabilistic error analysis for some approximation schemes to optimal control problems. Syst. Control Lett. 137, 104619 (2020)
30. Yang, I.: A convex optimization approach to dynamic programming in continuous state and action spaces. arXiv preprint arXiv:1810.03847 (2018)
31. Abate, A., Prandini, M., Lygeros, J., Sastry, S.: Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems. Automatica 44, 2724–2734 (2008)
32. Yang, I.: A dynamic game approach to distributionally robust safety specifications for stochastic systems. Automatica 94, 94–101 (2018)
33. Dantzig, G.B.: Linear Programming and Extensions. Princeton University Press, Princeton (1998)
34. Bertsimas, D., Tsitsiklis, J.N.: Introduction to Linear Optimization. Athena Scientific, Belmont (1997)
35. Sethi, S.P., Thompson, G.L.: Optimal Control Theory: Applications to Management Science and Economics. Springer, New York (2000)
36. Dubins, L.E.: On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. Am. J. Math. 79(3), 497–516 (1957)
37. Magron, V., Garroch, P.-L., Henrion, D., Thirion, X.: Semidefinite approximations of reachable sets for discrete-time polynomial systems. SIAM J. Control Optim. 57(4), 2799–2820 (2019)
38. Kurzhanskiy, A.A., Varaiya, P.: Reach set computation and control synthesis for discrete-time dynamical systems with disturbances. Automatica 47, 1414–1426 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.