Optimal bounds for numerical approximations of infinite horizon problems based on dynamic programming approach

Javier de Frutos* Julia Novo†

September 30, 2022

Abstract

In this paper we get error bounds for fully discrete approximations of infinite horizon problems via the dynamic programming approach. It is well known that considering a time discretization with a positive step size h an error bound of size h can be proved for the difference between the value function (viscosity solution of the Hamilton-Jacobi-Bellman equation corresponding to the infinite horizon) and the value function of the discrete time problem. However, including also a spatial discretization based on elements of size k an error bound of size $O(k/h)$ can be found in the literature for the error between the value functions of the continuous problem and the fully discrete problem. In this paper we revise the error bound of the fully discrete method and prove, under similar assumptions to those of the time discrete case, that the error of the fully discrete case is in fact $O(h+k)$ which gives first order in time and space for the method. This error bound matches the numerical experiments of many papers in the literature in which the behaviour $1/h$ from the bound $O(k/h)$ have not been observed.

Key words. Dynamic programming, Hamilton-Jacobi-Bellman equation, optimal control, error analysis.

1 Introduction

The numerical approximation of optimal control problems is of importance for many applications such as aerospace engineering, chemical processing and resource economics, among others. In this paper, we consider the dynamic programming approach to the solution of optimal control problems driven by dynamical systems in $\mathbb{R}^n$. We refer to [5] for a monograph on this subject.

*Instituto de Investigación en Matemáticas (IMUVA), Universidad de Valladolid, Spain. Research supported by Spanish MINECO under grant PID2019-104141GB-I00 and by Junta de Castilla y León under grant VA169P20 co-finanzed by FEDER (EU) funds (frutos@mac.uva.es)

†Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain. Research supported by Spanish MINECO under grant PID2019-104141GB-I00 and by Junta de Castilla y León under grant VA169P20 co-finanzed by FEDER (EU) funds (julia.novo@uam.es)

The value function of an optimal control problem is known to be usually only Lipschitz continuous even when the data is regular. The characterization of the value function is obtained in terms of a first-order nonlinear Hamilton-Jacobi-Bellman (HJB) partial differential equation. A bottleneck in the computation of the value function comes from the need of approaching a nonlinear partial differential equation in dimension n, which is a challenging problem in high dimensions. Several approximation schemes have been proposed in the literature, ranging from finite differences to semi-Lagrangian and finite volume methods, see e.g. [8], [1], [14], [7]. Some of these algorithms converge to the value function but their convergence is slow. The curse of dimensionality is mitigated in [4], [2] by means of a reduced-order model based on proper orthogonal decomposition. A new accelerated algorithm which can produce an accurate approximation of the value function in a reduced amount of time in comparison to other available methods is introduced in [3].

In the present paper, our concern is about the error bounds available in the literature for the fully discrete semi-lagrangian method approaching the value function, the viscosity solution of the HJB equation corresponding to the infinite horizon. For a method with a positive time step size h and spatial elements of size k an error bound of size $O(k/h)$ can be found in [9, Corollary 2.4], [11, Theorem 1.3]. However, the behaviour $1/h$ in the error bound of the fully discrete method has never been observed in the numerical experiments, see for example [4]. Based on this fact, we reconsider the error analysis of the fully discrete method.

In this paper we prove a bound of size $O(h + k)$ which gives first order in time and space for the method. This rate of convergence is the same appearing in [9, Corollary 2.4]. However, as stated in [15], [10], the proof of this corollary was based on an identification which does not hold in the example shown in [15]. The idea of the present paper is to imitate the analysis of the discrete-time method for which the value function is characterized as the minimum of a functional, see [5, Proposition 4.1, Chapter VI (appendix)]. To this end, we define a fully discrete cost functional that differs from that of the discrete-time approximation in the use of spatial interpolator operators. Then, in Theorem 3 we prove that the fully discrete approximation can also be characterized as the minimum of this fully discrete cost functional. In the proof we use that the fully discrete approximation defined by means of a discrete dynamic programming principle is unique, as it is proved in [5, Theorem 1.1, Appendix A].

Finally, thanks to this new characterization we can extend the ideas from [5, Lemma 1.2, Chapter VI] (for the semi-discrete case) to the fully discrete case and we are able to prove first order of convergence for the method both is space and time. First order of convergence in time is linked to the assumption of Lipschitz continuity of the controls in the intervals defined excluding a finite number of jump discontinuities. In case we have less regularity we obtain weaker error bounds. For example, for uniformly continuous controls, allowing again a finite number of discontinuities, the error goes to zero as h goes to zero but the first order of convergence is not achieved. Intermediate rates of convergence or order α in time, for $0 < \alpha < 1$ are equally proved for α-Hölder continuous controls.

Following the arguments in [6] we can also prove convergence arguing with piecewise constants controls and under weaker regularity assumptions (only some convexity assumptions are needed but no extra regularity assumptions for the controls).

However, adapting the arguments in [6] written for finite horizon problems to our infinite horizon case we loose the full first order in time. We develop this argument at the end of the paper.

We think that the new characterization we introduce in this paper, based on optimality arguments, could have potential to be used in other types of Hamilton-Jacobi-Bellman equations where the convergence rates are still suboptimal.

To conclude this section we want to mention some other related references. The paper [13] contains a first suboptimal convergence of rate $\log(h)h^{1/2}$ for a similar scheme. The monograph of Falcone and Ferreti [12] contains a few of the difference convergence rates for HJB partial differential equations in control and games.

The outline of the paper is as follows. In section 2 we state the model problem and some preliminary results. In Section 3 we prove several error bounds for the method under different regularity requirements. More precisely, in Subsection 3.1 we assume some regularity assumptions for the controls and show how the first order in time and space can be achieved. In Subsection 3.2, we follow the arguments in [6] to prove convergence under weaker regularity assumptions.

2 Model problem and Preliminary results

Along this section we follow the notation in [4]. For a nonlinear mapping

$$f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n,$$

and a given initial condition $y_0 \in \mathbb{R}^n$ let us consider the controlled nonlinear dynamical system

$$\dot{y}(t) = f(y(t), u(t)) \in \mathbb{R}^n, \quad t > 0, \quad y(0) = y_0 \in \mathbb{R}^n, \tag{1}$$

together with the infinite horizon cost functional

$$J(y, u) = \int_0^\infty g(y(t), u(t))e^{-\lambda t}\, dt. \tag{2}$$

In (2) $\lambda > 0$ is a given weighting parameter and

$$g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}.$$

The set of admissible controls is

$$\mathbb{U}_{\text{ad}} = \{u \in \mathbb{U} \mid u(t) \in U_{\text{ad}} \text{ for almost all } t \geq 0\},$$

where $\mathbb{U} = L^2(0, \infty; \mathbb{R}^m)$ and $U_{\text{ad}} \subset \mathbb{R}^m$ is a compact convex subset.

As in [4, Assumption 2.1] we assume the following hypotheses:

- The right-hand side f in (1) is continuous and globally Lipschitz-continuous in both the first and second arguments; i.e., there exists a constant $L_f > 0$ satisfying

$$\begin{aligned} \|f(y, u) - f(\tilde{y}, u)\|_2 &\leq L_f\|y - \tilde{y}\|_2, \quad \forall y, \tilde{y} \in \mathbb{R}^n, u \in U_{\text{ad}} && (3)\\ \|f(y, u) - f(y, \tilde{u})\|_2 &\leq L_f\|u - \tilde{u}\|_2, \quad \forall u, \tilde{u} \in U_{\text{ad}}, y \in \mathbb{R}^n. && (4) \end{aligned}$$

- The right-hand side f in (1) satisfies that there exists a constant $M_f > 0$ such that the following bound holds

$$\|f(y,u)\|_\infty = \max_{1\le i\le n} |f_i(y,u)| \le M_f, \quad \forall y \in \overline{\Omega} \subset \mathbb{R}^n, u \in U_{\rm ad}, \tag{5}$$

  where $\overline{\Omega}$ is a bounded polyhedron such that for sufficiently small $h > 0$ the following inward pointing condition on the dynamics holds

$$y + hf(y,u) \in \overline{\Omega}, \quad \forall y \in \overline{\Omega}, u \in U_{\rm ad}. \tag{6}$$

- The running cost g is continuous and globally Lipschitz-continuous in both the first and second arguments; i.e., there exists a constant $L_g > 0$ satisfying

$$|g(y,u) - g(\tilde{y},u)| \le L_g\|y-\tilde{y}\|_2, \quad \forall y,\tilde{y} \in \mathbb{R}^n, u \in U_{\rm ad} \tag{7}$$
$$|g(y,u) - g(y,\tilde{u})| \le L_g\|u-\tilde{u}\|_2, \quad \forall u,\tilde{u} \in U_{\rm ad}, y \in \mathbb{R}^n. \tag{8}$$

- Moreover, there exists a constant $M_g > 0$ such that

$$|g(y,u)| \le M_g, \quad \forall (y,u) \in \overline{\Omega} \times U_{\rm ad}. \tag{9}$$

From the assumptions made on f there exists a unique solution of (1) $y = y(y_0,u)$ defined on $[0,\infty)$ for every admissible control $u \in \mathbb{U}_{\rm ad}$ and for every initial condition $y_0 \in \mathbb{R}^n$, see [5, Chapter 3]. We define the reduced cost functional as follows:

$$\hat{J}(y_0,u) = J(y(y_0,u),u), \quad \forall u \in \mathbb{U}_{\rm ad}, \quad y_0 \in \mathbb{R}^n, \tag{10}$$

where $y(y_0,u)$ solves (1). Then, the optimal control can be formulated as follows: for given $y_0 \in \mathbb{R}^n$ we consider

$$\min_{u\in\mathbb{U}_{\rm ad}} \hat{J}(y_0,u).$$

The value function of the problem is defined as $v : \mathbb{R}^n \to \mathbb{R}$ as follows:

$$v(y) = \inf\left\{\hat{J}(y,u) \mid u \in \mathbb{U}_{\rm ad}\right\}, \quad y \in \mathbb{R}^n. \tag{11}$$

This function gives the best value for every initial condition, given the set of admissible controls $U_{\rm ad}$. It is characterized as the viscosity solution of the HJB equation corresponding to the infinite horizon optimal control problem:

$$\lambda v(y) + \sup_{u\in U_{\rm ad}} \{-f(y,u)\cdot\nabla v(y) - g(y,u)\} = 0, \quad y \in \mathbb{R}^n. \tag{12}$$

The solution of (12) is unique for sufficiently large λ, $\lambda > \max(L_g, L_f)$, [5]. To construct the approximation scheme, as in [9], let us consider first a time discretization where h is a strictly positive step size. We consider the following semidiscrete scheme for (12):

$$v_h(y) = \min_{u\in U_{\rm ad}} \{(1-\lambda h)v_h(y+hf(y,u)) + hg(y,u)\}, \quad y \in \mathbb{R}^n. \tag{13}$$

Under the assumptions (3), (5), (7) and (9) the function v_h is Lipschitz-continuous and satisfies (see [11, p. 473])

$$|v_h(y) - v_h(\tilde{y})| \le \frac{L_g}{\lambda - L_f}\|y-\tilde{y}\|_2, \quad \forall y,\tilde{y} \in \overline{\Omega},\ h \in [0, 1/\lambda).$$

The following convergence result for the semidiscrete approximation [9, Theorem 2.3] requires that for $(y, \tilde{y}, u) \in \mathbb{R}^n \times \mathbb{R}^n \times U_{\text{ad}}$

$$\begin{aligned} \|f(y+\tilde{y},u) - 2f(y,u) + f(y-\tilde{y},u)\|_2 &\leq C_f \|\tilde{y}\|_2^2, \quad (14) \\ \|g(y+\tilde{y},u) - 2g(y,u) + g(y-\tilde{y},u)\|_2 &\leq C_g \|\tilde{y}\|_2^2. \quad (15) \end{aligned}$$

Theorem 1 *Let assumptions* (3), (5), (6), (7), (9), (14) *and* (15) *hold and let* $\lambda > \max(2L_g, L_f)$. *Let* v *and* v_h *be the solutions of* (12) *and* (13), *respectively. Then, there exists a constant* $C \geq 0$, *that can be bounded explicitly, such that the following bound holds*

$$\sup_{y \in \mathbb{R}^n} |v(y) - v_h(y)| \leq Ch, \quad h \in [0, 1/\lambda). \quad (16)$$

Following [9], [11] we introduce a fully discrete approximation to (12). Let Ω a bounded polyhedron in $\mathbb{R}^n$ satisfying (6). Let $\{S_j\}_{j=1}^{m_s}$ be a family of simplices which defines a regular triangulation of Ω

$$\overline{\Omega} = \bigcup_{j=1}^{m_s} S_j, \quad k = \max_{1 \leq j \leq m_s} (\text{diam } S_j).$$

We assume we have n_s vertices/nodes $y^1, \ldots, y^{n_s}$ in the triangulation. Let V^k be the space of piecewise affine functions from $\overline{\Omega}$ to $\mathbb{R}$ which are continuous in $\overline{\Omega}$ having constant gradients in the interior of any simplex S_j of the triangulation. Then, a fully discrete scheme for the HJB equations is given by

$$v_{h,k}(y^i) = \min_{u \in U_{\text{ad}}} \left\{ (1 - \lambda h) v_{h,k}(y^i + hf(y^i, u)) + hg(y^i, u) \right\}, \quad (17)$$

for any vertex $y^i \in \overline{\Omega}$. Clearly, a solution to (13) satisfies (17). There exists a unique solution of (17) in the space V^k, see [5, Theorem 1.1, Appendix A]. The following result can be found in [9, Corollary 2.4], [11, Theorem 1.3], it also requires the semiconcavity assumptions (14) and (15).

Theorem 2 *Let assumptions* (3), (5), (6), (7), (9), (14) *and* (15) *hold. Let* v, v_h *and* $v_{h,k}$ *be the solutions of* (12), (13) *and* (17), *respectively. For* $\lambda > L_f$ *the following bound holds*

$$\|v_h - v_{h,k}\|_{C(\overline{\Omega})} \leq \frac{L_g}{\lambda(\lambda - L_f)} \frac{k}{h}, \quad h \in [0, 1/\lambda).$$

For $\lambda > \max(2L_g, L_f)$ *the following bound holds*

$$\|v - v_{h,k}\|_{C(\overline{\Omega})} \leq Ch + \frac{L_g}{\lambda(\lambda - L_f)} \frac{k}{h}, \quad h \in [0, 1/\lambda),$$

where C *is the constant in* (16).

As we can observe in the theorem the error bound of the fully discrete method deteriorates when the time step h tends to zero. However, this behaviour of the method has not been observed in the literature. In next section, we improve the bound of Theorem 2 proving that the error behaves as $O(h + k)$ which gives first order both in space and time, as expected, for a method based on a first order discretization in time (Euler method) and a piecewise linear approximation in space.

3 Optimal error bounds for the fully discrete approximations

The key point to improve the above error bounds is to use a new characterization for the function $v_{h,k}$. The characterization is based on the analogous characterization for the semi-discrete approximation v_h that can be found in [5, Proposition 4.1, Chapter VI (appendix)]. Let us define the space $\mathcal{U}$ of all sequences $\boldsymbol{u} = \{u_0, u_1, \dots, \}$ such that $u_j \in U_{\rm ad}$ and $\mathcal{SU} = \{\boldsymbol{u}^s\} = \left\{\left\{\boldsymbol{u}^i\right\}_{i=1}^{n_S}\right\}$, with $\boldsymbol{u}_i \in \mathcal{U}$.

Let us define the following fully discrete cost functional. For any node y^i we define the value of the functional at the node as follows

$$\hat{J}_{h,k}(y^i, \boldsymbol{u}^s) = \hat{J}_{h,k}(y^i, \boldsymbol{u}^i) := h\sum_{n=0}^{\infty} \delta_h^n I_k g(\hat{y}_n^i, u_n^i), \quad \delta_h = (1-\lambda h), \tag{18}$$

$$\hat{y}_{n+1}^i = \hat{y}_n^i + hI_k f(\hat{y}_n^i, u_n^i), \quad \hat{y}_0 = y^i, \tag{19}$$

where $I_k g(\hat{y}_n^i, u_n^i)$, respectively $I_k f(\hat{y}_n^i, u_n^i)$, is the interpolant taking the values $\left\{g(y^i, u_n^i)\right\}_{i=1}^{n_S}$, respectively $\left\{f(y^i, u_n^i)\right\}_{i=1}^{n_S}$, evaluated at $\hat{y}_n^i$. Now, we define $w_{h,k} \in V^k$ by

$$w_{h,k}(y) = \inf_{\boldsymbol{u}^s \in \mathcal{SU}} \hat{J}_{h,k}(y, \boldsymbol{u}^s), \tag{20}$$

where for $y = \sum_{j\in J_y} \mu_j y^j$, with $\mu_j > 0$ and $\sum_{j\in J_y} \mu_j = 1$ then

$$w_{h,k}(y) = \inf_{\boldsymbol{u}^s \in \mathcal{SU}} \sum_{j\in J_y} \mu_j \hat{J}_{h,k}(y^j, \boldsymbol{u}^s) = \sum_{j\in J_y} \mu_j \inf_{\boldsymbol{u}^j \in \mathcal{U}} J_{h,k}(y^j, \boldsymbol{u}^j).$$

Theorem 3 *The function $w_{h,k} \in V_k$ defined by* (20) *satisfies equation* (17) *for $i = 1, \dots, n_S$ which implies $w_{h,k} = v_{h,k}$ is the unique solution of the fully discrete problem.*

Proof The proof follows the argument of the proof of [5, Proposition 4.1, Chapter VI (appendix)]. Let us take $\boldsymbol{u}^s \in \mathcal{SU}$ and let us define $\overline{\boldsymbol{u}}^s = \left\{\left\{\overline{\boldsymbol{u}}^i\right\}_{i=1}^{n_S}\right\}$ where $\overline{\boldsymbol{u}}^i = \left\{u_1^i, u_2^i, \dots, \right\}$. We first observe that

$$\begin{aligned}
\hat{J}_{h,k}(y^i, \boldsymbol{u}^s) = \hat{J}_{h,k}(y^i, \boldsymbol{u}^i) &= hg(y^i, u_0^i) + h\sum_{n=1}^{\infty} \delta_h^n I_k g(\hat{y}_n^i, u_n^i) \\
&= hg(y^i, u_0^i) + \delta_h \sum_{n=0}^{\infty} \delta_h^n I_k g(\hat{y}_{n+1}^i, u_{n+1}^i) \\
&= hg(y^i, u_0^i) + \delta_h \hat{J}_{h,k}(y^i + hf(y^i, u_0^i), \overline{\boldsymbol{u}}^s).
\end{aligned}$$

Let us write $y = y^i + hf(y^i, u_0) = \sum_{j\in J_y} \mu_j y_j$, where $\sum_{j\in J_y} \mu_j = 1$ and $0 \le \mu_j \le 1$. Now, by definition of $J_{h,k}$ and $w_{h,k}$

$$\begin{aligned}
\hat{J}_{h,k}(y^i + hf(y^i, u_0^i), \overline{\boldsymbol{u}}^s) &= \sum_{j\in J_y} \mu_j J_{h,k}(y^j, \overline{\boldsymbol{u}}^j) \ge \sum_{j\in J_y} \mu_j \inf_{\boldsymbol{u}^j \in \mathcal{U}} J_{h,k}(y^j, \boldsymbol{u}^j) \\
&= w_{h,k}(y^i + hf(y^i, u_0^i)).
\end{aligned}$$

And then

$$\hat{J}_{h,k}(y^i, \boldsymbol{u}^s) \geq hg(y^i, u_0^i) + \delta_h w_{h,k}(y^i + hf(y^i, u_0^i)).$$

So that,

$$w_{h,k}(y^i) = \inf_{\boldsymbol{u}^i \in \mathcal{U}} \hat{J}_{h,k}(y^i, \boldsymbol{u}^i) \geq \inf_{\boldsymbol{u}^i \in \mathcal{U}} \left\{ hg(y^i, u_0^i) + \delta_h w_{h,k}(y^i + hf(y^i, u_0^i)) \right\}.$$

Finally, since the right-hand side above depends only on u_0^i

$$\inf_{\boldsymbol{u}^i \in \mathcal{U}} \left\{ hg(y^i, u_0^i) + w_{h,k}(y^i + hf(y^i, u_0^i)) \right\} = \inf_{u \in U_{\text{ad}}} \left\{ hg(y^i, u) + \delta_h w_{h,k}(y^i + hf(y^i, u)) \right\} \tag{21}$$

and then

$$w_{h,k}(y^i) \geq \inf_{u \in U_{\text{ad}}} \left\{ hg(y^i, u) + \delta_h w_{h,k}(y^i + hf(y^i, u)) \right\}.$$

Now we take any $u_0 \in U_{\text{ad}}$ and denote by $z = y^i + hf(y^i, u_0)$. With the same notation as before $z = \sum_{j \in J_z} \mu_j y^j$ with μ_j are the barycentric coordinates of z so that

$$w_{h,k}(z) = \sum_{j \in J_z} \mu_j w_{h,k}(y^j) = \sum_{j \in J_z} \mu_j \inf_{\boldsymbol{u}^j \in \mathcal{U}} \hat{J}_{h,k}(y^j, \boldsymbol{u}^j).$$

Now, we observe that for any $\epsilon > 0$ there exists $\boldsymbol{u}^\epsilon \in \mathcal{SU}$ such that

$$w_{h,k}(z) + \epsilon \geq \hat{J}_{h,k}(z, \boldsymbol{u}^\epsilon).$$

If this were not the case then for all $\boldsymbol{u}^s \in \mathcal{SU}$ and L the cardinal of J_z

$$w_{h,k}(z) + \epsilon = \sum_{j \in J_z} \mu_j \left(\inf_{\boldsymbol{u}^j \in \mathcal{U}} \hat{J}_{h,k}(y^j, \boldsymbol{u}^j) + \epsilon / L \right) < \sum_{j \in J_z} \mu_j J_{h,k}(y^j, \boldsymbol{u}^s).$$

If $L = 1$ we have obtained a contradiction, If $L > 1$ let us fix any $j_0 \in J_z$ and let us choose $\boldsymbol{u}^s \in \mathcal{SU} = \left\{ \boldsymbol{u}_m^1, \dots \boldsymbol{u}_m^{j_0 - 1}, \boldsymbol{u}, \boldsymbol{u}_m^{j_0 + 1}, \dots, \boldsymbol{u}_m^{n_s} \right\}$ where $\boldsymbol{u}_m^i$ is the argument giving the minimum in $\inf_{\boldsymbol{u} \in \mathcal{U}} \hat{J}_{h,k}(y^i, \boldsymbol{u})$. Then we get

$$\inf_{\boldsymbol{u}^{j_0} \in \mathcal{U}} \hat{J}_{h,k}(y^{j_0}, \boldsymbol{u}^{j_0}) + \epsilon < J_{h,k}(y^{j_0}, \boldsymbol{u})$$

and taking the minimum on the right-hand side we get a contradiction.

Let us now denote by

$$\boldsymbol{u} = \{u_0, \boldsymbol{u}^\epsilon\}.$$

Arguing as before, we get

$$\hat{J}_{h,k}(y^i, \boldsymbol{u}) = hg(y^i, u_0) + \delta_h \hat{J}_{h,k}(z, \boldsymbol{u}^\epsilon),$$

where, as before, we have applied the same extended definition of $J_{h,k}$ as before for z different from a node. And then

$$\hat{J}_{h,k}(y^i, \boldsymbol{u}) \leq hg(y^i, u_0) + \delta_h w_{h,k}(y^i + hf(y^i, u_0)) + \epsilon.$$

Arguing as before

$$w_{h,k}(y^i) = \inf_{\boldsymbol{u} \in \mathcal{U}} \hat{J}_{h,k}(y^i, \boldsymbol{u}) \leq \inf_{\boldsymbol{u} \in \mathcal{U}} \left\{ hg(y^i, u_0) + \delta_h w_{h,k}(y^i + hf(y^i, u_0)) \right\} + \epsilon.$$

And since, on the one hand, (21) holds and, on the other, the above inequality is valid for any $\epsilon > 0$ we get

$$w_{h,k}(y^i) \leq \inf_{u \in U_{\text{ad}}} \left\{ hg(y^i, u) + \delta_h w_{h,k}(y^i + hf(y^i, u)) \right\}.$$

Then

$$w_{h,k}(y^i) = \inf_{u \in U_{\text{ad}}} \left\{ hg(y^i, u) + \delta_h w_{h,k}(y^i + hf(y^i, u)) \right\}.$$

□

In the rest of the paper we apply Theorem 3 with two different scenarios. In the first subsection, assuming enough regularity for the controls, we can get the full first order in time and space. In the second one, following [6], we make a proof using piecewise constants controls. However, adapting [6], which is written in the context of finite horizon to our infinite horizon problem, we loose the full order in time in the rate of convergence.

3.1 Error analysis assuming some regularity for the controls

Let us denote by

$$M_u := \max_n \max_{s \in [nh, (n+1)h)} \|u(s) - u(t_n)\|_2. \tag{22}$$

In the proof of next lemma we assume that the following condition holds for the controls

$$\lim_{h \to 0} M_u = 0. \tag{23}$$

Let us observe that assuming the controls are uniformly continuous condition (23) always holds.

Lemma 1 *Let $\hat{J}$ and $\hat{J}_{h,k}$ be the functionals defined in* (10) *and* (18) *respectively. Assume conditions* (3), (4), (5), (7), (8), (9) *and* (23) *hold. Let $y_0 = y^i$ for any $i = 1, \ldots, n_S$. Then*

$$\lim_{h \to 0, k \to 0} |\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| = 0, \tag{24}$$

where $u \in \mathbb{U}_{\text{ad}}$ and $\boldsymbol{u} = \{u_0, u_1, \ldots, \} = \{u(t_0), u(t_1), \ldots, \}$, $t_i = ih$, $i = 0, 1, \ldots,$.

Proof We argue similarly as in [5, Lemma 1.2, Chapter VI]. Let $y(t)$ be the solution of (1) and let us denote by $\tilde{y}(t) = \hat{y}_k$, $k = [t/h]$ where $\hat{y}_k$ is the solution of (19) with $\hat{y}_0 = y_0$. Let us denote by

$$\overline{u}(t) = u_k = u(t_k), \quad t \in [kh, (k+1)h). \tag{25}$$

Then, $\tilde{y}$ can be expressed as

$$\tilde{y}(t) = y_0 + \int_0^{[t/h]h} I_k f(\tilde{y}(s), \overline{u}(s))\, ds.$$

And,

$$y(t) - \tilde{y}(t) = \int_0^{[t/h]h} (f(y(s), u(s)) - I_k f(\tilde{y}(s), \overline{u}(s)))\ ds + \int_{[t/h]h}^{t} f(y(s), u(s))\ ds.$$

From the above equation, applying (5), we get

$$\|y(t) - \tilde{y}(t)\|_\infty \leq \int_0^{[t/h]h} \|f(y(s), u(s)) - I_k f(\tilde{y}(s), \overline{u}(s))\|_\infty\ ds + M_f h. \tag{26}$$

Let us bound now the term inside the integral. Adding and subtracting terms we get

$$\begin{aligned} \|f(y(s), u(s)) - I_k f(\tilde{y}(s), \overline{u}(s))\|_\infty &\leq \|f(y(s), u(s)) - f(y(s), \overline{u}(s))\|_\infty \\ &+ \|f(y(s), \overline{u}(s)) - I_k f(y(s), \overline{u}(s))\|_\infty + \|I_k f(y(s), \overline{u}(s)) - I_k f(\tilde{y}(s), \overline{u}(s))\|_\infty. \end{aligned} \tag{27}$$

For the first term on the right-hand side of (27) using (4) and (49) and assuming $s \in [kh, (k+1)h)$ we get

$$\begin{aligned} \|f(y(s), u(s)) - f(y(s), \overline{u}(s))\|_\infty \leq L_f \|u(s) - \overline{u}(s)\|_2 &= L_f \|u(s) - u(t_k)\|_2 \\ &\leq L_f M_u, \end{aligned} \tag{28}$$

for M_u defined in (22). Let us observe that assuming condition (23) holds the above term goes to zero as h goes to zero.

To bound the second term on the right-hand side of (27) arguing as in [4] we observe that for any $y \in \overline{\Omega}$ there exists an index l with $y \in \overline{S}_l \subset \overline{\Omega}$. Let us denote by J_l the index subset such that $y_i \in S_l$ for $i \in J_l$. Writing

$$y = \sum_{i=1}^{n_S} \mu_i y_i, \quad 0 \leq \mu_i \leq 1, \quad \sum_{i=1}^{n_S} \mu_i = 1,$$

it is clear that $\mu_i = 0$ holds for any $i \notin J_l$. Now, we observe that for any $u \in U_{\text{ad}}$ and $j = 1, \ldots, n$, applying (3) we get

$$\begin{aligned} |f_j(y, u) - I_k f_j(y, u)| &= |\sum_{i=1}^{n_S} \mu_i f_j(y, u) - \sum_{i=1}^{n_S} \mu_i I_k f_j(y_i, u)| \\ &= |\sum_{i \in J_l} \mu_i (f_j(y, u) - f_j(y_i, u)| \\ &\leq \sum_{i \in J_l} \mu_i L_f \|y - y_i\|_2 \leq L_f k, \end{aligned} \tag{29}$$

where in the last inequality we have applied $\|y - y_i\|_2 \leq k$, for $i \in J_l$. From the above inequality we get for the second term on the right-hand side of (27)

$$\|f(y(s), \overline{u}(s)) - I_k f(y(s), \overline{u}(s))\|_\infty \leq L_f k. \tag{30}$$

For the third term on the right-hand side of (27) we observe that the difference of the interpolation operator evaluated at two different points can be bounded in

terms of the constant gradient of the interpolant in the element to which those points belong times the difference of them, i.e.,

$$I_k f_j(y,u) - I_k f_j(\tilde{y},u) = \nabla I_k f_j(\tilde{y},u)\cdot(y-\tilde{y}) \leq \|\nabla I_k f_j(\tilde{y},u)\|_2 \|y-\tilde{y}\|_2.$$

Moreover, $\nabla I_k f_k$ can be bounded in terms of the lipschitz constant of f, L_f, more precisely, $\|\nabla I_k f_j(\tilde{y},u)\|_2 \leq C\sqrt{n}L_f$. Then,

$$|I_k f_j(y,u) - I_k f_j(\tilde{y},u)| \leq C L_f \sqrt{n}\|y-\tilde{y}\|_2.$$

As a consequence, for the third term on the right-hand side of (27) we get

$$\begin{aligned} \|I_k f(y(s),\overline{u}(s)) - I_k f(\tilde{y}(s),\overline{u}(s))\|_\infty &\leq C\sqrt{n}L_f\|y(s)-\tilde{y}(s)\|_2 \\ &\leq CnL_f\|y(s)-\tilde{y}(s)\|_\infty. \end{aligned} \tag{31}$$

Inserting (28), (30) and (31) into (26) we get for

$$\overline{L} = CnL_f, \tag{32}$$

$$\|y(t)-\tilde{y}(t)\|_\infty \leq \overline{L}\int_0^t \|y(s)-\tilde{y}(s)\|_\infty \, ds + tL_f(M_u+k) + M_f h.$$

Applying Gronwall's lemma we obtain

$$\|y(t)-\tilde{y}(t)\|_\infty \leq \frac{e^{\overline{L}t}}{\overline{L}}\left(tL_f(M_u+k)+M_f h\right). \tag{33}$$

Applying (9) and taking into account that $\|I_k g(y,u)\|_\infty \leq \|g(y,u)\|_\infty$ (where $\|\cdot\|_\infty$ refers to the L^∞ norm respect to the first argument) it is easy to check that

$$|\hat{J}(y_0,u) - J_{h,k}(y_0,\boldsymbol{u})| \leq X_1 + X_2 + X_3, \tag{34}$$

with

$$\begin{aligned} X_1 &= \left| h\sum_{n=0}^{[T/h]-1} \delta_h^n I_k g(\hat{y}_n,u_n) - \int_0^T g(y(s),u(s))e^{-\lambda s}\, ds \right|, \\ X_2 &= \left| h\sum_{n=[T/h]}^{\infty} M_g \delta_h^n \right|, \quad X_3 = \int_T^\infty M_g e^{-\lambda s}\, ds, \end{aligned}$$

and $T>0$ arbitrary. Now, we will estimate X_i. It is easy to see that

$$X_2 + X_3 \leq M_g h \frac{\delta_h^{[T/h]}}{\lambda h} + M_g \frac{e^{-\lambda T}}{\lambda}.$$

Since $\delta_h^{[T/h]} \to e^{-\lambda T}$ when $h\to 0$ then for any $\epsilon>0$ there exists $\overline{h}=\overline{h}(\epsilon,\lambda,M_g)>0$ and $\overline{T}=\overline{T}(\epsilon,\lambda,M_g)>0$ such that

$$X_2 + X_3 \leq \epsilon, \quad \text{for all } 0<h\leq\overline{h}, \quad T\geq\overline{T}. \tag{35}$$

In the argument that follows we fix $T = \overline{T}$. We observe that

$$X_1 = \left| \int_0^{[T/h]h} I_k g(\tilde{y}(s), \overline{u}(s)) \delta_h^{[s/h]} \, ds - \int_0^T g(y(s), u(s)) e^{-\lambda s} \, ds \right|.$$

Then, we can write

$$\begin{aligned} X_1 &\leq X_{1,1} + X_{1,2} + X_{1,3} + X_{1,4} + X_{1,5} \qquad (36) \\ &:= \left| \int_0^{[T/h]h} I_k g(\tilde{y}(s), \overline{u}(s)) (\delta_h^{[s/h]} - e^{-\lambda s}) \, ds \right| \\ &\quad + \left| \int_0^{[T/h]h} I_k \left(g(\tilde{y}(s), \overline{u}(s)) - I_k g(y(s), \overline{u}(s)) \right) e^{-\lambda s} \, ds \right| \\ &\quad + \left| \int_0^{[T/h]h} \left(I_k g(y(s), \overline{u}(s)) - g(y(s), \overline{u}(s)) \right) e^{-\lambda s} \, ds \right| \\ &\quad + \left| \int_0^{[T/h]h} \left(g(y(s), \overline{u}(s)) - g(y(s), u(s)) \right) e^{-\lambda s} \, ds \right| \\ &\quad + \left| \int_{[T/h]h}^T g(y(s), u(s)) e^{-\lambda s} \, ds \right|. \end{aligned}$$

We will bound the terms on the right-hand side of (36). To bound the first term we will apply again $\|I_k g(y,u)\|_\infty \leq \|g(y,u)\|_\infty$ and (9) to obtain

$$\begin{aligned} X_{1,1} &= \left| \int_0^{[T/h]h} I_k g(\tilde{y}(s), \overline{u}(s)) (\delta_h^{[s/h]} - e^{-\lambda s}) \, ds \right| \\ &\leq \int_0^T |I_k g(\tilde{y}(s), \overline{u}(s))| |\delta_h^{[s/h]} - e^{-\lambda s}| \, ds \leq M_g \int_0^T |\delta_h^{[s/h]} - e^{-\lambda s}| \, ds. \end{aligned}$$

Now we write $\delta_h^{[s/h]} = e^{-\lambda \theta [s/h] h}$, for $\theta = -\log(\delta_h)/(\lambda h)$. Applying the mean value theorem to the function $e^{-\lambda s}$ and taking into account that since $[s/h]h \leq s \leq [s/h]h + h$ then $|s - \theta[s/h]h| \leq (\theta - 1)T + \theta h$ and that $\theta \to 1$ when $h \to 0$ then we get

$$X_{1,1} \leq M_g T \lambda ((\theta - 1)T + \theta h) \leq \epsilon, \qquad (37)$$

for $h \leq \overline{h}$, with $\overline{h} = \overline{h}(\epsilon, \overline{T}, \lambda, M_g)$.

To bound the next term we argue as in (31) and use (33) to get

$$X_{1,2} \leq \frac{CnL_g}{\overline{L}} \int_0^T e^{\overline{L}s} \left(sL_f(M_u + k) + M_f h \right) e^{-\lambda s} \, ds \leq C_1 h + C_2 k + C_3 M_u,$$

where

$$C_1 = \frac{CnL_g}{\overline{L}} \int_0^T e^{(\overline{L} - \lambda)s} M_f \, ds, \quad C_2 = \frac{C\sqrt{n}L_g}{\overline{L}} \int_0^T e^{(\overline{L} - \lambda)s} sL_f \, ds,$$

and

$$C_3 = \frac{CnL_g}{\overline{L}} \int_0^T e^{(\overline{L} - \lambda)s} sL_f \, ds.$$

Then, assuming condition (23) holds to assure convergence for the third term in the error bound of $X_{1,2}$

$$X_{1,2} \leq \epsilon, \quad h \leq \overline{h}, \ k \leq \overline{k}, \tag{38}$$

where $\overline{h} = \overline{h}(\epsilon, \overline{T}, \lambda, C, \overline{L}, n, L_g, L_f, M_u, M_f)$, $\overline{k} = \overline{k}(\epsilon, \overline{T}, \lambda, C, \overline{L}, n, L_g, L_f)$. For the third term, arguing as in (29) we get

$$X_{1,3} \leq L_g k \int_0^T e^{-\lambda s} \, ds \leq \epsilon, \quad k \leq \overline{k} = \overline{k}(\epsilon, \overline{T}, \lambda, L_g). \tag{39}$$

To bound the forth term we apply (7) and (49) to get

$$|g(y(s), \overline{u}(s)) - g(y(s), u(s))| \leq L_g \|u(s) - u(t_k)\|_2 \leq L_g M_u, \quad s \in [kh, (k+1)h).$$

And then, assuming again condition (23) holds

$$X_{1,4} \leq L_g M_u \int_0^T e^{-\lambda s} \, ds \leq \epsilon, \quad h \leq \overline{h} = \overline{h}(\epsilon, \overline{T}, \lambda, L_g, M_u). \tag{40}$$

Finally, for the last term on the right-hand side of (36), applying (9),

$$X_{1,5} \leq M_g h \leq \epsilon, \quad h \leq \overline{h} = \overline{h}(\epsilon, M_g). \tag{41}$$

Inserting (37), (38), (39), (40) and (41) into (36) and taking into account (34) and (35) we conclude (24). □

Lemma 2 *Let $\hat{J}$ and $\hat{J}_{h,k}$ be the functionals defined in* (10) *and* (18) *respectively. Assume conditions* (3), (4), (5), (7), (8), (9) *hold. Assume $\lambda > \overline{L}$ with $\overline{L}$ the constant in* (32). *Then, for $0 \leq h \leq 1/(2\lambda)$ there exist positive constants $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$ and $C_2 = C_2(\lambda, L_f, L_g)$ such that for $y_0 = y^i$, $i = 1, \ldots, n_S$*

$$|\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| \leq C_1(h + k) + C_2 M_u, \tag{42}$$

where $u \in \mathbb{U}_{\text{ad}}$ and $\boldsymbol{u} = \{u_0, u_1, \ldots, \} = \{u(t_0), u(t_1), \ldots, \}$, $t_i = ih$, $i = 0, 1, \ldots,$ and M_u is defined in (22).

Proof We argue as in [5, Lemma 1.2, Chapter VI]. Let $y(t)$ be the solution of (1) and let us denote by $\tilde{y}(t) = \hat{y}_k$, $k = [t/h]$ where $\hat{y}_k$ is the solution of (19) with $\hat{y}_0 = y_0$. Let us denote by

$$\overline{u}(t) = u_k = u(t_k), \quad t \in [kh, (k+1)h).$$

We first observe that

$$|\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| \leq X + Y,$$

where

$$\begin{aligned} X &= \int_0^\infty |g(y(s), u(s)) - I_k g(\tilde{y}(s), \overline{u}(s))| e^{-\lambda s} \, ds, \\ Y &= \int_0^\infty |I_k g(\tilde{y}(s), \overline{u}(s))| |e^{-\lambda s} - e^{-\lambda \theta [s/h] h}| \, ds, \end{aligned}$$

and, as in Lemma 1, $\theta = -\log(\delta_h)/(\lambda h)$. To bound X we decompose

$$\begin{aligned}|g(y(s),u(s)) - I_k g(\tilde{y}(s),\overline{u}(s))| \;\leq\; & |g(y(s),u(s)) - g(y(s),\overline{u}(s))| \\ & + |g(y(s),\overline{u}(s)) - I_k g(y(s),\overline{u}(s))| \\ & + |I_k g(y(s),\overline{u}(s)) - I_k g(\tilde{y}(s),\overline{u}(s))|.\end{aligned}$$

Arguing as in Lemma 1 we get

$$|g(y(s),u(s)) - I_k g(\tilde{y}(s),\overline{u}(s))| \;\leq\; L_g M_u + L_g k + CnL_g \|y(s) - \tilde{y}(s)\|_\infty$$

and then applying (33) we obtain

$$X \leq \frac{1}{\lambda}\left(L_g M_u + L_g k\right) + C\frac{nL_g}{\overline{L}} \int_0^\infty \left(sL_f(M_u + k) + M_f h\right) e^{(\overline{L}-\lambda)s}\, ds.$$

Then, for $\lambda > \overline{L}$ there exist constants $C_1 = C_1(\lambda, M_f, L_f, L_g, L_u)$ and $C_2(\lambda, L_f, L_g)$ such that

$$X \leq C_1(h+k) + C_2 M_u.$$

To bound Y we first observe that, arguing as before, $|I_k g(\tilde{y}(s),\overline{u}(s))| \leq M_g$ and then

$$Y \leq M_g \int_0^\infty |e^{-\lambda s} - e^{-\lambda\theta[s/h]h}|\, ds.$$

Applying the mean value theorem

$$Y \leq M_g \int_0^\infty \lambda \max\left\{e^{-\lambda s}, e^{-\lambda\theta[s/h]h}\right\} |\lambda s - \lambda\theta[s/h]h|\, ds.$$

Now, since $|s - \theta[s/h]h| \leq (\theta-1)s + \theta h$ we get

$$Y \leq M_g \lambda^2 e^{\theta\lambda h} \int_0^\infty e^{-\lambda s}((\theta-1)s + \theta h)\, ds,$$

and since both

$$\int_0^\infty s e^{-\lambda s}\, ds, \quad \int_0^\infty e^{-\lambda s}\, ds,$$

are bounded and $\lim_{h\to 0}(\theta-1)h = \lambda/2$ (for $0 \leq h \leq 1/(2\lambda)$ the function $(\theta-1)/h$ is an increasing function bounded by its value at $h = 1/(2\lambda)$, i.e., $(\theta-1)/h \leq 2\lambda(2\log(2)-1)$, we conclude

$$Y \leq Ch, \quad C = C(M_g, \lambda) > 0.$$

□

Theorem 4 *Assume conditions* (3), (4), (5), (7), (8) *and* (9) *hold. Assume* $\lambda > \overline{L}$ *with* $\overline{L}$ *the constant in* (32). *Then, for* $0 \leq h \leq 1/(2\lambda)$ *there exists positive constants* $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$ *and* $C_2 = C_2(\lambda, L_f, L_g)$ *such that*

$$|v(y) - v_{h,k}(y)| \leq C_1(h+k) + L_v k + C_2 M_u, \quad y \in \mathbb{R}^n, \tag{43}$$

and M_u *is defined in* (22).

Proof For any $y \in \overline{\Omega}$ there exists an index l with $y \in \overline{S}_l \subset \overline{\Omega}$. Let us denote by J_l the index subset such that $y_i \in S_l$ for $i \in J_l$. We can write

$$y = \sum_{i \in J_l} \mu_i y_i, \quad 0 \leq \mu_i \leq 1, \quad \sum_{i \in J_l} \mu_i = 1.$$

Now, let us observe that

$$|v(y) - \sum_{i \in J_l} \mu_i v(y_i)| = |\sum_{i \in J_l} \mu_i v(y) - \sum_{i \in J_l} \mu_i v(y_i)| = |\sum_{i \in J_l} \mu_i (v(y) - v(y_i))| \leq L_v k, \quad (44)$$

where L_v is the Lipschitz constant of v. Then

$$\begin{aligned} v(y) - v_{h,k}(y) &= \left(v(y) - \sum_{i \in J_l} \mu_i v(y_i) \right) + \left(\sum_{i \in J_l} \mu_i (v(y_i) - v_{h,k}(y_i)) \right) \\ &\leq L_v k + \left(\sum_{i \in J_l} \mu_i (v(y_i) - v_{h,k}(y_i)) \right). \end{aligned} \quad (45)$$

For any $i \in J_l$, in view of (20) let us denote by $\overline{\boldsymbol{u}}^i \in \mathcal{U}$ a control giving the minimum

$$v_{h,k}(y_i) = \hat{J}_{h,k}(y_i, \overline{\boldsymbol{u}}^i).$$

Then

$$v(y_i) - v_{h,k}(y_i) \leq \hat{J}(y_i, \overline{u}^i) - J_{h,k}(y_i, \overline{\boldsymbol{u}}^i),$$

where $\overline{u}^i \in \mathbb{U}_{\rm ad}$ such that $\overline{u}^i(t_j) = \overline{u}^i_j$, $\overline{\boldsymbol{u}}^i = \left\{ \overline{u}^i_0, \overline{u}^i_1, \dots, \right\}$. Applying (42), there exist positive constants $C_1 = C(\lambda, M_f, M_g, L_f, L_g)$ and $C_2 = C(\lambda, L_f, L_g)$ such that

$$v(y_i) - v_{h,k}(y_i) \leq C_1(h + k) + C_2 M_u. \quad (46)$$

Consequently, since $\sum_{i \in J_l} \mu_i = 1$, from (45) and (46)

$$\begin{aligned} v(y) - v_{h,k}(y) &\leq L_v k + \left(\sum_{i \in J_l} \mu_i (v(y_i) - v_{h,k}(y_i)) \right) \\ &\leq L_v k + C_1(h + k) + C_2 M_u. \end{aligned} \quad (47)$$

Now, for any $i \in J_l$ let us denote by $\underline{u}^i \in \mathbb{U}_{\rm ad}$ the control giving the minimum in (11) and let us denote by $\underline{\boldsymbol{u}}^i = \left\{ \underline{u}^i(t_0), \underline{u}^i(t_1), \dots, \right\}$. Then, arguing as before there exist positive constants $C_1 = C(\lambda, M_f, M_g, L_f, L_g)$ and $C_2 = C(\lambda, L_f, L_g)$ such that

$$v_{h,k}(y_i) - v(y_i) \leq \hat{J}_{h,k}(y_i, \underline{\boldsymbol{u}}^i) - \hat{J}(y_i, \underline{u}^i) \leq C_1(h + k) + C_2 M_u, \quad (48)$$

Arguing as before, we can write

$$v_{h,k}(y) - v(y) = \sum_{i \in J_l} \mu_i v_{h,k}(y_i) - v(y) \leq \sum_{i \in J_l} \mu_i (v_{h,k}(y_i) - v(y_i)) + \sum_{i \in J_l} \mu_i v(y_i) - v(y)$$

Applying (44) and (48) we get

$$v_{h,k}(y) - v(y) \leq L_v k + C_1(h + k) + C_2 M_u$$

which together with (55) implies (43). □

Remark 1 We observe that Theorem 4 gives a bound for the error without assuming any regularity on the controls. The error bound (43) has two terms. The first one is always first order convergent both in time and space. The second one depends on the size of M_u defined in (22), which depends on the regularity of the controls. As it is always the case when one applies numerical methods, some regularity is required to achieve the best rate of convergence as possible. In Theorems 5 and 6 below we get as a consequence of Theorem 4 two possible scenarios. In Theorem 5, assuming condition (23) holds for the controls (which is true for uniformly continuous controls) we prove convergence. In Theorem 6, assuming that the controls are Lipschitz-continuous, (49), we prove that M_u in (22) behaves as h, so that, in (43), the optimal rate of convergence of order one both in time and space is recovered. Although in a concrete problem one could not have enough regularity for the controls to achieve first order of convergence in time (observe that the rate of convergence is always one in space) the error bound (43) can always be applied. This error bound allow us to identify the different sources of the error in the method and consequently is able to explain the behavior of the method. As a conclusion, we can discard a behaviour for the rate of convergence as $O(k/h)$, as the existing bounds in the literature had predicted.
In the following theorem we deduce convergence as h goes to zero assuming condition (23) holds.

Theorem 5 *Assume conditions* (3), (4), (5), (7), (8), (9) *and* (23) *hold. Assume* $\lambda > \overline{L}$ *with* $\overline{L}$ *the constant in* (32). *Then,*

$$\lim_{h\to 0,k\to 0} |v(y) - v_{h,k}(y)| = 0, \quad y \in \mathbb{R}^n.$$

Proof The conclusion is obtained as a corollary of Theorem 4 applying (23) to bound the second term in (43). □

If we assume that the controls are Lipschitz-continuous; i.e., there exists a positive constant $L_u > 0$ such that

$$\|u(t) - u(s)\|_2 \leq L_u|t - s|. \tag{49}$$

we can prove first order of convergence both in time and space.

Theorem 6 *Assume conditions* (3), (4), (5), (7), (8), (9) *and* (49) *hold. Assume* $\lambda > \overline{L}$ *with* $\overline{L}$ *the constant in* (32). *Then, for* $0 \leq h \leq 1/(2\lambda)$ *there exist positive constants* $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$ *and* $C_2 = C_2(\lambda, L_f, L_g, L_u)$ *such that*

$$|v(y) - v_{h,k}(y)| \leq C_1(h + k) + L_v k + C_2 h, \quad y \in \mathbb{R}^n.$$

Proof The conclusion is obtained as a corollary of Theorem 4 applying (49) to bound the second term in (43). □

Remark 2 The regularity requirements on the controls can still be weakened. Let us assume that the controls have a finite number of jump discontinuities: $t_1^* < t_2^* < \ldots < t_l^*$,. For a fixed h let us denote by $I_j^* = [m_j h, (m_j + 1)h)$, $m_j \in \mathbb{N}$, the interval

such that $t_j^* \in I_j^*$, $j = 1, \ldots, l$. Then, the arguments of Lemma 2 can be adapted to get instead of (42) the following bound

$$|\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| \leq C_1(h+k) + C_2 h + C_3 M_u^*, \quad y_0 = y^i,\ i = 1, \ldots, n_S, \tag{50}$$

where $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$, $C_2 = C_2(\lambda, L_f, L_g, M_s)$ and $C_3 = C_3(\lambda, L_f, L_g)$ and

$$\begin{aligned} M_u^* &:= \max_n \max_{s \in [nh,(n+1)h), s \notin I} \|u(s) - u(t_n)\|_2, \quad I = I_1^* \cup \ldots \cup I_l^*, \\ M_s &:= \max_{1 \leq j \leq l} \max_{s \in I_j^*} \|u(s) - u(t_{m_j})\|_2. \end{aligned}$$

The idea is to use the additive property of integrals to isolate those corresponding to intervals I_j^*, $j = 1, \ldots, l$. To bound the terms involving integrals over I_j^* one can apply the boundedness of the integrand together with the fact that the diameter of I_j^* is equal h for $j = 1, \ldots, l$. The union of the bounds corresponding to integrals over I_j^* gives rise to the second term on the right-hand-side of (50).

Accordingly, instead of (43) in Theorem 4, one can prove

$$|v(y) - v_{h,k}(y)| \leq C_1(h+k) + L_v k + C_2 h + C_3 M_u^*, \quad y \in \mathbb{R}^n, \tag{51}$$

with C_1, C_2 and C_3 the constants in (50).

From (51), and depending on the regularity of the controls, one gets either convergence or convergence of order one, arguing as in Theorems 5 and 6 but with the assumptions on the controls restricted to the finite number of intervals:

$$[0, t_1^*), \ldots [t_j^*, t_{j+1}^*), \ldots, [t_l^*, \infty).$$

Moreover, assuming that the controls are Hölder continuous over those intervals of order α, for any $0 < \alpha < 1$, one gets an error bound in time of size $O(h^\alpha)$, applying (51).

3.2 Error analysis arguing with piecewise constants controls

In this section we adapt the error analysis in [6] for finite horizon problems to our context of infinite horizon problems to get a weaker result for the rate of convergence but weakening also the regularity assumptions over the controls.

Let us denote by

$$\mathbb{U}_{\text{ad}}^{pc} = \{u \in \mathbb{U} \mid u(t) = u_k \in U_{\text{ad}},\ t \in [t_k, t_{k+1})\},$$

with u_k constant. Let us observe that we can consider the continuous problem for controls in $\mathbb{U}_{\text{ad}}^{pc}$. The following lemmas are a direct consequence of Lemmas 1 and 2

Lemma 3 *Let $\hat{J}$ and $\hat{J}_{h,k}$ be the functionals defined in* (10) *and* (18) *respectively. Assume conditions* (3), (4), (5), (7), (8), (9). *Then*

$$\lim_{h \to 0, k \to 0} |\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| = 0, \quad y_0 = y^i,\ i = 1, \ldots, n_S, \tag{52}$$

where $u \in \mathbb{U}_{\text{ad}}^{pc}$ and $\boldsymbol{u} = \{u_0, u_1, \ldots, \}$ with $u_k = u(t)$, $t \in [t_k, t_{k+1})$.

Lemma 4 *Let $\hat{J}$ and $\hat{J}_{h,k}$ be the functionals defined in* (10) *and* (18) *respectively. Assume conditions* (3), (4), (5), (7), (8), (9) *hold. Assume $\lambda > \overline{L}$ with $\overline{L}$ the constant in* (32). *Then, for $0 \le h \le 1/(2\lambda)$ there exists a positive constant $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$ such that*

$$|\hat{J}(y_0, u) - \hat{J}_{h,k}(y_0, \boldsymbol{u})| \le C_1(h+k), \quad y_0 = y^i,\ i = 1, \dots, n_S, \tag{53}$$

where $u \in \mathbb{U}_{\rm ad}^{pc}$ and $\boldsymbol{u} = \{u_0, u_1, \dots, \}$ with $u_k = u(t)$, $t \in [t_k, t_{k+1})$.

The proof of Lemmas 3 and 4 is obtained taking in the proofs of Lemmas 1 and 2 $M_u = 0$ since for piecewise controls it holds, see (25),

$$\overline{u}(t) = u_k = u(t),\ \forall t \in [kh, (k+1)h).$$

The following theorem is analogous to Theorem 4. For the proof we need to assume an additional convexity assumption, see [6, (A4)],

- (CA) For every $y \in \mathbb{R}^n$,

  $$\{f(y,u), g(y,u), \quad u \in U_{\rm ad}\}$$

  is a convex subset of $\mathbb{R}^{n+1}$.

Theorem 7 *Assume conditions* (3), (4), (5), (7), (8), (9) *and* (CA) *hold. Assume $\lambda > \overline{L}$ with $\overline{L}$ the constant in* (32). *Then, for $0 \le h \le 1/(2\lambda)$ there exist positive constants $C_1 = C_1(\lambda, M_f, M_g, L_f, L_g)$ and $C_2 = C_2(\lambda, M_f, M_g, L_f, L_g)$ such that for $y \in \mathbb{R}^n$*

$$|v(y) - v_{h,k}(y)| \le C_1(h+k) + C_2 \frac{1}{(1+\beta)^2\lambda^2} (\log(h))^2 h^{\frac{1}{1+\beta}}, \quad \beta = \frac{\sqrt{n} L_f}{\lambda}. \tag{54}$$

Proof For simplicity we will assume $y = y^i$ for any $i = 1, \dots n_S$ since the general case can be proved arguing exactly as in Theorem 4.

In view of (20) let us denote by $\overline{\boldsymbol{u}} \in \mathcal{U}$ a control giving the minimum

$$v_{h,k}(y) = \hat{J}_{h,k}(y, \overline{\boldsymbol{u}}).$$

Then

$$v(y) - v_{h,k}(y) \le \hat{J}(y, \overline{u}) - J_{h,k}(y, \overline{\boldsymbol{u}}),$$

where $\overline{u} \in \mathbb{U}_{\rm ad}^{pc}$ such that $\overline{u}(t) = \overline{u}_i$, $t \in [t_i, t_{i+1})$. Applying (53), there exists a positive constant $C_1 = C(\lambda, M_f, M_g, L_f, L_g)$ such that

$$v(y) - v_{h,k}(y) \le C_1(h+k). \tag{55}$$

Now, let us denote by $\underline{u} \in \mathbb{U}_{\rm ad}$ the control giving the minimum in (11) so that

$$v(y) = \hat{J}(y, \underline{u}) = \int_0^\infty g(y(t), \underline{u}(t)) e^{-\lambda t} dt. \tag{56}$$

The following argument is taken from [6, Appendix B].

For any t_k we can write

$$y(t) = y(t_k) + \int_{t_k}^{t} f(y(s), \underline{u}(s))ds.$$

Applying (5)

$$\|y(t) - y(t_k)\|_\infty \leq M_f h.$$

Then, for any $t \in [t_k, t_{k+1}]$, using the above inequality and (3) we obtain

$$\left\| \int_{t_k}^{t} f(y(s), \underline{u}(s)) - f(y(t_k), \underline{u}(s))ds \right\|_\infty \leq \sqrt{n} L_f M_f h^2.$$

As a consequence, we get

$$y(t) = y(t_k) + \int_{t_k}^{t} f(y(t_k), \underline{u}(s))ds + \epsilon_k, \quad \|\epsilon_k\|_\infty \leq \sqrt{n} L_f M_f h^2. \tag{57}$$

On the other hand, as in [6, (B.6a), (B.6b)], thanks to (CA), for any k, there exists $\underline{u}_k$ such that

$$\int_{t_k}^{t_{k+1}} f(y(t_k), \underline{u}(s))ds = h f(y(t_k), \underline{u}_k) \tag{58}$$

$$\int_{t_k}^{t_{k+1}} g(y(t_k), \underline{u}(s))e^{-\lambda s}ds \leq h g(y(t_k), \underline{u}_k))e^{-\lambda t_k}. \tag{59}$$

From (59) and (9) we get

$$\begin{aligned} \int_{t_k}^{t_{k+1}} (g(y(t_k), \underline{u}(s) - g(y(t_k), \underline{u}_k))e^{-\lambda s}ds &\leq \int_{t_k}^{t_{k+1}} g(y(t_k), \underline{u}_k)(e^{-\lambda t_k} - e^{-\lambda s})ds \\ &\leq \lambda M_g h^2. \end{aligned} \tag{60}$$

From (57) and (58)

$$y(t) = y(t_k) + \int_{t_k}^{t} f(y(t_k), \underline{u}_k)ds + \epsilon_k, \quad \|\epsilon_k\|_\infty \leq \sqrt{n} L_f M_f h^2.$$

Let us denote by y^{pc} the time-continuous trayectory solution with the same initial condition as y associated to the control $\underline{u}^{pc}(t) = \underline{u}_k, \ \forall t \in [t_k, t_{k+1})$.

Arguing as above we get

$$y^{pc}(t) = y^{pc}(t_k) + \int_{t_k}^{t} f(y^{pc}(t_k), \underline{u}_k)ds + \epsilon'_k, \quad \|\epsilon'_k\|_\infty \leq \sqrt{n} L_f M_f h^2. \tag{61}$$

Subtracting (61) from (57) and using (3) we obtain

$$\begin{aligned} \|y(t_k) - y^{pc}(t_k)\|_\infty &\leq \|y(t_{k-1}) - y^{pc}(t_{k-1})\|_\infty + \sqrt{n} h L_f \|y(t_{k-1}) - y^{pc}(t_{k-1})\|_\infty \\ &\quad + \|\epsilon_{k-1}\|_\infty + \|\epsilon'_{k-1}\|_\infty \\ &\leq (1 + h\sqrt{n} L_f)\|y(t_{k-1}) - y^{pc}(t_{k-1})\|_\infty + \|\epsilon_{k-1}\|_\infty + \|\epsilon'_{k-1}\|_\infty. \end{aligned}$$

Since $y(y_0) = y^{pc}(t_0)$ by standard recursion we get

$$\|y(t_k) - y^{pc}(t_k)\|_\infty \leq e^{t_k\sqrt{n}L_f} \sum_{0\leq l\leq k-1} \left(\|\epsilon_l\|_\infty + \|\epsilon_l'\|_\infty\right) \leq 2t_k e^{\sqrt{n}t_k L_f}\sqrt{n}L_f M_f h. \tag{62}$$

For the control $\underline{u} \in \mathbb{U}_{\rm ad}$ giving the minimum in (11) and for $\underline{\boldsymbol{u}} = \{\underline{u}_0, \ldots \underline{u}_k, \ldots, \}$. we obtain

$$v_{h,k}(y) - v(y) \leq \hat{J}_{h,k}(y, \underline{\boldsymbol{u}}) - \hat{J}(y, \underline{u}) = \hat{J}_{h,k}(y, \underline{\boldsymbol{u}}) - \hat{J}(y^{pc}, \underline{u}^{pc}) + \hat{J}(y^{pc}, \underline{u}^{pc}) - \hat{J}(y, \underline{u}).$$

The first term on the right-hand side above is bounded in Lemma 4 so that

$$v_{h,k}(y) - v(y) \leq C_1(h+k) + \hat{J}(y^{pc}, \underline{u}^{pc}) - \hat{J}(y, \underline{u}).$$

To conclude we need to bound the second term. We write

$$\begin{aligned}\hat{J}(y^{pc}, \underline{u}^{pc}) - \hat{J}(y, \underline{u}) &= \int_0^T \left(g(y^{pc}(s), \underline{u}^{pc}(s)) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds \\ &\quad + \int_T^\infty \left(g(y^{pc}(s), \underline{u}^{pc}(s)) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds. \end{aligned} \tag{63}$$

For the second term on the right-hand side above, applying (9) we get

$$\left| \int_T^\infty \left(g(y^{pc}(s), \underline{u}^{pc}(s)) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds \right| \leq 2M_g \int_T^\infty e^{-\lambda s} ds = 2M_g \frac{e^{-\lambda T}}{\lambda}.$$

Let

$$e^{-\lambda T} = h^{1/(1+\beta)}, \quad \beta = \frac{\sqrt{n}L_f}{\lambda}.$$

Then

$$T = \log(h^{-1/(1+\beta)\lambda}).$$

We fix the above value of T so that from (63) we get

$$\begin{aligned}\left|\hat{J}(y^{pc}, \underline{u}^{pc}) - \hat{J}(y, \underline{u})\right| &\leq \int_0^T \left(g(y^{pc}(s), \underline{u}^{pc}(s)) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds \\ &\quad + \frac{2M_g}{\lambda} h^{\frac{1}{1+\beta}}. \end{aligned} \tag{64}$$

To conclude we will bound the first term on the right-hand side of (64).

Also, for simplicity we assume there exists an integer N such that $T = N\Delta t$. For the first term on the right-hand side of (64) we have

$$\begin{aligned}&\int_0^T \left(g(y^{pc}(s), \underline{u}^{pc}(s)) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds = \\ &\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \left(g(y^{pc}(s), \underline{u}_k) - g(y(s), \underline{u}(s))\right) e^{-\lambda s} ds.\end{aligned}$$

Adding and subtracting terms we get

$$\begin{aligned}
&\int_{t_k}^{t_{k+1}} \left(g(y^{pc}(s),\underline{u}_k) - g(y(s),\underline{u}(s))\right) e^{-\lambda s} ds = \\
&\quad \int_{t_k}^{t_{k+1}} \left(g(y^{pc}(s),\underline{u}_k) - g(y^{pc}(t_k),\underline{u}_k)\right) e^{-\lambda s} ds \\
&\quad + \int_{t_k}^{t_{k+1}} \left(g(y^{pc}(t_k),\underline{u}_k) - g(y(t_k),\underline{u}_k)\right) e^{-\lambda s} ds \\
&\quad + \int_{t_k}^{t_{k+1}} \left(g(y(t_k),\underline{u}_k) - g(y(t_k),\underline{u}(s))\right) e^{-\lambda s} ds \\
&\quad + \int_{t_k}^{t_{k+1}} \left(g(y(t_k),\underline{u}(s)) - g(y(s),\underline{u}(s))\right) e^{-\lambda s} ds
\end{aligned}$$

Applying (7), (62) and (60) we get

$$\begin{aligned}
&\int_{t_k}^{t_{k+1}} \left(g(y^{pc}(s),\underline{u}_k) - g(y(s),\underline{u}(s))\right) e^{-\lambda s} ds \le 2\sqrt{n} h^2 L_g M_f + \lambda h^2 M_g \\
&\quad + 2nL_g h^2 t_k e^{\sqrt{n} t_k L_f} L_f M_f.
\end{aligned}$$

And then

$$\begin{aligned}
&\int_0^T \left(g(y^{pc}(s),\underline{u}^{pc}(s)) - g(y(s),\underline{u}(s))\right) e^{-\lambda s} ds \le Th\left(2\sqrt{n} L_g M_f + \lambda M_g\right) \\
&\quad + 2nL_g T^2 h e^{\sqrt{n} T L_f} L_f M_f.
\end{aligned}$$

To conclude, we observe that with the definition of T we get

$$h e^{\sqrt{n} T L_f} = h^{\frac{1}{1+\beta}},$$

since we have chosen T and β to optimize the rate of convergence. The above term, together with the last term in (64) are the terms that produce a reduction in the rate of convergence compared with the finite horizon case. We finally obtain

$$v_{h,k}(y) - v(y) \le C_1(h+k) + C_2 \frac{1}{(1+\beta)^2 \lambda^2} (\log(h))^2 h^{\frac{1}{1+\beta}}, \quad \beta = \frac{\sqrt{n} L_f}{\lambda}.$$

□

Remark 3 Let us observe that in view of (54) and taking into account that β is smaller than 1 then we loose at most half an order in the rate of convergence in time of the method up to a logarithmic term. This comes from adapting the arguments in [6] in the context of finite horizon problems to our infinite horizon case.

References

[1] M. Akian, S. Gaubert, A. Lakhoua, *The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis*, SIAM J. Control Optim. 47 (2008), 817–848.

[2] A. Alla, B. Haasdonk & A. Schmidt, *Feedback control of parametrized PDEs via model order reduction and dynamic programming principle*, Adv. Comput. Math. 46 (2020), 28 pp.

[3] A. Alla, M. Falcone & D. Kalise, *An efficient policy iteration algorithm for dynamic programming equations*, SIAM J. Sci. Comput., 37 (2015), A181–A200.

[4] A. Alla, M. Falcone & S. Volkwein, *Error analysis for POD approximations of infinite horizon problems via the dynamic programming approach*, SIAM J. Control Optim., 55 (2017), 3091–3115.

[5] M. Bardi & I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Belmann Equations*, Springer Science+Business Media, LLC, New York, 1997.

[6] O. Bokanowski, N. Gammoudi & H. Zidani, *Optimistic planning algorithms for state-constrained optimal control*, Comput. Math. Appl. 109 (2022), 158–179.

[7] O. Bokanowski, J. Garcke, M. Griebel & I. Klompmaker, *An adaptive sparse grid semi-Lagrangian scheme for first order Hamilton-Jacobi Bellman equations*, J. Sci. Comput. 55 (2013), 575–605.

[8] E. Carlini, M. Falcone & R. Ferretti, *An efficient algorithm for Hamilton-Jacobi equations in high dimension*, Comput. and Visualization in Science, 7 (2004), 15–29.

[9] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), 1–13.

[10] M. Falcone, *Corrigenda: "A numerical approach to the infinite horizon problem of deterministic control theory" [App. Math. Optim. 15 (1987), 1–13]*, Appl. Math. Optim., 23 (1991), 213–214.

[11] M. Falcone, *Numerical solution of dynamic programming equations*, in Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Belmann Equations, Springer Science+Business Media, LLC, New York, 1997, 471-504.

[12] M. Falcone, R. Ferretti, *Semi-Lagrangian Approximation Schemes for Linear and Hamilton-Jacobi Equations*, SIAM, Philadelphia, 2014, ISBN: 978-1-611973-04-4.

[13] R. Gonzalez & E. Rofman, *On deterministic control problems: an approximation procedure for the optimal cost I. The stationary problem.*, SIAM J. Control Optimization, 23 (1985), 242–266.

[14] B-Z. Guo & T-T, Wu, *Approximation of optimal feedback control: a dynamic programming approach*, J. Global Optim. 46 (2010), 395–422.

[15] M. M. Tidball, *Comments on "A numerical approach to the infinite horizon problem of deterministic control theory"* , Appl. Math. Optim., 23 (1991), 209–211.