PRIMAL-DUAL REGRESSION APPROACH FOR MARKOV DECISION PROCESSES WITH GENERAL STATE AND ACTION SPACES

DENIS BELOMESTNY[1] AND JOHN SCHOENMAKERS[2]

Abstract. We develop a regression-based primal-dual martingale approach for solving discrete time, finite horizon MDPs with state and action spaces that are general in the sense that they may be finite or infinite (but regular enough) subsets of Euclidean space. As a result, our method allows for the construction of tight upper and lower biased approximations of the value functions and provides tight approximations to the optimal policy. In particular, we prove error bounds for the estimated duality gap featuring polynomial dependence on the time horizon and sublinear dependence of the stochastic part of the error on the cardinality/dimension of the state and action spaces. From a computational point of view, the proposed method is efficient since, in contrast to the usual duality-based methods for optimal control problems in the literature, the Monte Carlo procedures involved here do not require nested simulations.

1. Introduction

Markov decision processes (MDPs) provide a general framework for modeling sequential decision-making under uncertainty. A large number of practical problems from various areas such as economics, finance, and machine learning can be viewed as MDPs. For a classical reference we refer to [22], and for MDPs with application to finance, see [5]. The aim is usually to find an optimal policy that maximizes the expected accumulated rewards (or minimizes the expected accumulated costs). In principle, these Markov decision problems can be solved by a dynamic programming approach; however, in practice, this approach suffers from the so-called "curse of dimensionality" and the "curse of horizon" meaning that the complexity of the program increases exponentially in the dimension of the problem (dimensions of the state and action spaces) and the horizon (at least for problems without discounting). While the curse of dimensionality is known to be unavoidable in the case of general continuous state/action spaces, the possibility of beating the curse of the horizon remains an open issue, see [32] for recent results and discussions.

A natural performance metric is given by the value function V^π which is the expected total reward of the agent following π. Unfortunately, even a precise knowledge of V^π does not provide reliable information on how far is the policy π from the optimal one. To address this issue a popular quality measure is the *regret* of the algorithm which is the difference between the total sum of rewards accumulated when following the optimal policy and the sum of rewards obtained when following the current policy π. In the setting of finite state- and action space MDPs, there is a variety of regret bounds for popular Reinforcement Learning (RL) algorithms like Q-learning [20], optimistic value iteration [3], and many others. Unfortunately, regret bounds beyond the discrete setup are much less common in the literature. Even more crucial drawback of the regret-based comparison is that regret bounds are typically pessimistic and rely on the unknown quantities of the underlying MDP's such as maximal value of rewards or smoothness of the action-value function functions . A simpler, but related, quantity is the *suboptimality gap (policy error)* $\Delta_\pi(x) := V^\star(x) - V^\pi(x)$. Since we do not know $V^\star$, the suboptimality gap can not be calculated directly. There is a vast amount of literature devoted to theoretical guarantees for $\Delta_\pi(x)$, see e.g. [2], [29], [21] and references therein. However, these bounds share the same drawbacks as the regret bounds. Moreover, known bounds do not apply to the general policy π and depend heavily on the particular algorithm which produced it. For instance, in Approximate

2010 *Mathematics Subject Classification.* 90C40 and 65C05 and 62G08.

Key words and phrases. Markov decision processes, Reinforcement learning, dual representation, pseudo regression, Stein control functionals.

J.S. gratefully acknowledges financial support from the German science foundation (DFG) via the cluster of excellence MATH+, project AA4-2.

Policy Iteration (API, [12]) all existing bounds for $\Delta_\pi(x)$ depend on the one-step error induced by the approximation of the action-value function. This one-step error is difficult to quantify since it depends on the unknown smoothness properties of the action-value function. Similarly, in policy gradient methods (see e.g. [28]), there is always an approximation error due to the choice of the family of policies that can be hardly quantified.

Methods devoted to constructing (sub-)optimal policies for optimal control or optimal stopping times in the case of standard or multiple optimal stopping, are usually termed *primal*. Such methods typically result in a lower biased estimate of the corresponding MDP, i.e., a lower bound in mean on the optimal expected reward.[1] The accuracy of a suboptimal policy obtained via a primal method, is generally not known, however. Using an upper biased estimate, i.e. an upper bound in mean, due to a *dual* approach makes it possible to decide whether or not the suboptimal policy is "tight" in the sense that the gap between the upper and lower estimates is "small". Hence the lack of theoretical guarantees on a suboptimal policy can be addressed by providing a dual bound, that is, an upper bound (or lower bound) on the optimal expected reward (or cost). The last decades have seen a high development of duality approaches for optimal stopping and control problems, initiated by the works of [25] and [19] in the context of pricing of American and Bermudan options. Essentially, in the dual approach one minimizes a certain *dual martingale representation* corresponding to the problem under consideration, for instance a single or multiple stopping problem, an MDP, or a more general control problem, over a set of martingales or martingale type elements. In general terms, the dual version of an optimal control problem $V_0^* = \sup_\alpha \mathbb{E}[R(\alpha)]$ for a reward R depending on adapted policies α may be formulated as

$$V_0^* = \inf_{\text{martingales } M(\boldsymbol{a})} \mathbb{E}[\sup_{\boldsymbol{a} \text{ in control space}} (R(\boldsymbol{a}) - M(\boldsymbol{a}))].$$

Thus, in the dual approach one seeks for optimal martingales rather than optimal policies. For optimal stopping problems, [1] showed how to compute martingales using stopping rules via nested Monte Carlo simulations. In [24], the dual representation for optimal stopping (hence American options) was generalized to Markovian control problems. Somewhat later [14] presented a dual representation for quite general control problems in terms of the so-called information relaxation and martingale penalties. On the other hand, the dual representation for optimal stopping was generalized to multiple stopping in [26] and [11]. As a numerical approach to [24], [10] applied regression methods to solve Markov decision problems that can be seen, in a sense, as a generalization of [1]. However, it should be noted that in the convergence analysis of [10], the primal value function estimates showed exponential dependence on the time horizon, and the corresponding dual algorithm was based on nested simulations while its convergence was not analyzed there. Generally speaking, to the best of our knowledge, all error bounds for the primal/dual value function estimates available in the literature so far show exponential dependence on the horizon at least in the case of finite horizon undiscounted optimal control problems, e.g. see also [31].

In this paper, we propose a novel approach towards constructing valid dual upper bounds on the optimal value function via simulations and pseudo regression in the case of finite horizon MDPs with general (possibly continuous) state and action spaces. This approach includes the construction of primal value functions via a backwardly structured pseudo regression procedure based on a properly chosen reference distribution (measure). In contrast to standard regression, where the conditional expectation $\mathbb{E}[Y|X]$ is estimated using a simulation of pairs (X,Y), in pseudo-regression one simulates X according to a suitably chosen reference measure, and then simulates Y given each realization of X. As a result, in the estimate for $\mathbb{E}[Y|X]$ one may use the explicitly known covariance matrix of X due to the reference measure, and so avoid the delicate problem of inverting the empirical covariance matrix of X. In particular, pseudo regression is carried out with respect to the state variable to approximate conditional expectations of the value function (or its estimate). Let us note that in the context of optimal stopping, a similar primal procedure was proposed in [6], though with accuracy estimates exploding with the number of exercise dates or time horizon. As for the dual part of our algorithm, we avoid nested Monte

[1] Problems of minimizing expected costs can be easily recast to our setting by changing signs.

Carlo simulation that were used in many dual-type methods proposed in the literature so far, see for instance the path-wise optimization approach for MDPs in [15] and [13] for an overview. Instead, for constructing the dual martingale increments, we propose to combine a point-wise pseudo regression approach for estimating the martingale parts of the primal estimates, involving a linear system of elementary martingale functions, with a suitable interpolation method such that the martingale property is preserved. Furthermore, we provide a rigorous convergence analysis showing that the error of approximating the true value function via the estimated dual value function (duality gap) depends at most polynomially on the time horizon. Moreover, we show that the stochastic part of the error depends sublinearly on the state's dimension (or cardinality in the finite case) and action spaces. Let us also mention the work [33] for another approach to avoid nested simulations when estimating the conditional expectations, hence the martingale functions, inside the dual representation. However, [33] left the issue of bounding the duality gap in terms of the error bounds on the primal value functions as an open problem. In this respect, we have solved this problem within the context of the algorithm proposed in this paper.

The paper is organized as follows. The basic setup of the Markov Decision Process and the well-known representations for its maximal expected reward is given in Section 2. Section 3 recalls the dual representation for an MDP from the literature. The primal pseudo regression algorithm for the value functions is described in Section 4, whereas the dual regression algorithm is presented in Section 5. Section 6 and Section 7 are dedicated to the convergence analysis of the primal and dual algorithm, respectively. In Section 8, we illustrate the primal-dual approach proposed above in the context of a particular MDP framework with infinite state and action space, which is popular in various practical applications. Appendix A introduces some auxiliary notions needed to formulate an auxiliary result in Appendix B stemming from the theory of empirical processes.

2. Setup and basic properties of the Markov Decision Process

We consider discrete time finite horizon Markov Decision Process (MDP), given by the tuple

$$\mathcal{M} = (\mathsf{S}, \mathsf{A}, (P_h)_{h\in]H]}, (R_h)_{h\in[H[}, F, H),$$

made up by the following items:

- a measurable state space $(\mathsf{S}, \mathcal{S})$;
- a measurable action space $(\mathsf{A}, \mathcal{A})$;
- an integer H which defines the horizon of the problem;
- for each $h \in]H]$, with $]H] := \{1, \dots, H\}$[2], a time dependent transition function $P_h : \mathsf{S} \times \mathsf{A} \to \mathcal{P}(\mathsf{S})$ where $\mathcal{P}(\mathsf{S})$ is the space of probability measures on $(\mathsf{S}, \mathcal{S})$;
- a time dependent reward function $R_h : \mathsf{S}\times\mathsf{A} \to \mathbb{R}$, where $R_h(x, a)$ is the immediate reward associated with taking action $a \in \mathsf{A}$ in state $x \in \mathsf{S}$ at time step $h \in [H[$;
- a terminal reward $F : \mathsf{S} \to \mathbb{R}$.

Let $\mathfrak{S} := \big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[H]}, \mathbb{P}\big)$ be a filtered probability space. For a fixed policy $\boldsymbol{\pi} = (\pi_0, \dots, \pi_{H-1})$ with $\pi_t : \mathsf{S} \to \mathcal{P}(\mathsf{A})$, we consider an adapted controlled process $\big(S_t, A_t\big)_{t=0,\dots,H}$ on $\mathfrak{S}$ satisfying $S_0 \in \mathsf{S}$, $A_0 \sim \pi_0(S_0)$, and

$$S_{t+1} \sim P_{t+1}(\cdot\,|\, S_t, A_t), \quad A_t \sim \pi_t(S_t), \quad t = 0, \dots, H-1.$$

Assumption 1. In the sequel, we shall assume that the controlled chain (S_t, A_t) comes from a system of the so-called *random iterative functions*:

$$S_t = \mathcal{K}_t(S_{t-1}, A_{t-1}, \varepsilon_t) \sim P_t(\cdot\,|\, S_{t-1}, A_{t-1}), \quad t \in]H],$$

where $\mathcal{K}_t : \mathsf{S}\times\mathsf{A}\times\mathsf{E} \to \mathsf{S}$ is a measurable map with E being a measurable space, and $(\varepsilon_t, t \in]H])$ is an adapted i.i.d. sequence of E-valued random variables on $\mathfrak{S}$ with distribution $\mathcal{P}_{\mathsf{E}}$. Furthermore, ε_t is assumed to be independent of $\mathcal{F}_{t-1}$ for $t \in]H]$.

The expected reward of this MDP due to the chosen policy $\boldsymbol{\pi}$ is given by

$$V_0^{\boldsymbol{\pi}}(x) := \mathbb{E}_{\boldsymbol{\pi},x}\left[\sum_{t=0}^{H-1} R_t(S_t, A_t) + F(S_H)\right], \quad x \in \mathsf{S}$$

[2]We further write $[H] := \{0, 1, \dots, H\}$ etc.

where $\mathbb{E}_{\boldsymbol{\pi},x}$ stands for expectation induced by the policy $\boldsymbol{\pi}$ and transition kernels P_t, $t \in [H]$, conditional on the event $S_0 = x$. The goal of the Markov decision problem is to determine the maximal expected reward:

$$V_0^\star := \sup_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi},x_0}\left[\sum_{t=0}^{H-1} R_t(S_t, A_t) + F(S_H)\right] = \sup_{\boldsymbol{\pi}} V_0^{\boldsymbol{\pi}}(x_0). \tag{2.1}$$

Let us introduce for a generic time $h \in [H]$, the value function due to the policy $\boldsymbol{\pi}$,

$$V_h^{\boldsymbol{\pi}}(x) := \mathbb{E}_{\boldsymbol{\pi},x}\left[\sum_{t=h}^{H-1} R_t(S_t, A_t) + F(S_H)\middle| S_h = x\right], \quad x \in \mathsf{S}.$$

Furthermore, let

$$V_h^\star(x) := \sup_{\boldsymbol{\pi}} V_h^{\boldsymbol{\pi}}(x) \tag{2.2}$$

be the optimal value function at $h \in [H]$. It is well known that under weak conditions, there exists an optimal policy solving (2.2) which depends on S_t in a deterministic way. In this case, we shall write $\boldsymbol{\pi}^\star = (\pi_t^\star(S_t))$ for some mappings $\pi_t^\star : \mathsf{S} \to \mathsf{A}$. One has the following result, see [22].

Theorem 1. *Let $x \in \mathsf{S}$ be fixed. It holds $V_H^\star(x) = F(x)$, and*

$$V_h^\star(x) = \sup_{a\in A}\left(R_h(x,a) + \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V_{h+1}^\star(S_{h+1})\right]\right), \quad h = H-1,\dots,0. \tag{2.3}$$

Furthermore, if R_h is continuous and the action space is compact, the supremum in (2.3) is attained at some deterministic optimal action $a^\star = \pi_h^\star(x)$.

Let us further introduce recursively $Q_H^\star(x,a) = F(x)$ and

$$Q_h^\star(x,a) := R_h(x,a) + \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[\sup_{a'\in A} Q_{h+1}^\star(S_{h+1}, a')\right], \quad h = H-1,\dots,0.$$

Then $Q_h^\star(x,a)$ is called the *optimal state-action* function (Q-function) and one thus has

$$V_h^\star(x) = \sup_{a\in A} Q_h^\star(x,a), \quad \pi_h^\star(x) \in \arg\max_{a\in\mathsf{A}} Q_h^\star(x,a), \quad \text{for} \quad h \in [H].$$

Finally, note that the optimal value function $V^\star$ satisfies due to Theorem 1,

$$V_h^\star(x) = T_h V_{h+1}^\star(x), \quad h \in [H[,$$

where $T_h V(x) := \sup_{a\in A}\left(R_h(x,a) + P_{h+1}^a V(x)\right)$ with $P_{h+1}^a V(x) := \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V(S_{h+1})\right]$.

3. Dual representation

Loosely speaking, the aim of the dual approach is, instead of maximizing over policies as in the primal approach (2.1), to minimize a particular dual stochastic representation over a family of the so-called martingale penalties. Due to its very nature, the dual approach allows for computing upper bounds on the value function, in contrast to lower bounds due to a suboptimal policy obtained by the primal method. A popular interpretation of the dual representation presented below is that one considers the expected maximal reward in a perfect foresight penalized with a possibly "optimal" control-dependent martingale. As such, the penalization of the reward by the martingale corrects for ideal foresight.

Let us denote by $a_{<t}$ the deterministic vector of actions $a_{<t} = (a_0,\dots,a_{t-1}) \in \mathsf{A}^t$, similarly $a_{\le t}$ etc., and denote by $S_t \equiv (S_t(a_{<t}))_{t\in\{0,\dots,H\}}$ the process defined by

$$S_0 = x, \quad S_{t+1} \equiv S_{t+1}(a_{<t+1}) := \mathcal{K}_t(S_t(a_{<t}), a_t, \varepsilon_{t+1}) \sim P_{t+1}(\cdot|S_t(a_{<t}), a_t), \tag{3.1}$$

$t = 0,\dots,H-1$. Let us also denote by Ξ the class of H-tuples $\boldsymbol{\xi} = (\xi_t(\cdot,\cdot),\, t \in]H])$ consisting of $\mathcal{A}^{\otimes t} \times \mathcal{F}_t$ measurable real valued random variables

$$\mathsf{A}^t \times \Omega \ni (a_{<t}, \omega) \to \xi_t(a_{<t}, \varepsilon_{\le t}) \equiv \xi_t(a_{<t}, \varepsilon_{\le t}(\omega)),$$

satisfying

$$\mathbb{E}\left[\xi_t(a_{<t}, \varepsilon_{\le t})\middle|\, \mathcal{F}_{t-1}\right] = 0, \quad \text{for all } a_{<t} \in \mathsf{A}^t, \quad t \in \{1,\dots,H\}. \tag{3.2}$$

The next duality theorem, essentially due to [24], may be seen as a generalization of the dual representation theorem for optimal stopping, developed independently in [25] and [19], to Markov decision processes. For a more general dual representations in terms of information relaxation, see [14]. Let us further mention dual representations in the context of multiple stopping developed in [26], [8], and applications to flexible caps studied in [4].

Theorem 2. *The following statements hold.*
(i) *For any $\boldsymbol{\xi} \in \Xi$ and any $x \in \mathsf{S}$, we have $V_0^{\mathrm{up}}(x;\boldsymbol{\xi}) \geq V_0^\star(x)$ with*

$$V_0^{\mathrm{up}}(x;\boldsymbol{\xi}) := \mathbb{E}_x\left[\sup_{a_{\geq 0}\in \mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \xi_{t+1}(a_{<t+1})\right) + F(S_H(a_{<H}))\right)\right] \tag{3.3}$$

where under the expectation the dependence on $\varepsilon_{\leq H} \equiv \varepsilon_{\leq H}(\omega)$ is suppressed for notational simplicity. Hence $V_0^{\mathrm{up}}(x;\boldsymbol{\xi})$ is an upper (upper-biased) bound for $V_0^\star(x)$.
(ii) *If we set $\boldsymbol{\xi}^\star = (\xi_t^\star,\, t \in [H]) \in \Xi$ with*

$$\begin{aligned}\xi_{t+1}^\star(a_{<t+1}) &:= V_{t+1}^\star(S_{t+1}(a_{<t+1})) - \mathbb{E}_{\mathcal{F}_t}\left[V_{t+1}^\star(S_{t+1}(a_{<t+1}))\right] \\ &= V_{t+1}^\star(S_{t+1}(a_{<t+1})) - \mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t(a_{<t}),a_t)}\left[V_{t+1}^\star(S'_{t+1})\right]\end{aligned} \tag{3.4}$$

by (3.1), for $t = 0,\ldots,H-1$, then, almost surely,

$$V_0^\star(x) = \sup_{a_{\geq 0}\in \mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \xi_{t+1}^\star(a_{<t+1})\right) + F(S_H(a_{<H}))\right). \tag{3.5}$$

Remark 3.

- In Theorem 2 and further below, supremum should be interpreted as essential supremum in case it concerns the supremum over an uncountable family of random variables.
- The random variables (ξ_t) are called martingale increments or martingale functions due to the property (3.2). They are the building blocks for the controlled "dual martingale penalty" $\sum_{t=0}^{H-1}\xi_{t+1}(a_{<t+1})$.

In principle, Theorem 2 may be inferred from [24, Theorem 3] or [14, Theorem 2.1]. Nonetheless, also for the convenience of the reader, we here give a concise proof in terms of the present notation and terminology.

Proof. (i) Since for any $\boldsymbol{\xi} \in \Xi$ and policy $\boldsymbol{\pi}$ in (2.1) one has that

$$\mathbb{E}_{\boldsymbol{\pi},x}\left[\xi_{t+1}(A_{\leq t})\right] = \mathbb{E}_{\boldsymbol{\pi},x}\mathbb{E}_{\boldsymbol{\pi}}\left[\xi_{t+1}(A_{\leq t})\middle|\,\mathcal{F}_t\right] = 0,$$

for $t = h, ..., H-1$, it follows that

$$V_0^\star(x) = \sup_{\boldsymbol{\pi}}\mathbb{E}_{\boldsymbol{\pi},x}\left[\sum_{t=0}^{H-1}\left(R_t(S_t(A_{<t}),A_t) - \xi_{t+1}(A_{\leq t})\right) + F(S_H(A_{<H}))\right],$$

from which (3.3) follows immediately.

(ii) Due to (3.4) we may write for any $a_{\geq 0} \in \mathsf{A}^H$,

$$\begin{aligned}
&\sum_{t=0}^{H-1} \left(R_t(S_t\left(a_{<t}\right), a_t) - \xi^\star_{t+1}(a_{\leq t})\right) + F(S_H(a_{<H})) \\
&= \sum_{t=0}^{H-1} R_t(S_t\left(a_{<t}\right), a_t) - \sum_{t=0}^{H-1} V^\star_{t+1}(S_{t+1}(a_{\leq t})) \\
&+ \sum_{t=0}^{H-1} \mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t(a_{<t}),a_t)} \left[V^\star_{t+1}(S'_{t+1})\right] + F(S_H(a_{<H})) \\
&= \sum_{t=0}^{H-1} \Big(R_t(S_t\left(a_{<t}\right), a_t) + \mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t(a_{<t}),a_t)} \left[V^\star_{t+1}(S'_{t+1})\right] - V^\star_t(S_t(a_{<t}))\Big) \\
&+ F(S_H(a_{<H})) - V^\star_H(S_H(a_{<H})) + V^\star_0(S_0(a_{<0})) \\
&\leq V^\star_0(S_0(a_{<0})) = V^\star_0(x),
\end{aligned}$$

where the latter inequality follows from the Bellman principle, see Theorem 1. The statement (3.5) now follows by taking the supremum over $a_{\geq 0} \in \mathsf{A}^H$ on the left-hand-side, then taking the expectation, next applying (3.3), and finally using the sandwich property. □

In the primal approach, one typically constructs a sequence of approximate (continuation) value functions. In practice, however, assessing the quality of these estimated value functions for a problem at hand is difficult. The same is true for the lower biased estimate simulated using the approximated policy. The very benefit of the dual approach is that one may construct an upper-biased estimate based on the policy constructed by the primal method. Then the gap between the upper and lower bounds can be used to assess the quality of the primal policy.

4. Primal regression algorithm for the value function

In Section 5, we will describe regression based martingale methods for computing dual upper bounds based on Theorem 2. However, these methods require as an input a sequence of (approximate) value functions V_h, $h \in [H]$. Below we describe a regression-based algorithm for approximating the value functions $V^\star_h$, $h \in [H]$, backwardly in time. In fact, in contrast to usual regression, the proposed algorithm is based on a kind of "pseudo" or "quasi" regression procedure due to N drawings from a measure $\mu_h(dx)$ at each time $h \in [H[$, where μ_h is the so-called reference measure. Furthermore, we consider a vector of basis functions

$$\boldsymbol{\gamma}_K := (\gamma_1, \ldots, \gamma_K)^\top, \quad \gamma_k : \mathsf{S} \to \mathbb{R}, \quad k = 1, \ldots, K,$$

such that the matrix

$$\Sigma \equiv \Sigma_{h,K} := \mathbb{E}_{X\sim\mu_h} \left[\boldsymbol{\gamma}_K(X)\boldsymbol{\gamma}_K^\top(X)\right] \tag{4.1}$$

is analytically known and invertible. This basically means that the choice of basis functions is adapted to the choice of the reference measure. For example, if μ_h is Gaussian one can chose basis functions to be polynomials or trigonometric polynomials. [3] The algorithm reads then as follows. At $h = H$ we set $V_{H,N}(x) = V^\star_H(x) = F(x)$. Suppose that for some $h \in [H[$, the approximations $V_{t,N}$ of $V^\star_t$, $h+1 \leq t \leq H$, are already obtained. We next simulate, in view of Assumption 1, independent random variables $X^h_i \sim \mu_h$, $\varepsilon^i_{h+1} \sim \mathcal{P}_{\mathsf{E}}$, $i = 1, \ldots, N$, and set

$$Y^a_i = \mathcal{K}_{h+1}(X^h_i, a, \varepsilon^i_{h+1}) \sim P_{h+1}(\cdot|X^h_i, a), \qquad a \in \mathsf{A}.$$

We underline that we can use the same drawings of the random variables X^h and ε_{h+1} to sample Y^a for different a. That is, we only need to recalculate the mapping $\mathcal{K}_{h+1}$.

Next approximate $V^\star_h$ via

$$V_{h,N}(x) = T_{h,N}V_{h+1,N}(x) := \sup_{a\in\mathsf{A}}(R_h(x,a) + \widetilde{P}^a_{h+1,N}V_{h+1,N}(x)), \tag{4.2}$$

[3] By taking $\boldsymbol{\gamma}_K$ depending on h one may even achieve that Σ is the identity for each h.

where

$$\widetilde{P}^a_{h+1,N}V(x) := \mathcal{T}_{\widetilde{L}_{h+1}}[\beta_{N,a}^\top \boldsymbol{\gamma}_K](x) := \max\big(-\widetilde{L}_{h+1}, \min\big(\widetilde{L}_{h+1}, \beta_{N,a}^\top \boldsymbol{\gamma}_K(x)\big)\big) \tag{4.3}$$

with $\widetilde{L}_{h+1}$ being a positive constant depending on h, which will be defined later, and

$$\beta_{N,a} := \frac{1}{N}\sum_{i=1}^N U_i^a, \quad U_i^a := Z_i^a \Sigma^{-1}\boldsymbol{\gamma}_K(X_i^h), \quad Z_i^a := V(Y_i^a), \quad i = 1, \dots, N. \tag{4.4}$$

Note that $\beta_a := \mathbb{E}\left[\beta_{N,a}\right] = \mathbb{E}\left[V(Y_1^a)\Sigma^{-1}\boldsymbol{\gamma}_K(X_1^h)\right]$ solves the minimization problem

$$\inf_{\beta\in\mathbb{R}^K} \mathbb{E}\left[\Big(V(Y_1^a) - \beta^\top \boldsymbol{\gamma}_K(X_1^h)\Big)^2\right].$$

Thus, the quantity $\widetilde{P}^a_{h+1,N}V_{h+1,N}(x)$ aims to approximate the conditional expectation

$$x \to \mathbb{E}_{S'\sim P_{h+1}(\cdot|x,a)}\left[V_{h+1,N}(S')\right], \quad a \in \mathsf{A}.$$

The use of clipping at level $\widetilde{L}_{h+1}$ is done to avoid large values of $\beta_{N,a}^\top \boldsymbol{\gamma}_K(x)$. Note that the approximate state-action function

$$Q_{h,N}(x,a) = R_h(x,a) + \widetilde{P}^a_{h+1,N}V_{h+1,N}(x),$$

due to (4.3) and (4.4), determines a greedy policy solving (4.2),

$$a_h(x) \in \operatorname*{argsup}_{a\in\mathsf{A}} Q_{h,N}(x,a),$$

so that $V_{h,N}(x) = Q_{h,N}(x, a_h(x))$. After H steps of the above procedure we obtain the estimates $V_{H,N}, \dots, V_{0,N}$.[4] The greedy policies $a_h(\cdot)$, $h = H-1, \dots, 0$, may be utilized for computing a lower biased estimate for $V_0^\star(x_0)$,

$$V_0^{\mathrm{low}}(x_0) := \frac{1}{N_{\mathrm{test}}}\sum_{n=1}^{N_{\mathrm{test}}}\sum_{t=0}^{H-1} R_t(S_t^{(n)}, a_t(S_t^{(n)})) + F(S_H^{(m)}) \quad \text{with} \tag{4.5}$$

$$S_{t+1}^{(n)} = \mathcal{K}(S_t^{(n)}, a_t(S_t^{(n)}), \widetilde{\varepsilon}_{t+1}^{(n)}), \qquad S_0^{(n)} = x_0,$$

using an independent sequence $\widetilde{\varepsilon}_t^{(n)}$, $t \in [H[$, $n = 1, \dots, N_{\mathrm{test}}$. In fact, the main advantage of the dual approach is that it allows one to assess the quality of the lower estimate (4.5) by constructing an upper biased estimate based on $V_{h,N}$, $h \in [H]$, see the next section.

Note that the above regression algorithm and its convergence analysis in Section 6 are similar in spirit to the least-squares regression algorithms for optimal stopping problems and BSDEs, see [27], [16], [18] and [17]. In the latter work, the authors proposed a novel method of stratification. This method involves efficient storage of simulations and minimizes memory requirements. In our pseudo-regression algorithm, we avoid the inversion of the empirical covariance matrix, thus significantly reducing the computational cost.

5. Dual regression algorithm

In this section, we outline how to construct an upper biased estimate based on Theorem 2 from a given sequence of approximations V_t, $t \in [H]$ obtained, for example, as described in Section 4.

Theorem 2-(ii) implies that we can restrict our attention to processes $\boldsymbol{\xi} = (\xi_t)_{t\in[H]}$ where the $t+1$ component of $\boldsymbol{\xi}$ is of the form

$$\xi_{t+1}(a_{\le t}) = m(S_{t+1}(a_{\le t}); S_t(a_{<t}), a_t) \tag{5.1}$$

for a deterministic real valued function $m(\cdot; x, a)$ satisfying

$$\int m(y; x, a) P_{t+1}(dy|x, a) = 0, \tag{5.2}$$

[4] Actually, for computing $V_0(x_0)$ we may replace the above procedure by a standard Monte Carlo simulation when going from V_1 to V_0.

for all $(x,a) \in \mathsf{S}\times\mathsf{A}$. Note that the condition (5.2) is time dependent. We shall denote by $\mathcal{M}_{t+1,x,a}$ the set of "martingale" functions m on S that satisfy (5.2) for time $t+1$, a state x, and a control a. In this section, we develop an algorithm approximating $\boldsymbol{\xi}^\star$ via regression of V_{t+1} on a properly chosen finite dimensional subspace of $\mathcal{M}_{t+1,x,a}$. The idea of approximating $\boldsymbol{\xi}^\star$ via regression can be explained as follows. Equation (3.4) and (5.1) imply that, for a particular $t \in [H[$, the component $\xi^\star_{t+1}(a_{\le t})$ of the random vector $\boldsymbol{\xi}^\star$ is given by $\xi^\star_{t+1}(a_{\le t}) = m^\star_{t+1}(S_{t+1}(a_{\le t}); S_t(a_{<t}), a_t)$, where, for each $(x,a) \in \mathsf{S}\times\mathsf{A}$, $m^\star_{t+1}(\cdot;x,a)$ solves the optimization problem

$$\operatorname*{arginf}_{m\in\mathcal{M}_{t+1,x,a}} \mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|x,a)}\left[\left(V^\star_{t+1}(S'_{t+1}) - m(S'_{t+1};x,a)\right)^2\right] = \operatorname*{arginf}_{m\in\mathcal{M}_{t+1,x,a}} \mathrm{Var}_{S'_{t+1}\sim P_{t+1}(\cdot|x,a)}\left[V^\star_{t+1}(S'_{t+1}) - m(S'_{t+1};x,a)\right]. \tag{5.3}$$

By generating a sample $Y_1^{x,a},\dots,Y_N^{x,a}$ from $P_{t+1}(\cdot|x,a)$ we readily obtain a computable approximation of $m^\star_{t+1}(\cdot;x,a)$, that is, the solution of (5.3), by

$$\operatorname*{arginf}_{m\in\mathcal{M}'_{t+1,x,a}} \left\{\frac{1}{N}\sum_{i=1}^N \left(V_{t+1}(Y_i^{x,a}) - m(Y_i^{x,a})\right)^2\right\} \tag{5.4}$$

where $\mathcal{M}'_{t+1,x,a}$ is some "large enough" finite-dimensional subset of $\mathcal{M}_{t+1,x,a}$.

Let us now discuss possible constructions of the martingale functions m satisfying (5.2). Assume that $\mathsf{S} \subseteq \mathbb{R}^d$ and that the conditional distribution $P_{t+1}(\cdot|x,a)$ possesses a smooth density $p_{t+1}(\cdot|x,a)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. Furthermore, assume that $p_{t+1}(\cdot|x,a)$ does not vanish on any compact set in $\mathbb{R}^d$, and that $p_{t+1}(y|x,a) \to 0$ for $|y| \to \infty$. Now consider, for any fixed (x,a), functions of the form (*Stein* control functionals)

$$m_{t+1,\phi}(\cdot;x,a) := \langle \nabla \log(p_{t+1}(\cdot|x,a)), \phi\rangle + \mathrm{div}(\phi)$$

with $\phi : \mathsf{S} \to \mathbb{R}^d$ being a smooth and bounded mapping with bounded derivatives. It is then not difficult to check that

$$\int_{\mathsf{S}} p_{t+1}(y|x,a)\phi_i(y)\partial_{y_i}\log(p_{t+1}(y|x,a))\,dy = -\int_{\mathsf{S}} p_{t+1}(y|x,a)\partial_{y_i}\phi_i(y)\,dy, \quad i=1,\dots,d,$$

and hence $m_{t+1,\phi}$ satisfies (5.2) for all $(x,a) \in \mathsf{S}\times\mathsf{A}$. This means that in (5.4), we can take $\mathcal{M}'_{t+1,x,a} = \{m_{t+1,\phi}(\cdot;x,a) : \phi \in \Phi\}$ where Φ is the linear space of mappings $\mathbb{R}^d \to \mathbb{R}^d$, which are smooth, bounded, and have bounded derivatives. Since $\phi \to m_{t+1,\phi}(\cdot;x,a)$ is linear in ϕ we moreover have that $\mathcal{M}'_{t+1,x,a}$ is a linear space of real valued functions. So the problem (5.4) can be casted into a standard linear regression problem after choosing a system of basis functions $(m_{t+1,\varphi_k}(\cdot;x,a))_{k\in\mathbb{N}}$ due to some basis $(\varphi_k)_{k\in\mathbb{N}}$ in Φ. Needles to say that the problem (5.4) can only be solved on some finite grid, $(x_l,a_l)_{l=1,\dots,L} \in \mathsf{S}\times\mathsf{A}$ say, yielding solutions $\phi_k(\cdot) := \phi(\cdot;x_k,a_k)$ and the corresponding martingale functions $m_{t+1,\phi_k}(\cdot;x_k,a_k)$. In order to obtain a martingale function $m_{t+1} \equiv m_{t+1}(\cdot;x,a)$ for a generic pair (x,a), we may apply some suitable interpolation procedure. Loosely speaking, if (x,a) is an interpolation between (x_k,a_k) and $(x_{k'},a_{k'})$ we may interpolate $\phi(\cdot;x,a)$ between ϕ_k and $\phi_{k'}$ correspondingly, and set $m_{t+1} = m_{t+1,\phi}(\cdot;x,a)$. For details regarding suitable interpolation procedures, we refer to Section 7.

Let now, for each $t \in [H[$, and $(x,a) \in \mathsf{S}\times\mathsf{A}$, the martingale function $m_{t+1}(\cdot;x,a)$ be an approximate solution of (5.4). Then we can construct an upper bound (upper biased estimate) for $V_0^\star(x_0)$, via a standard Monte Carlo estimate of the expectation

$$V_0^{\mathrm{up}}(x) = \mathbb{E}_{\boldsymbol{\pi},x}\left[\sup_{a_{\ge 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{\ge t}),a_t) - m_{t+1}(S_{t+1}(a_{\le t});S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right]. \tag{5.5}$$

Another way of constructing $\boldsymbol{\xi} \in \Xi$ is based on Assumption 1. Let us assume that $(\psi_k,\, k\in\mathbb{N}_0)$ is a system in $L^2(\mathsf{E},\mathcal{P}_\mathsf{E})$ satisfying

$$\int \psi_k(\varepsilon)\,d\mathcal{P}_\mathsf{E} = 0, \quad k \in \mathbb{N}.$$

By then letting

$$\eta_{t+1,K}(x,a) \equiv \eta_{t+1,K}(x,a,\varepsilon_{t+1}) = \sum_{k=1}^{K} c_k(x,a)\psi_k(\varepsilon_{t+1}) \tag{5.6}$$

for some natural $K > 0$ and "nice" functions $c_k : \mathsf{S} \times \mathsf{A} \to \mathbb{R}$, $k = 1, \dots, K$, we have that

$$\xi_{t+1,K}(a_{\leq t}) := \eta_{t+1,K}(S_t(a_{<t}), a_t)$$

is $\mathcal{F}_{t+1}$-measurable, and, since $\int \psi_k(\varepsilon)\, d\mathcal{P}_{\mathsf{E}}(\varepsilon) = 0$ for $k \in \mathbb{N}$, it holds that $\mathbb{E}\left[\xi_{t+1,K}(a_{\leq t}) \middle| \mathcal{F}_t\right] = 0$. Hence, we have that $\boldsymbol{\xi}_K = (\xi_{t+1,K}(a_{\leq t}), t \in [H[) \in \Xi$. In this case, we can consider the least-squares problem

$$\inf_{(c_1,\dots,c_K)} \mathbb{E}\left[\left(V_{t+1}(Z^{x,a}) - \sum_{k=1}^{K} c_k\psi_k(\varepsilon_{t+1})\right)^2\right], \quad Z^{x,a} \equiv \mathcal{K}_{t+1}(x,a,\varepsilon_{t+1}), \tag{5.7}$$

for estimating the coefficients in (5.6). Let us further denote $\Sigma_{\mathsf{E},K} := \mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}\left[\boldsymbol{\psi}_K(\varepsilon)\boldsymbol{\psi}_K^{\top}(\varepsilon)\right]$ with $\boldsymbol{\psi}_K(\varepsilon) := [\psi_1(\varepsilon), \dots, \psi_K(\varepsilon)]^{\top}$. The minimization problem (5.7) is then explicitly solved by

$$\bar{\mathbf{c}}_K(x,a) := \Sigma_{\mathsf{E},K}^{-1}\mathbb{E}\left[V_{t+1}(Z^{x,a})\boldsymbol{\psi}_K(\varepsilon)\right]. \tag{5.8}$$

In the sequel we assume that $\Sigma_{\mathsf{E},K}$ is known and invertible. This assumption is not particularly restrictive, as we choose the basis $\boldsymbol{\psi}$ ourselves. In order to compute (5.8), we can construct a new sample $U_m(x,a) = V_{t+1}(Z_m^{x,a})\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon_m)$ with $\varepsilon_m \sim \mathcal{P}_{\mathsf{E}}$, $Z_m^{x,a} \equiv \mathcal{K}_{t+1}(x,a,\varepsilon_m)$, $m = 1, \dots, M$, and estimate its mean $\bar{\mathbf{c}}_K(x,a)$ by the empirical mean

$$\mathbf{c}_{K,M}(x,a) = [c_{1,M}(x,a), \dots, c_{K,M}(x,a)]^{\top} := \frac{1}{M}\sum_{m=1}^{M} U_m(x,a). \tag{5.9}$$

We so obtain as martingale functions in (5.6),

$$\eta_{t+1,K,M}(x,a,\varepsilon_{t+1}) := \mathbf{c}_{K,M}^{\top}(x,a)\boldsymbol{\psi}_K(\varepsilon_{t+1}) = \sum_{k=1}^{K} c_{k,M}(x,a)\psi_k(\varepsilon_{t+1}). \tag{5.10}$$

Also note that the problem (5.7) may only numerically be solved on a grid, and a suitable interpolation procedure is required to obtain (5.10) for generic $(x,a) \in \mathsf{S} \times \mathsf{A}$ (for details see Section 7). Finally, an upper biased estimate for $V_0^{\star}(x)$, hence an upper bound in mean, can be obtained via an independent standard Monte Carlo estimate of the expectation

$$V_0^{\mathrm{up}}(x) = \mathbb{E}_{\boldsymbol{\pi},x}\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{\geq t}),a_t) - \eta_{t+1,K,M}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right]. \tag{5.11}$$

In Section 7 we will give a detailed convergence analysis of the dual estimator (5.11). It is anticipated that a similar analysis can be carried out for the dual estimator (5.5), but this analysis is omitted due to space restrictions.

6. Convergence analysis of the primal algorithm

In this section, we carry out the convergence analysis of the primal algorithm designed in Section 4, under some mild assumptions.

Assumption 2. Assume that (1) holds. In this case $P_h^a f(x) = \mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}[f(\mathcal{K}_h(x,a,\varepsilon))]$, $(x,a) \in \mathsf{S}\times\mathsf{A}$. Also assume that the kernels $\mathcal{K}_h$ are Lipschitz continuous:

$$|\mathcal{K}_h(x,a,\varepsilon) - \mathcal{K}_h(x',a',\varepsilon)| \leq L_{\mathcal{K}}\,\rho((x,a),(x',a')), \quad (x,a),(x',a') \in \mathsf{S}\times\mathsf{A}, \quad \varepsilon \in \mathsf{E}, \tag{6.1}$$

for some constant $L_{\mathcal{K}}$ not depending on h. In (6.1), the metric $\rho \equiv \rho_{\mathsf{S}\times\mathsf{A}}$ on $\mathsf{S}\times\mathsf{A}$ is considered to be of the form

$$\rho_{\mathsf{S}\times\mathsf{A}}((x,a),(x',a')) = \left\|\left(\rho_{\mathsf{S}}(x,x'),\rho_{\mathsf{A}}(a,a')\right)\right\|,$$

where ρ_{S} and ρ_{A} are suitable metrics on S and A, respectively, and $\|(\cdot,\cdot)\|$ is a fixed but arbitrary norm on $\mathbb{R}^2$. In order to avoid an overkill of notation, we will henceforth drop the subscripts S, A, and $\mathsf{S}\times\mathsf{A}$, whenever it is clear from the arguments which metric is considered.

Assumption 3. Assume that $\sup_{(x,a)\in \mathsf{S}\times\mathsf{A}}\{|R_h(x,a)| \vee |F(x)|\} \le R_{\max}$ and

$$\sup_{a\in\mathsf{A}} |R_h(x,a) - R_h(x',a)| \le L_R \rho(x,x')$$

for some constants $R_{\max}$ and L_R not depending on $h \in [H[$.

We now set

$$\widetilde{L}_h := (H-h+1)R_{\max}, \quad h\in[H], \quad V^\star_{\max} := \widetilde{L}_0 = (H+1)R_{\max}. \tag{6.2}$$

Assumption 4. Assume that $|\Sigma^{-1}_{h,K}\boldsymbol{\gamma}_K(x)|_\infty \le \Lambda_K$ for all $x\in\mathsf{S}$, $h\in[H[$, and

$$|\boldsymbol{\gamma}_K(x) - \boldsymbol{\gamma}_K(x')| \le L_{\gamma,K}\rho(x,x')$$

for a constant $L_{\gamma,K}>0$, where $|\cdot|$ denotes the Euclidian norm and $|\cdot|_\infty$ stands for the ℓ_∞ norm.

Note that due to (4.2) and (6.2), one has that $|V_{h,N}| \le \widetilde{L}_h$, $h\in[H]$, and that under Assumptions 2, 3, and 4 one has

$$\begin{aligned}
|T_{h,N}V_{h+1,N}(x) - T_{h,N}V_{h+1,N}(x')| &\le L_R\rho(x,x') + \sup_{a\in\mathsf{A}} |\widetilde{P}^a_{h+1,N}V_{h+1,N}(x) - \widetilde{P}^a_{h+1,N}V_{h+1,N}(x')| \\
&\le L_R\rho(x,x') + \sup_{a\in\mathsf{A}} |\beta_{N,a}||\boldsymbol{\gamma}_K(x) - \boldsymbol{\gamma}_K(x')| \\
&\le L_R\rho(x,x') + \frac{1}{N}\sum_{n=1}^N \sup_{a\in\mathsf{A}} |Z^a_n||\Sigma^{-1}_{h,K}\boldsymbol{\gamma}_K(X_n)||\boldsymbol{\gamma}_K(x) - \boldsymbol{\gamma}_K(x')| \\
&\le [L_R + V^\star_{\max}\Lambda_K\sqrt{K}L_{\gamma,K}]\rho(x,x').
\end{aligned}$$

Let us denote $L_{V,K} := L_R + V^\star_{\max}\Lambda_K L_{\gamma,K}\sqrt{K}$. The above estimates imply that $V_{h,N}\in \mathrm{Lip}(L_{V,K})$, and so the function $f(x,a,\varepsilon) := V_{h,N}(\mathcal{K}_h(x,a,\varepsilon))$ satisfies

$$\big|f(x,a,\varepsilon) - f(x',a',\varepsilon)\big| \le L_{V,K}L_{\mathcal{K}}\rho((x,a),(x',a')). \tag{6.3}$$

The next assumption concerns the measures $\mu_0,\dots,\mu_{H-1}$.

Assumption 5. Let us consider for any $0\le h<l<H$, the Radon-Nikodym derivative

$$\mathfrak{R}_{h,l}(x'|x,\boldsymbol{\pi}) := \frac{P^{\pi_h}_{h+1}\dots P^{\pi_{l-1}}_{l}(dx'|x)}{\mu_l(dx')},$$

where we define for a generic policy $\boldsymbol{\pi} = (\pi_0,\dots,\pi_{H-1})$,

$$P^{\pi_h}_{h+1}(dx'|x) := P_{h+1}(dx'|x,\pi_h(x)).$$

Assume that

$$\mathfrak{R}^{\max} := \sup_{0\le h<l<H,\boldsymbol{\pi}} \left(\int \mu_h(dx)\int \mathfrak{R}^2_{h,l}(x'|x,\boldsymbol{\pi})\mu_l(dx')\right)^{1/2} < \infty. \tag{6.4}$$

By the very construction of $V_{h,N}$ from $V_{h+1,N}$, $h\in[H[$, as outlined in Section 4, $V_{h,N}$ may be seen as random (Lipschitz continuous) function. In particular, for each $x\in\mathsf{S}$, $V_{h,N}(x)$ is measurable with respect to the σ-algebra

$$\mathcal{D}^N_h := \sigma\big\{\mathbf{Y}^{h;N},\dots,\mathbf{Y}^{H-1;N}\big\} \quad \text{with} \quad \mathbf{Y}^{h;N} := \big(\big(X^h_1,\varepsilon^1_{h+1}\big),\dots,\big(X^h_N,\varepsilon^N_{h+1}\big)\big), \tag{6.5}$$

where the pairs $\big(X^h_i,\varepsilon^i_{h+1}\big) \sim \mu_h\otimes\mathcal{P}_{\mathsf{E}}$ are i.i.d. for $h\in[H[$, $i=1,\dots,N$, and Monte Carlo simulated under the measure $\mathsf{P}\equiv\mathsf{P}_{\mathsf{N}} := \bigotimes_{h=0}^{H-1}(\mu_h\otimes\mathcal{P}_{\mathsf{E}})^{\otimes N}$. The following theorem provides an upper bound for the difference between $V_{h,N}$ and $V^\star_h$.

Theorem 4. *Suppose that* $\mathbb{E}_{X\sim\mu_h}\big[|\boldsymbol{\gamma}_K(X)|^2\big] \le \varrho^2_{\gamma,K}$ *for all* $h\in[H[$. *Then for* $h\in[H]$,

$$\begin{aligned}
&\|V^\star_h(\cdot) - V_{h,N}(\cdot)\|_{L^2(\mu_h\otimes\mathsf{P})} \\
&\quad\lesssim \mathfrak{R}^{\max}\left((H-h)\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A}) + L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A}) + V^\star_{\max})\sqrt{\frac{K}{N}} + \sum_{l=h}^{H-1}\mathcal{R}_{K,l}\right),
\end{aligned}$$

where $\lesssim$ *denotes* $\le$ *up to an absolute constant,* $I_{\mathcal{D}}(\mathsf{A})$ *is the metric entropy of* A, $\mathsf{D}(\mathsf{A})$ *is the diameter of* A *as defined in Appendix A, and*

$$\mathcal{R}_{K,h} := \sup_{\boldsymbol{\zeta}\in\mathbb{R}^{K\times\mathsf{A}}} \mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}\Big(\beta_{a,\boldsymbol{\zeta}}^\top\boldsymbol{\gamma}_K(X) - P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}(X)\Big)^2\right]^{1/2}, \quad \textit{where}$$

$$\mathcal{V}_{h,\boldsymbol{\zeta}}(x) := \sup_{a\in\mathsf{A}}\big(R_h(x,a) + \mathcal{T}_{\widetilde{L}_{h+1}}[\zeta_a^\top\boldsymbol{\gamma}_K(x)]\big) \ \textit{ for } 0\le h<H, \quad \mathcal{V}_{H,\boldsymbol{\zeta}}(x) := F(x), \quad \textit{and}$$

$$\beta_{a,\boldsymbol{\zeta}} := \operatorname*{arginf}_{\beta\in\mathbb{R}^K} \mathbb{E}_{X\sim\mu_h}\left[\Big(\beta^\top\boldsymbol{\gamma}_K(X) - P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}(X)\Big)^2\right].$$

Discussion.

- The quantity $\mathcal{R}_{K,h}$ is related to the error of approximating the conditional expectation $P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}$ via a linear combination of the basis functions $\gamma_1,\dots,\gamma_K$ in a worst case scenario, that is, for the most unfavorable choice of $\boldsymbol{\zeta}$. Note that this way of measuring the approximation error differs from one typically used in the convergence analysis of the least-squares approaches, e.g., [32]. Usually, one assesses the approximation error based on the smoothness of the actual conditional expectation $P_{h+1}V_{h+1}^\star$. Let us suppose, for illustration, that A is finite and take some $h<H-1$. One then has

$$\mathcal{R}_{K,h} \le \sum_{a\in\mathsf{A}}\sup_{\boldsymbol{\zeta}\in\mathbb{R}^{K\times|\mathsf{A}|}} \mathbb{E}_{X\sim\mu_h}\left[\Big(\beta_{a,\boldsymbol{\zeta}}^\top\boldsymbol{\gamma}_K(X) - P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}(X)\Big)^2\right]^{1/2} \tag{6.6}$$

  where $\beta_{a,\boldsymbol{\zeta}}^\top\boldsymbol{\gamma}_K$ is the $L^2(\mu_h)$ projection of $P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}$ on $\operatorname{span}(\boldsymbol{\gamma}_{\mathrm{K}})$ with the corresponding projection error

$$\mathcal{E}_{K,h}(a,\boldsymbol{\zeta}) := \mathbb{E}_{X\sim\mu_h}\left[\Big(\beta_{a,\boldsymbol{\zeta}}^\top\boldsymbol{\gamma}_K(X) - P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}(X)\Big)^2\right]^{1/2}. \tag{6.7}$$

  Under mild conditions on P_{h+1}^a, (6.7) converges to zero uniformly in $\boldsymbol{\zeta}$, at a rate depending on the choice of $\boldsymbol{\gamma}_K$. For example, if the system $\gamma_1,\gamma_2,\dots$ is an orthonormal base in $L^2(\mu_h)$ then $\max_{a\in\mathsf{A}}\sup_{\boldsymbol{\zeta}\in\mathbb{R}^{K\times\mathsf{A}}}\mathcal{E}_{K,h}(a,\boldsymbol{\zeta}) \lesssim K^{-\beta}$, $\beta>0$, provided that the series

$$\sum_{k=1}^\infty k^{2\beta}\,\mathbb{E}_{X\sim\mu_h}[\gamma_k(X)P_{h+1}^a\mathcal{V}_{h+1,\boldsymbol{\zeta}}(X)]^2$$

  is uniformly bounded in $\boldsymbol{\zeta}\in\mathbb{R}^{K\times\mathsf{A}}$ and $a\in\mathsf{A}$. Hence, then $\mathcal{R}_{K,h}\lesssim|\mathsf{A}|K^{-\beta}\to 0$ for $K\to\infty$. Note that (6.6) is a worst case estimate, which may be very rough in general.
- Suppose that $P_{h+1}^a =: P_{h+1}$ does not depend on $a\in\mathsf{A}$, and that $\gamma_1,\gamma_2,\dots$ are bounded eigenfunctions (corresponding to nonnegative eigenvalues) of P_{h+1}. Let further $F(x)=\beta^\top\boldsymbol{\gamma}_K(x)$ for some $\beta\in\mathbb{R}^K$ and $R_t(x,a)=R_{1,t}(x)R_{2,t}(a)$ with $R_{1,t}(x)=c_t^\top\boldsymbol{\gamma}_K(x)\ge 0$, then for $\widetilde{L}_{h+1}$ large enough, $\mathcal{R}_{K,h}=0$ (in this case we may take ζ_a independent of a in the definition of $\mathcal{V}_{h+1,\boldsymbol{\zeta}}$) and only the stochastic part of the error remains:

$$\|V_h^\star - V_{h,N}\|_{L^2(\mu_h\otimes\mathsf{P})} \lesssim H\mathfrak{R}^{\max}\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A}) + L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A}) + V_{\max}^\star)\sqrt{\frac{K}{N}}. \tag{6.8}$$

- Let us consider the stochastic error (6.8) in more details for an example where $\mathsf{A}=[0,1]^{d_{\mathsf{A}}}$ for some $d_{\mathsf{A}}\in\mathbb{N}$. One then has $\mathsf{D}(\mathsf{A})=\sqrt{d_{\mathsf{A}}}$ and $I_{\mathcal{D}}(\mathsf{A})\lesssim\sqrt{d_{\mathsf{A}}}$. In this example the bound (6.8) depends sub-linearly in d_{A}. If in addition all basis functions (γ_k) are uniformly bounded and the infinity matrix norm (i.e. the maximum absolute row sum) of $\Sigma_{h,K}$ is uniformly bounded from below for all $K\in\mathbb{N}$ and $h\in[H]$, then $\varrho_{\gamma,K}\lesssim K^{1/2}$, $\Lambda_K\lesssim 1$, $L_{V,K}\lesssim L_{\gamma,K}HK^{1/2}$, $V_{\max}^\star\lesssim H$, and the bound in Theorem 4 transforms to

$$\|V_h^\star - V_{h,N}\|_{L^2(\mu_h\otimes\mathsf{P})} \lesssim \frac{(H-h)H\mathfrak{R}^{\max}\sqrt{d_{\mathsf{A}}}L_{\gamma,K}K^{3/2}}{\sqrt{N}} + \mathfrak{R}^{\max}\sum_{l=h}^{H-1}\mathcal{R}_{K,l}, \tag{6.9}$$

where $\lesssim$ means inequality up to a constant not depending on H, N, K and A. Another relevant situation is the case of finite A. Here $I_{\mathcal{D}}(\mathsf{A}) = \sqrt{\log|\mathsf{A}|}$ and $\mathsf{D}(\mathsf{A}) = 1$. Hence (6.9) changes to

$$\|V_h^\star - V_{h,N}\|_{L^2(\mu_h\otimes\mathsf{P})} \lesssim \frac{(H-h)H\mathfrak{R}^{\max}\sqrt{\log|\mathsf{A}|}L_{\gamma,K}K^{3/2}}{\sqrt{N}} + \mathfrak{R}^{\max}\sum_{l=h}^{H-1}\mathcal{R}_{K,l}. \tag{6.10}$$

Let us point out to a logarithmic dependence of (6.10) on $|\mathsf{A}|$.

- Let us remark on Assumption 5 and discuss the quantity $\mathfrak{R}^{\max}$. Consider $\mathsf{S} = \mathbb{R}^d$ and assume that the transition kernels are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, that is,

$$P_{h+1}^{\pi_h}\cdot\ldots\cdot P_l^{\pi_{l-1}}(dy|x) = p_{h+1}^{\pi_h}\cdot\ldots\cdot p_l^{\pi_{l-1}}(y|x)\,dy.$$

Further assume that

$$\sup_{0\le h<l<H,\boldsymbol{\pi}} p_{h+1}^{\pi_h}\ldots p_l^{\pi_{l-1}}(y|x) \le Ce^{-\alpha_H|y-x|^2} \text{ for some } C,\alpha_H>0, \tag{6.11}$$

and consider some absolutely continuous reference measures $\mu_h(dx) = \mu_h(x)\,dx$, $h\in[H]$. For the bound (6.4), we then have

$$\begin{aligned}(\mathfrak{R}^{\max})^2 &= \sup_{0\le h<l<H,\boldsymbol{\pi}}\int\int\frac{\mu_h(x)}{\mu_l(y)}\left(P_{h+1}^{\pi_h}\ldots P_l^{\pi_{l-1}}(y|x)\right)^2\,dx\,dy\\ &\le C^2\max_{0\le h<l<H}\int\int\frac{\mu_h(x)}{\mu_l(x+u)}e^{-2\alpha_H|u|^2}\,dx\,du.\end{aligned}$$

The latter expression can be easily bounded by choosing μ_h to be Gaussian with an appropriate variance structure depending on h. For example, set

$$\mu_h(x) = \sqrt{\frac{\alpha_H}{\pi(h+1)}}^d e^{-\frac{\alpha_H}{h+1}|x|^2},\quad h\in[H[, \tag{6.12}$$

then the straightforward calculations yield

$$\mathfrak{R}^{\max} \le C\sqrt{\max_{0\le h<l<H}\frac{(l+1)\pi}{\alpha_H\sqrt{2(l-h)-1}}}^d \le C\sqrt{\frac{H\pi}{\alpha_H}}^d. \tag{6.13}$$

If α_H^{-1} is polynomial in H, then the bound of Theorem 4 also grows polynomially in H as opposed to most bounds available in the literature. Also note that this bound is obtained under rather general assumptions on the sets S and A. In particular, we don't assume that either S or A is finite. Note that α_H^{-1} is polynomial in H, if $p_l^a(y|x) \ge C_1e^{-C_2l^r|y-x|^2}$ for all $a\in\mathsf{A}$ and some constants $C_1, C_2, r>0$.

7. Convergence analysis of the dual algorithm

7.1. **Convergence of martingale functions.** For the dual representation (5.11) we construct an H-tuple of martingale functions $\widetilde{\boldsymbol{\eta}} := (\widetilde{\eta}_{t+1,K,M}(x,a),\, t\in[H[)$, as outlined in Section 5, from a given pre-computed H-tuple of approximate value functions $(V_{t+1,N},\, t\in[H[)$ based on sampled data $\mathcal{D}_1^N$ (see (6.5)), and a system of $K_{\rm pr}$ basis functions $\boldsymbol{\gamma}_{K_{\rm pr}}$, see Section 4.

Let us consider a fixed time $t\in[H[$ and suppress time subscripts where notationally convenient. We fix two possibly random grids $\mathsf{S}_L := \{x_1,\ldots,x_L\}$ and $\mathsf{A}_L := \{a_1,\ldots,a_L\}$ on S and A, respectively, and obtain values of the coefficient functions $c_{k,M}$ on $\mathsf{S}_L\times\mathsf{A}_L$ due to M simulations, see (5.9). Next, we construct

$$\eta_{t+1,K,M}(x,a) \equiv \eta_{t+1,K,M}(x,a,\varepsilon) = \mathbf{c}_{K,M}^\top(x,a)\boldsymbol{\psi}(\varepsilon) =: \sum_{k=1}^K c_{k,M}(x,a)\psi_k(\varepsilon),$$

for $(x,a)\in\mathsf{S}_L\times\mathsf{A}_L$. To approximate $\eta_{t+1,K,M}(x,a)$ for $(x,a)\notin\mathsf{S}_L\times\mathsf{A}_L$, we suggest to use an appropriate interpolation procedure described below, which is particularly useful for our situation where the function to be interpolated is only Lipschitz continuous (due to the presence of the

maximum). The *optimal central interpolant* for a function $f \in \mathrm{Lip}_\rho(\mathcal{L})$ on $\mathsf{S}\times\mathsf{A}$ with respect to some metric ρ on $\mathsf{S}\times\mathsf{A}$ is defined as

$$I[f](x,a) := (H_f^{\mathrm{low}}(x,a) + H_f^{\mathrm{up}}(x,a))/2,$$

where

$$H_f^{\mathrm{low}}(x,a) := \max_{(x',a')\in \mathsf{S}_L\times\mathsf{A}_L} (f(x',a') - \mathcal{L}\rho((x,a),(x',a')),$$
$$H_f^{\mathrm{up}}(x,a) := \min_{(x',a')\in \mathsf{S}_L\times\mathsf{A}_L} (f(x',a') + \mathcal{L}\rho((x,a),(x',a')).$$

Note that $H_f^{\mathrm{low}}(x,a) \le f(x,a) \le H_f^{\mathrm{up}}(x,a)$, $H_f^{\mathrm{low}}, H_f^{\mathrm{up}} \in \mathrm{Lip}_\rho(\mathcal{L})$ and hence $I[f] \in \mathrm{Lip}_\rho(\mathcal{L})$. An efficient algorithm to compute the values of the interpolant $I[f]$ without knowing $\mathcal{L}$ in advance can be found in [7]. The so constructed interpolant achieves the bound

$$\begin{aligned} \|f - I[f]\|_\infty &\le \mathcal{L}\rho_L(\mathsf{S},\mathsf{A}) \\ &:= \mathcal{L} \sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} \min_{(x',a')\in\mathsf{S}_L\times\mathsf{A}_L} \rho((x,a),(x',a')). \end{aligned} \tag{7.1}$$

The quantity $\rho_L(\mathsf{S},\mathsf{A})$ is usually called covering radius (also known as the mesh norm or fill radius) of $\mathsf{S}_L\times\mathsf{A}_L$ with respect to $\mathsf{S}\times\mathsf{A}$. We set

$$\widetilde{\eta}_{t+1,K,M}(x,a) \equiv \widetilde{\eta}_{t+1,K,M}(x,a,\varepsilon) := \sum_{k=1}^K \widetilde{c}_{k,M}(x,a)\psi_k(\varepsilon) \quad \text{with} \quad \widetilde{c}_{k,M} := I[c_{k,M}]. \tag{7.2}$$

The coefficients $\widetilde{c}_{k,M}(x,a)$ in (7.2) are considered as random, which are measurable with respect to $\mathcal{D}_{t+1}^N \vee \mathcal{G}_{t+1}^M$ with $\mathcal{G}_{t+1}^M := \sigma\left\{\widetilde{\varepsilon}_1^{t+1},\dots,\widetilde{\varepsilon}_M^{t+1}\right\}$, where $\widetilde{\varepsilon}_m^{t+1} \sim \mathcal{P}_\mathsf{E}$, $m=1,\dots,M$, $t \in [H[$, denote the i.i.d. random drawings used in (5.9). Let us denote the simulation measure (for both primal and dual) with $\mathsf{P} \equiv \mathsf{P}_{N,M} := \mathsf{P}_N \otimes \mathcal{P}_\mathsf{E}^{\otimes HM}$ (while slightly abusing notation) with $\mathsf{P}_\mathsf{N} = (\mu_h \otimes \mathcal{P}_\mathsf{E})^{\otimes HN}$.

Furthermore, denote by $\mathbf{c}_K(x,a) = [c_1(x,a),\dots,c_K(x,a)]^\top$ the unique solution of the minimization problem

$$\inf_{c_1,\dots,c_K} \mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[\left(V_{t+1}^\star(\mathcal{K}_{t+1}(x,a,\varepsilon)) - \sum_{k=1}^K c_k\psi_k(\varepsilon)\right)^2\right] \tag{7.3}$$

for any $(x,a) \in \mathsf{S}\times\mathsf{A}$, and define $\eta_{t+1,K}(x,a) := \mathbf{c}_K^\top(x,a)\boldsymbol{\psi}_K(\varepsilon)$. As such, $\eta_{t+1,K}(x,a)$ is the projection of the optimal martingale function $\eta_{t+1}^\star(x,a)$ on $\mathrm{span}(\psi_1,\dots,\psi_K)$.

Assumption 6. Assume that $|\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon)|_\infty \le \Lambda_{\mathsf{E},K}$ for all $\varepsilon\in\mathsf{E}$, and that $\mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[|\boldsymbol{\psi}_K(\varepsilon)|^2\right] \le \varrho_{\psi,K}^2$.

The following theorem provides a bound on the difference between the projection $\eta_{t+1,K}(x,a)$ and its estimate (7.2).

Theorem 5. *Under Assumptions 2, 3, 4, and 6 it holds that*

$$\begin{aligned} \mathbb{E}_{\mathcal{P}_\mathsf{E}\otimes\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} |\eta_{t+1,K}(x,a,\cdot) - \widetilde{\eta}_{t+1,K,M}(x,a,\cdot)|^2\right] \lesssim \\ \varrho_{\psi,K}^2 \frac{K(L_{V,K_{pr}}L_\mathcal{K} I_\mathcal{D}(\mathsf{S}\times\mathsf{A}) + L_{V,K_{pr}}L_\mathcal{K}\mathsf{D}(\mathsf{S}\times\mathsf{A}) + V_{\max}^\star)^2\Lambda_{\mathsf{E},K}^2}{M} \\ + K\Lambda_{\mathsf{E},K}^2\varrho_{\psi,K}^2 \sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} \left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_\infty \|V_{t+1}^\star - V_{t+1,N}\|_{L^2(\mu_{t+1}\otimes\mathsf{P})}^2 \\ + K\varrho_{\psi,K}^2 L_{V,K_{pr}}^2 L_\mathcal{K}^2 \Lambda_{\mathsf{E},K}^2 \rho_L^2(\mathsf{S},\mathsf{A}), \end{aligned}$$

where $\lesssim$ denotes $\le$ up to a natural constant, the constants $L_{V,K_{pr}}$, $L_\mathcal{K}$, and the measure μ_{t+1} are inferred from the primal procedure in Section 6.

Let us now consider the approximation error

$$\mathcal{E}^2_{K,t} := \mathbb{E}_{\varepsilon\sim\mathcal{L}_{\mathsf{E}}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta_{t+1,K}(x,a)-\eta^\star_{t+1}(x,a)\right|^2\right]$$

with

$$\eta^\star_{t+1}(x,a) = V^\star_{t+1}(\mathcal{K}_{t+1}(x,a,\varepsilon)) - \mathbb{E}\left[V^\star_{t+1}(\mathcal{K}_{t+1}(x,a,\varepsilon))\right], \quad (x,a)\in\mathsf{S}\times\mathsf{A}, \quad t\in[H[.$$

Suppose that one has pointwise

$$\eta^\star_{t+1}(x,a) = \sum_{k=1}^{\infty} c^\star_{k,t+1}(x,a)\psi_k(\varepsilon_{t+1}), \quad (x,a)\in\mathsf{S}\times\mathsf{A}, \quad t\in[H[.$$

If $\|\psi_k\|_\infty \le \psi^*_k$ for all $k\in\mathbb{N}$, then

$$\mathcal{E}^2_{K,t} \;=\; \mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\sum_{k=K+1}^{\infty} c^\star_{k,t+1}(x,a)\psi_k(\varepsilon_t)\right|^2\right] \;\le\; \sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left(\sum_{k=K+1}^{\infty}|c^\star_{k,t+1}(x,a)|\psi^*_k\right)^2.$$

If

$$\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\sum_{k=1}^{\infty} k^{\beta_\psi}|c^\star_{k,t+1}(x,a)|\psi^*_k \le C<\infty \tag{7.4}$$

for some $\beta_\psi>0$, then

$$\mathcal{E}^2_{K,t} \le C^2K^{-2\beta_\psi}. \tag{7.5}$$

Discussion.

- Let us discuss the quantity $\rho_L(\mathsf{S},\mathsf{A})$. Let $\mathsf{S}=[0,1]^{d_\mathsf{S}}$, $\mathsf{A}=[0,1]^{d_\mathsf{A}}$ for some $d_\mathsf{S},d_\mathsf{A}\in\mathbb{N}$ and let the points S_L (A_L) be uniformly distributed on S (A). Moreover set, $\rho((x,a),(x',a')) = |x-x'|+|a-a'|$. Then, similarly to [23] it can be shown that

$$\mathbb{E}[\rho^p_L(\mathsf{S}\times\mathsf{A})]^{1/p} \lesssim \sqrt{d_\mathsf{S}}\left(\frac{p\log L}{L}\right)^{1/d_\mathsf{S}} + \sqrt{d_\mathsf{A}}\left(\frac{p\log L}{L}\right)^{1/d_\mathsf{A}}, \tag{7.6}$$

  where $\lesssim$ stands for inequality up to a constant not depending on L. Using the Markov inequality, we can derive a high probability bound for $\rho_L(\mathsf{S},\mathsf{A})$. Note that if S and A are finite we need not to interpolate and $\rho_L=0$.
- Assume that all basis functions (ψ_k) are uniformly bounded and that the infinity matrix norm (i.e. the maximum absolute row sum) of the matrix $\Sigma_{\mathsf{E},K}$ is uniformly bounded from below for all $K\in\mathbb{N}$. In this case, $\varrho_{\psi,K}\lesssim K^{1/2}$, $\Lambda_{\mathsf{E},K}\lesssim 1$, $L_{V,K_{\rm pr}}\lesssim L_{\gamma,K_{\rm pr}}HK^{1/2}_{\rm pr}$. Suppose also that the quantities $\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_\infty$ are uniformly bounded for all $t>0$. Then using the bound (6.9) and the bound of Theorem 5, we arrive at

$$\mathbb{E}_{\mathcal{P}_\mathsf{E}\otimes\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta^\star_{t+1}(x,a,\cdot)-\widetilde{\eta}_{t+1,K,M}(x,a,\cdot)\right|^2\right]^{1/2} \lesssim D_{t+1}(H,K,K_{\rm pr},L)+$$
$$\frac{HKK^{1/2}_{\rm pr}L_{\gamma,K_{\rm pr}}(\sqrt{d_\mathsf{A}}+\sqrt{d_\mathsf{S}})}{\sqrt{M}} + \frac{(H-t-1)HKK^{3/2}_{\rm pr}\mathfrak{R}^{\max}L_{\gamma,K_{\rm pr}}\sqrt{d_\mathsf{A}}}{\sqrt{N}} \tag{7.7}$$

  where $D_{t+1}(H,K,K_{\rm pr},L)$ denotes the deterministic part of the error reflecting the approximation properties of the systems $\boldsymbol{\gamma}_{K_{\rm pr}}$, $\boldsymbol{\psi}_K$ and the interpolation error due to finite L (see the above discussions for some quantitative estimates). Under the above assumptions, including (7.4), one obtains from (7.5), Theorem 4, and Theorem 5,

$$D_{t+1}(H,K,K_{\rm pr},L) \lesssim K^{-\beta_\psi} + HKK^{1/2}_{\rm pr}L_{\gamma,K_{\rm pr}}\rho_L(\mathsf{S},\mathsf{A}) + K\mathfrak{R}^{\max}\sum_{l=t+1}^{H-1}\mathcal{R}_{K_{\rm pr},l},$$

where $\lesssim$ means inequality up to a constant not depending on H, N, K, $K_{\rm pr}$, and L. This bound is again polynomial in H, provided that $\mathfrak{R}^{\max}$ depends polynomially on H (see the discussion after Theorem 4).

7.2. **Convergence of upper bounds.** Suppose that the estimates $\widetilde{\boldsymbol{\eta}} = (\widetilde{\eta}_{t+1}(x,a),\, t \in [H[)$ of the optimal martingale tuple $\eta^\star = (\eta^\star_t(x,a),\, t \in]H])$ are constructed based on the sampled data $\mathcal{D}_1^N \vee \mathcal{G}_1^M \vee \ldots \vee \mathcal{G}_H^M$ such that Theorem 5 holds. Consider for $\widetilde{\boldsymbol{\xi}} := (\widetilde{\eta}_{t+1}(S_t(a_{<t}), a_t),\, a_{<t} \in \mathsf{A}^t,\, t \in [H[) \in \Xi$, $S_0 = x$, the upper bias

$$\begin{aligned}
V_0^{\rm up}(x;\widetilde{\boldsymbol{\xi}}) - V_0^\star(x) &= \mathbb{E}_x\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \widetilde{\eta}_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right] \\
&\quad - \mathbb{E}_x\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \eta^\star_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right] \\
&\leq \mathbb{E}_x\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left|\sum_{t=0}^{H-1}\eta^\star_{t+1}(S_t(a_{<t}),a_t) - \sum_{t=0}^{H-1}\widetilde{\eta}_{t+1}(S_t(a_{<t}),a_t)\right|\right] \\
&\leq \sum_{t=0}^{H-1}\mathbb{E}_x\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|\right] \\
&\leq \sum_{t=0}^{H-1}\mathbb{E}_x\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|^2\right]^{1/2},
\end{aligned}$$

where $\mathbb{E}_x$ denotes the "all-in" expectation, i.e. including the randomness of the pre-simulation, and, the independently simulated trajectories $t \to S_t(a_{<t})$. Furthermore, similarly,

$$\begin{aligned}
&\mathrm{Var}\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \widetilde{\eta}_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right] \\
&\quad = \mathrm{Var}\begin{bmatrix} \sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \widetilde{\eta}_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right) \\ -\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \eta^\star_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\end{bmatrix} \\
&\quad \leq \mathbb{E}_x\left[\left(\sum_{t=0}^{H-1}\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|\right)^2\right].
\end{aligned}$$

Hence for the standard deviation we get by the triangle inequality,

$$\begin{aligned}
&\mathrm{Dev}\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t) - \widetilde{\eta}_{t+1}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right] \\
&\qquad \leq \sum_{t=0}^{H-1}\mathbb{E}_x\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|^2\right]^{1/2}.
\end{aligned}$$

Thus, for the Monte Carlo estimate of $V_0^{\rm up}(x;\widetilde{\boldsymbol{\xi}})$,

(7.8)
$$V^{\rm up}_{0,N_{\rm test}}(x;\widetilde{\boldsymbol{\xi}}) = \frac{1}{N_{\rm test}}\sum_{n=1}^{N_{\rm test}}\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t^{(n)}(a_{<t}),a_t) - \widetilde{\eta}_{t+1}(S_t^{(n)}(a_{<t}),a_t)\right) + F(S_H^{(n)})\right)$$

with

$$S_t^{(n)}(a_{<t}) = \mathcal{K}_t(S_{t-1}^{(n)}(a_{<t-1}), a_{t-1}, \varepsilon_t^{(n)}), \quad t \in]H], \quad S_0^{(n)} = x,$$

where for each $t \in]H]$, $\varepsilon_t^{(n)}$, $n = 1, \dots, N_{\text{test}}$ are i.i.d. drawings from the distribution $\mathcal{P}_{\mathsf{E}}$, we obtain

$$\begin{aligned}
\mathbb{E}_x\left[|V^{\text{up}}_{0,N_{\text{test}}}(x;\widetilde{\boldsymbol{\xi}}) - V_0^\star(x)|^2\right]^{1/2} &\le \mathbb{E}_x\left[|V^{\text{up}}_{0,N_{\text{test}}}(x;\widetilde{\boldsymbol{\xi}}) - V_0^{\text{up}}(x;\widetilde{\boldsymbol{\xi}})|^2\right]^{1/2} + V_0^{\text{up}}(x;\widetilde{\boldsymbol{\xi}}) - V_0^\star(x) \\
&\le \frac{1}{\sqrt{N_{\text{test}}}} \sum_{t=0}^{H-1} \mathbb{E}_x\left[\sup_{(x,a)\in \mathsf{S}\times\mathsf{A}} \left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|^2\right]^{1/2} \\
&+ \sum_{t=0}^{H-1} \mathbb{E}_x\left[\sup_{(x,a)\in \mathsf{S}\times\mathsf{A}} \left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|\right] \\
&\le \left(\frac{1}{\sqrt{N_{\text{test}}}} + 1\right) \sum_{t=0}^{H-1} \mathbb{E}_x\left[\sup_{(x,a)\in \mathsf{S}\times\mathsf{A}} \left|\eta^\star_{t+1}(x,a) - \widetilde{\eta}_{t+1}(x,a)\right|^2\right]^{1/2}.
\end{aligned}$$

By using the bound of Theorem 5 (or in more specific form (7.7)), we derive the corresponding bound $\mathbb{E}_x\left[|V^{\text{up}}_{0,N_{\text{test}}}(x;\widetilde{\boldsymbol{\xi}}) - V_0^\star(x)|^2\right]^{1/2}$. Note that this bound remains polynomial in H under rather general assumptions. Let us also remark that we can use the same interpolation points to construct $\widetilde{\eta}_{t+1}(S_t^{(n)}(a_{<t}), a_t)$ for all $n = 1, \dots, N$ and all $a_{\le t}$. Let us make several concluding remarks.

- The obtained bound remains polynomial in H under rather general assumptions.
- The bound holds for general spaces S and A featuring dependence on their complexity only through the quantities $I_{\mathcal{D}}(\mathsf{S}\times\mathsf{A})$ and $\mathsf{D}(\mathsf{S}\times\mathsf{A})$. Note that these quantities remain finite for the case of S and A being compact subsets of Euclidean space.
- The above bound clearly shows the presence of a deterministic part consisting of two terms $\mathcal{R}_{K,h}$ and $\mathcal{E}_{K,t}$ (approximation/interpolation errors) and a stochastic part containing two terms of the order $1/\sqrt{M}$ and $1/\sqrt{N}$, respectively. The stochastic component of the error depends sublinearly on the dimension of the underlying Euclidean space. Moreover, in the case of finite spaces, one has a sublinear dependence on their cardinality.
- Concerning numerical complexity of computing $c_{k,M}$, note that we can use the same coefficients $c_{k,M}$ at the same interpolation points to construct $\widetilde{\eta}_{t+1}(S_t^{(n)}(a_{<t}), a_t, \varepsilon_{t+1}^{(n)})$ for all $n = 1, \dots, N$, all $a_{\le t}$, and all $\varepsilon_{t+1}^{(n)}$, see (7.2).

8. Illustrative and representative example

In this section, we illustrate the primal-dual approach proposed above in the context of Linear Convex Control, a particular MDP framework with infinite state and action space, which is popular in various practical applications (for more details on applications, see [15], Section 6.2). The setup is as follows.

- State space $(\mathsf{S}, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$;
- Action space $(\mathsf{A}, \mathcal{A}) = ([-a_{\max}, a_{\max}]^m, \mathcal{B}([-a_{\max}, a_{\max}]^m))$, $a_{\max} > 0$;
- For matrices $D_t, G_t \in \mathbb{R}^{d\times d}$, $U_t \in \mathbb{R}^{d\times m}$, $t = 0, \dots, H-1$, the controlled process dynamics are given by

$$S_{t+1} = \mathcal{K}_{t+1}(S_t, a_t, \varepsilon_{t+1}) = G_t S_t + U_t a_t + D_t \varepsilon_{t+1}, \tag{8.1}$$

  where $(\varepsilon_{t+1})_{t=0,\dots,H-1}$ are i.i.d. standard Gaussian random vectors in $\mathbb{R}^d$;
- For all $t \in [0, H[$, $R_t(x,a)$ is concave in (x,a) and $F(x)$ is concave in x.

8.1. Choice of basis functions.

Let us consider appropriate choices of basis functions such that the matrices in (4.1) and (5.8) for the primal algorithm and the dual algorithm, respectively, simplify to the diagonal ones.

Basis functions for the dual algorithm. Since the density of each r.v. ε_{t+1} is given by

$$\phi_d(z) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{|z|^2}{2}\right) = \prod_{i=1}^d \phi_1(z_i),$$

it is natural to consider the system

$$\left\{ h_l(z) := \prod_{i=1}^{d} \frac{H_{l_i}(z_i)}{\sqrt{l_i!}} : l \in \mathbb{N}_0^d,\ z \in \mathbb{R}^d \right\}, \tag{8.2}$$

where $\{H_l, l \in \mathbb{N}_0\}$ is the set of the so-called probabilistic Hermite polynomials on $\mathbb{R}$ defined as

$$H_l(u) = (-1)^l e^{\frac{u^2}{2}} \frac{d^l}{du^l} e^{-\frac{u^2}{2}}, \quad l \in \mathbb{N}_0, \quad u \in \mathbb{R}.$$

It is well-known that (8.2), constitutes a complete orthonormal system in $L^2\left(\mathbb{R}^d, \phi_d(z)dz\right)$, and follows from Gram-Schmidt orthogonalization of the sequence 1, x, x^2, ... In particular,

$$\int h_l(z)\phi_d(z)dz = \prod_{i=1}^{d} \int 1 \cdot \frac{H_{l_i}(z_i)}{\sqrt{l_i!}} \phi_1(z_i)dz_i = \prod_{i=1}^{d} \delta_{0,l_i} = \delta_{(0,\ldots,0),(l_1,\ldots,l_d)}.$$

Let $\theta : \mathbb{N}_0 \to \mathbb{N}_0^d$ be an arbitrary bijection with $\theta(0) = (0, \ldots, 0)$. We then may consider in (5.6)-(5.7) the system

$$\psi_k(z) = h_{\theta(k)}(z), \quad k = 1, 2, \ldots, \tag{8.3}$$

and obtain $\Sigma_{\mathsf{E},K} = I_d$ in (5.9).

Basis functions for the primal algorithm. It is easy to see that the transition density corresponding to (8.1) is given by

$$p_{t+1}(y|x, a) = |D_t|^{-1} \phi_d\left(D_t^{-1}\left(y - G_t x - U_t a\right)\right). \tag{8.4}$$

Assume that we have determined C and α_H such that (6.11) is satisfied for (8.4) (more details below). Then choose μ_h as in (6.12), and

$$\gamma_k(z) = h_{\theta(k)}\left(z\sqrt{\frac{2\alpha_H}{h+1}}\right), \quad k = 0, 1, 2, \ldots$$

with $h_{\theta(k)}$ defined according to (8.3). Then we have under the choice (6.12),

$$\begin{aligned}
\int \gamma_k(z)\gamma_l(z)\mu(z)\,dz &= \int \gamma_k(z)\gamma_l(z) \left(\frac{\alpha_H}{\pi(h+1)}\right)^{d/2} e^{-\frac{\alpha_H}{h+1}|z|^2} dz \\
&= \int \gamma_k\left(y\sqrt{\frac{h+1}{2\alpha_H}}\right) \gamma_l\left(y\sqrt{\frac{h+1}{2\alpha_H}}\right) \frac{1}{(2\pi)^{d/2}} e^{-|y|^2/2} dy \\
&= \int h_{\theta(k)}(y)\, h_{\theta(l)}(y) \frac{1}{(2\pi)^{d/2}} e^{-|y|^2/2} dy = \delta_{k,l}, \quad k, l \geq 0.
\end{aligned}$$

Choice of α_H Let us here assume for illustration that the matrices in (8.1) are all equal to $I \equiv I_d$, that is,

$$p_{t+1}(y|x, a) = \phi_d(y - x - a).$$

Let us determine C and α_H such that (6.11) holds. Denote for $t \in \mathbb{N}$,

$$\phi_{d,t}(y - x - a) := t^{-d/2} \phi_d\left(\frac{y - x - a}{\sqrt{t}}\right) \quad \text{with} \quad \phi_{d,1} \equiv \phi_d.$$

Then it is easy to show by induction that for any $l \geq t + 1$,

$$\begin{aligned}
&p_{t+1}^{a_t} \cdot \ldots \cdot p_l^{a_{l-1}}(y|x) = \\
&\qquad \int p_{t+1}(z_{t+1}|x, a_t) p_{t+2}(z_{t+2}|z_{t+1}, a_{t+1}) \cdot \ldots \cdot p_l(y|z_{l-1}, a_{l-1})\, dz_{t+1} \ldots dz_{l-1} = \\
&\qquad = \phi_{d,l-t}\left(y - x - \sum_{j=t+1}^{l} a_{j-1}\right) = \frac{1}{(2\pi(l-t))^{d/2}} \exp\left[-\frac{\left|y - x - \sum_{j=t+1}^{l} a_{j-1}\right|^2}{2(l-t)}\right].
\end{aligned}$$

Let us set $\theta_{t,l}(a_{t\leq\cdot<l}) := \sum_{j=t+1}^{l} a_{j-1}$. Since $|a_j| \leq a_{\max}$, one has $|\theta_{t,l}(a_{t\leq\cdot<l})| \leq (l-t)\, a_{\max}$. Next note that

$$|y-x-\theta_{t,l}(a_{t\leq\cdot<l})|^2 \geq \big||y-x| - |\theta_{t,l}(a_{t\leq\cdot<l})|\big|^2 \geq (1-\delta)^2\,|y-x|^2 - ((1-\delta)/\delta)^2(l-t)^2 a_{\max}^2$$

for any $\delta \in (0,1)$. By taking $\delta = (l-t)/(1+(l-t))$ we arrive at

$$|y-x-\theta_{t,l}(a_{t\leq\cdot<l})|^2 \geq (l-t+1)^{-2}\,|y-x|^2 - a_{\max}^2.$$

Hence

$$p_{t+1}^{a_t}\cdot\ldots\cdot p_l^{a_{l-1}}(y|x) \leq e^{a_{\max}^2 - |x-y|^2/H^2} \max_{0\leq t<l<H} \frac{1}{(2\pi(l-t))^{d/2}} = \frac{e^{a_{\max}^2}}{(2\pi)^{d/2}} e^{-|x-y|^2/H^2},$$

yielding the estimate (see (6.13))

$$\mathfrak{R}_{\max} \leq e^{a_{\max}^2}(H^3/2)^{d/2}.$$

For general matrices D_t, G_t, and U_t one can carry out a similar analysis to derive an upper bound on $\mathfrak{R}_{\max}$.

8.2. **Computing the dual upper bound.** The dual upper bound (5.11) is approximated by the upper biased estimate (7.8). In (7.8) one is faced with the path-wise computation of the quantities

$$\sup_{a_{\geq 0}\in \mathsf{A}^H}\Big(\sum_{t=0}^{H-1}\big(R_t(S_t^{(n)}(a_{<t}),a_t) - \eta_{t+1,K,M}(S_t^{(n)}(a_{<t}),a_t,\varepsilon_{t+1}^{(n)})\big) + F(S_H^{(n)}(a_{<H}))\Big), \tag{8.5}$$

for $n = 1,\ldots,N^{\text{test}}$, based on a sequence $\varepsilon_{t+1}^{(n)}$, $t = 0,\ldots,H-1, n = 1,\ldots,N^{\text{test}}$, of newly generated random variables, independent of the functions $(\eta_{t+1,K,M})$ in (5.10). For a fixed n and (partially suppressed) sequence $(\varepsilon_1,\ldots,\varepsilon_H)$, (8.5) may be cast into the form

$$\operatorname*{argsup}_{\substack{x_1,\ldots,x_H\\ a_{\geq 0}\in \mathsf{A}^H}} \left(\sum_{t=0}^{H-1}\left(R_t(x_t,a_t) - \eta_{t+1,K,M}(x_t,a_t,\varepsilon_{t+1})\right) + F(x_H)\right) \tag{8.6}$$

$$\text{subject to } x_{t+1} = \mathcal{K}\left(x_t,a_t,\varepsilon_{t+1}\right),\ t\in[0,H[,\ x_0 = S_0, \tag{8.7}$$

assuming the evolution dynamics (1). Note that in order to handle the optimization problem (8.5) more conveniently, we introduced in (8.6) a tuple of "slack variables" $(x_1,\ldots,x_H)$.

For our present example we have that (8.7) is given by (8.1) and $\mathsf{A} = [-a_{\max}, a_{\max}]^m$. Hence the control space is convex and the constraints are linear. Thus, if the given reward functions R_t and F are concave in (x,a), (8.6)-(8.7) yields a concave maximization problem, provided that $\eta_{t+1,K,M}(x,a)$ is convex in (x,a). One then may apply many established methods, such as stochastic gradient methods for instance. Of course, in general convexity of the optimal functions $\eta^{\star}_{t+1,K,M}$ or their approximations is not guaranteed. However, in many practical situations it is possible to construct convex martingale increments that nonetheless provide relatively tight upper bounds, while thus keeping the optimization problem numerically tractable. Let us next demonstrate that one may construct martingale increments that are even affine in (x,a) (see also [15], Section 6.2).

Pragmatic martingales that are affine in (x,a). Based on the lower biased approximations V_{t+1}, $t = 0,\ldots,H-1$, obtained by our primal method, we consider a suitable quadratic approximation of the form:

$$V_{t+1}(x) \approx V_{t+1}^{(0)} + x^{\top}V_{t+1}^{(1)} + x^{\top}V_{t+1}^{(2)}x, \tag{8.8}$$

where w.l.o.g. $V_{t+1}^{(2)}$ is symmetric. For instance, in order to find coefficients in (8.8) one can solve the least squares problem

$$\operatorname*{arg\,min}_{V^{(0)},V^{(1)},V^{(2)}} \mathbb{E}\left[\int\left(V^{(0)} + x^{\top}V^{(1)} + x^{\top}V^{(2)}x - V_{t+1}(x)\right)^2 \mu_{t+1}(x)dx\right].$$

Due to (8.8), one can define a proxy for (5.8) via (we take $\Sigma_{\mathsf{E},K} = I$),

$$\overline{\mathbf{c}}_K^{\circ}(x,a) := \mathbb{E}\left[\boldsymbol{\psi}_K(\varepsilon)\,\varepsilon^\top\right]\left(D_t{}^\top V_{t+1}^{(1)} + 2D_t^\top V_{t+1}^{(2)}(G_t x + U_t a)\right) + \mathbb{E}\left[\varepsilon^\top D_t^\top V_{t+1}^{(2)} D_t \varepsilon \boldsymbol{\psi}_K(\varepsilon)\right], \tag{8.9}$$

by a little algebra and using $\mathbb{E}[\boldsymbol{\psi}_K(\varepsilon)] = 0_K$. Hence, (8.9) is affine in (x,a) and known in closed form. Next, (5.10) yields an approximation for $\eta_{t+1,K,M}$,

$$\begin{aligned}\eta_{t+1,K,M}^{\circ}(x,a,\varepsilon_{t+1}) := {} & 2\boldsymbol{\psi}_K^\top(\varepsilon_{t+1})\,\mathbb{E}_{\varepsilon\sim\mathcal{P}}\left[\boldsymbol{\psi}_K(\varepsilon)\,\varepsilon^\top\right] D_t^\top V_{t+1}^{(2)}(G_t x + U_t a) \\ & + \boldsymbol{\psi}_K^\top(\varepsilon_{t+1})\,\mathbb{E}_{\varepsilon\sim\mathcal{P}}\left[\boldsymbol{\psi}_K(\varepsilon)\,\varepsilon^\top\right] D_t{}^\top V_{t+1}^{(1)} \\ & + \boldsymbol{\psi}_K^\top(\varepsilon_{t+1})\,\mathbb{E}_{\varepsilon\sim\mathcal{P}}\left[\varepsilon^\top D_t^\top V_{t+1}^{(2)} D_t \varepsilon \boldsymbol{\psi}_K(\varepsilon)\right].\end{aligned}$$

Since the second and third term are independent of (x,a), (8.6)-(8.7) are equivalent with

$$\begin{aligned}&\operatorname*{argsup}_{\substack{x_1,\ldots,x_H \\ |a_{\geq 0}|\leq a_{\max}}} \left(F(x_H) + \sum_{t=0}^{H-1} R_t(x_t,a_t) - 2\boldsymbol{\psi}_K^\top(\varepsilon_{t+1})\,\mathbb{E}_{\varepsilon\sim\mathcal{P}}\left[\boldsymbol{\psi}_K(\varepsilon)\,\varepsilon^\top\right] D_t^\top V_{t+1}^{(2)}(G_t x_t + U_t a_t)\right) \\ &\qquad \text{subject to } x_{t+1} = G_t x_t + U_t a_t + D_t \varepsilon_{t+1},\ t \in [0,H[,\ x_0 = S_0.\end{aligned} \tag{8.10}$$

Generic martingale construction. It should be stressed that the pragmatic approach leading to the concave maximization problem (8.10) may be insufficient due to a non vanishing duality gap in the general case. To illustrate the approach proposed in this sequel, let us assume that $F(x)/(1+|x|)$ and $R_t(x,a)/(1+|x|)$, are uniformly bounded in $t \in [0,H[$, $x \in \mathbb{R}^d$, and $a \in \mathbb{R}^m$, and fix some radius $r_{\max} > 0$ large enough. We then generate interpolation points in the set

$$[-r_{\max}, r_{\max}]^d \times [-a_{\max}, a_{\max}]^m,$$

and apply the interpolation procedure of [7] constrained to this region while setting

$$\eta_{t+1,K,M}(x,a,\varepsilon_{t+1}) = 0, \quad x \notin [-r_{\max}, r_{\max}]^d.$$

It is possible to show along similar lines as in [9] that due to this spatial truncation an extra bias of order $O(e^{c_1 d} H^{c_2} e^{-c_3 r_{\max}^2/H})$ for some $c_{1,2,3} > 0$ appears.

Given that a Monte Carlo sample $\varepsilon_1, \ldots, \varepsilon_H$ from the distribution of ε is available, the cost of computing the value

$$\sum_{t=0}^{H-1} \left(R_t(x_t,a_t) - \eta_{t+1,K,M}(x_t,a_t,\varepsilon_{t+1})\right) + F(x_H) \tag{8.11}$$

in (8.6) for particular sequences $x_1, \ldots, x_H$ and $a_0, \ldots, a_{H-1}$, is of order $O(H(1 + K \log L))$, if one uses the interpolation algorithm of [7]. Note that, compared to the sub-simulation based upper bound algorithm in [15] for example, our method is typically more efficient as the latter approach would require $H N_{\mathsf{sub}}$ operations for each trial sequences $x_1, \ldots, x_H$ and $a_0, \ldots, a_{H-1}$ if N_{sub} sub-simulations are used. In the context of optimal stopping, [1] used about 10000 sub-simulations for a Bermudan max-call while typically $K \log L \ll 10000$.

Solving the inner optimization problem. The inner optimization problem (8.6) can be treated as a deterministic backward dynamic program and combined with Beliakov [7] interpolation. Let us introduce for $0 \leq i \leq H$,

$$\Theta_i(x_i) = \operatorname*{argsup}_{\substack{x_{i+1},\ldots,x_H \\ a_{\geq i} \in \mathsf{A}^{H-i}}} \left(\sum_{t=i}^{H-1} \left(R_t(x_t,a_t) - \eta_{t+1,K,M}(x_t,a_t,\varepsilon_{t+1})\right) + F(x_H)\right).$$

At $i = H$, we initialize $\Theta_H(x_H) = F(x_H)$, and for $0 \leq i < H$, we may write

$$\begin{aligned}\Theta_i(x_i) = \operatorname*{argsup}_{x_{i+1},\ a\in A} & \left(R_i(x_i,a) - \eta_{i+1,K,M}(x_i,a,\varepsilon_{i+1}) + \Theta_{i+1}(x_{i+1})\right) \\ & \text{subject to } x_{i+1} = \mathcal{K}(x_i,a,\varepsilon_{i+1}).\end{aligned}$$

Let $\Theta^{\mathsf{up}}_{i+1}$ be the upper Beliakov interpolation of an approximation $\Theta^{\mathsf{grid}}_{i+1}$ of Θ_{i+1} on a spatial grid S_L. We then compute for $x_i \in \mathsf{S}_L$,

$$\Theta^{\mathsf{grid}}_i(x_i) = \underset{x_{i+1},\ a\in A}{\operatorname{argsup}} \left(R_i(x_i,a) - \eta_{i+1,K,M}(x_i,a,\varepsilon_{i+1}) + \Theta^{\mathsf{up}}_{i+1}(x_{i+1})\right)$$
$$\text{subject to } x_{i+1} = \mathcal{K}\left(x_i,a,\varepsilon_{i+1}\right),$$

and take for Θ^{up}_i the upper Beliakov interpolation of Θ^{grid}_i. The choice of the upper Beliakov interpolation rather than the mid Beliakov interpolation ensures that the estimator Θ remains upper biased. We thus end up with Θ^{up}_0 as the solution of the pathwise (for a fixed sequence $(\varepsilon_1,\dots,\varepsilon_H)$) inner optimization problem.

Remark 6. Let us finally note once again that the number N_{test} of test simulations in (7.8), hence the sequences $(\varepsilon_1^{(n)},\dots,\varepsilon_H^{(n)})$, $n=1,\dots,N_{\mathsf{test}}$, can usually be relatively small when the martingale increments $\eta_{i+1,K,M}$ are accurate enough.

9. Proofs

9.1. **Proof of Theorem 4.** *One-step analysis:* Suppose that after h steps of the algorithm the estimates $V_{H,N},\dots,V_{h+1,N}$ of the value functions $V^\star_H,\dots,V^\star_{h+1}$, respectively, are constructed using sampled data $\mathcal{D}^N_{h+1}$, such that $\|V_{t,N}\|_\infty \le \widetilde{L}_t \le V^\star_{\max}$ a.s. for all $t=h+1,\dots,H$. Denote for $a\in\mathsf{A}$,

$$\ell^a(\beta) := \mathbb{E}\left[(Z^a - \beta^\top \boldsymbol{\gamma}_K(X))^2 \mid \mathcal{D}^N_{h+1}\right] \quad \text{with}$$
$$Z^a \sim V_{h+1,N}(Y^{a,X}), \quad Y^{a,X}\sim P_{h+1}(\cdot|X,a), \quad X\sim\mu_h.$$

The unique minimizer of $\ell^a(\beta)$ is given by the $\mathcal{D}^N_{h+1}$-measurable vector

$$\beta_a = \mathbb{E}\left[Z^a\Sigma^{-1}\boldsymbol{\gamma}_K(X)\mid\mathcal{D}^N_{h+1}\right] = \mathbb{E}_{X\sim\mu_h}\left[P^a_{h+1}V_{h+1,N}(X)\Sigma^{-1}\boldsymbol{\gamma}_K(X)\mid\mathcal{D}^N_{h+1}\right].$$

For the estimation of the $\mathcal{D}^N_h$-measurable vector $\beta_{N,a}$ in (4.4), see (4.2), (4.3), and Assumption 2, it then holds that

$$\begin{aligned}
&\mathbb{E}_{\mu_h\otimes\mathsf{P}}\left[\sup_{a\in\mathsf{A}}\left((\beta^\top_{N,a}-\beta^\top_a)\boldsymbol{\gamma}_K(X)\right)^2 \mid \mathcal{D}^N_{h+1}\right]\\
&\qquad\le \mathbb{E}_{\mathsf{P}}\left[\sup_{a\in\mathsf{A}}|\beta_{N,a}-\beta_a|^2 \mid \mathcal{D}^N_{h+1}\right]\mathbb{E}_{X\sim\mu_h}\left[|\boldsymbol{\gamma}_K(X)|^2\right]\\
&\qquad\le \sum_{k=1}^K \mathbb{E}_{\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\beta_{N,a,k}-\beta_{a,k})^2 \mid \mathcal{D}^N_{h+1}\right]\mathbb{E}_{X\sim\mu_h}\left[|\boldsymbol{\gamma}_K(X)|^2\right],
\end{aligned}$$

where according to Proposition 7, (component wise applied to the vector function $f(x,a,\varepsilon) = V_{h+1,N}(\mathcal{K}_{h+1}(x,a,\varepsilon))\Sigma^{-1}\boldsymbol{\gamma}_K(x)$ with $p=2$, see (6.3)) one has for $k=1,\dots,K$,

$$\mathbb{E}_{\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\beta_{N,a,k}-\beta_{a,k})^2 \mid \mathcal{D}^N_{h+1}\right] \lesssim \frac{(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A}) + L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A}) + V^\star_{\max})^2\Lambda^2_K}{N}. \tag{9.1}$$

Due to the very structure of $V_{h+1,N}$ (see (4.2)), we further have

$$\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta_a^\top\boldsymbol{\gamma}(X) - P^a_{h+1}V_{h+1,N}(X))^2 \mid \mathcal{D}^N_{h+1}\right] \le \mathcal{R}^2_{K,h}, \tag{9.2}$$

and then with (9.1) and (9.2) we have the estimate

$$\begin{aligned}(9.3)\quad \mathbb{E}_{\mu_h\otimes\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\widetilde{P}^a_{h+1,N}V_{h+1,N}(X)-P^a_{h+1}V_{h+1,N}(X))^2\mid\mathcal{D}^N_{h+1}\right]^{1/2}\leq\\ \mathbb{E}_{\mu_h\otimes\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\beta^\top_{N,a}\boldsymbol{\gamma}_K(X)-P^a_{h+1}V_{h+1,N}(X))^2\mid\mathcal{D}^N_{h+1}\right]^{1/2}\leq\\ \mathbb{E}_{\mu_h\otimes\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\beta^\top_{N,a}\boldsymbol{\gamma}_K(X)-\beta^\top_a\boldsymbol{\gamma}_K(X))^2\mid\mathcal{D}^N_{h+1}\right]^{1/2}\\ +\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta^\top_a\boldsymbol{\gamma}_K(X)-P^a_{h+1}V_{h+1,N}(X))^2\mid\mathcal{D}^N_{h+1}\right]^{1/2}\\ \leq\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A})+V^\star_{\max})\sqrt{\frac{K}{N}}+\mathcal{R}_{K,h}.\end{aligned}$$

Since the right-hand-side of (9.3) is deterministic, the conditioning on $\mathcal{D}^N_{h+1}$ may be dropped and we obtain

$$\begin{aligned}(9.4)\quad \mathbb{E}_{\mu_h\otimes\mathsf{P}}\left[\sup_{a\in\mathsf{A}}(\widetilde{P}^a_{h+1,N}V_{h+1,N}(X)-P^a_{h+1}V_{h+1,N}(X))^2\right]^{1/2}\leq\\ \leq\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A})+V^\star_{\max})\sqrt{\frac{K}{N}}+\mathcal{R}_{K,h}.\end{aligned}$$

Multi step analysis: Let us denote for $h\in[H[$,

$$(9.5)\qquad \Delta^a_{h,N}(x):=\widetilde{P}^a_{h+1,N}V_{h+1,N}(x)-P^a_{h+1}V_{h+1,N}(x)\quad\text{and}\quad\Delta_h(x):=\sup_{a\in\mathsf{A}}|\Delta^a_{h,N}(x)|.$$

Note that

$$P^{\pi_h}_{h+1}P^{\pi_{h'}}_{h'+1}(dx''|x)=\int_{\mathsf{S}}P^{\pi_h}_{h+1}(dx'|x)P^{\pi_{h'}}_{h'+1}(dx''|x').$$

We then have

$$\begin{aligned}V^\star_h(x)-V_{h,N}(x)&=\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+P^a_{h+1}V^\star_{h+1}(x)\right\}-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}^a_{h+1,N}V_{h+1,N}(x)\right\}\\ &=R_h(x,\pi^\star_h(x))+\int V^\star_{h+1}(x')P_{h+1}(dx'|x,\pi^\star_h(x))\\ &\quad-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}^a_{h+1,N}V_{h+1,N}(x)\right\}\\ &\leq\int\left(V^\star_{h+1}-V_{h+1,N}\right)(x')P_{h+1}(dx'|x,\pi^\star_h(x))\\ &\quad+\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+P^a_{h+1}V_{h+1,N}(x)\right\}-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}^a_{h+1,N}V_{h+1,N}(x)\right\}\\ &\leq P^{\pi^\star_h}_{h+1}\left(V^\star_{h+1}-V_{h+1,N}\right)(x)+\ \Delta_h(x)\qquad(9.6)\end{aligned}$$

and analogously,

$$(9.7)\qquad V^\star_h(x)-V_{h,N}(x)\geq P^{\pi_{h,N}}_{h+1}[V^\star_{h+1}-V_{h+1,N}](x)-\Delta_h(x).$$

By iterating (9.6) and (9.7) upwards, and using that $V_{H,N}=V^\star_H$, we obtain, respectively,

$$V^\star_h(x)-V_{h,N}(x)\leq\sum_{k=1}^{H-h-1}P^{\pi^\star_h}_{h+1}\dots P^{\pi^\star_{h+k-1}}_{h+k}[\Delta_{h+k}](x)+\Delta_h(x),\quad\text{and}$$

$$V^\star_h(x)-V_{h,N}(x)\geq-\sum_{k=1}^{H-h-1}P^{\pi_{h,N}}_{h+1}\dots P^{\pi_{h+k-1,N}}_{h+k}[\Delta_{h+k}](x)-\Delta_h(x).$$

We thus have pointwise,

$$|V_h^\star(x) - V_{h,N}(x)| \le \sum_{k=1}^{H-h-1} P_{h+1}^{\pi_h^\star} \dots P_{h+k}^{\pi_{h+k-1}^\star}[\Delta_{h+k}](x) + \sum_{k=1}^{H-h-1} P_{h+1}^{\pi_{h,N}} \dots P_{h+k}^{\pi_{h+k-1,N}}[\Delta_{h+k}](x) + \Delta_h(x)$$

which implies

$$\|V_h^\star - V_{h,N}\|_{L^2(\mu_h\otimes \mathsf{P})} \le 2 \sup_{\boldsymbol{\pi}} \sum_{k=1}^{H-h-1} \left\| P_{h+1}^{\pi_h} \dots P_{h+k}^{\pi_{h+k-1}}[\Delta_{h+k}] \right\|_{L^2(\mu_h\otimes \mathsf{P})} + \|\Delta_h\|_{L^2(\mu_h\otimes \mathsf{P})}\,.$$

Hence we have due to Assumption 5,

$$\|V_h^\star - V_{h,N}\|_{L^2(\mu_h\otimes \mathsf{P})} \le 2\mathfrak{R}^{\max} \sum_{l=h}^{H-1} \|\Delta_l\|_{L^2(\mu_l\otimes \mathsf{P})}$$

(note that $\mathfrak{R}^{\max} \ge 1$), and then, by the definitions (9.5) and the estimate (9.4), the statement of the theorem follows.

9.2. **Proof of Theorem 5.** For the unique minimizer of (7.3) one has that

$$\mathbf{c}_K(x,a) := \Sigma_{\mathsf{E},K}^{-1}\, \mathbb{E}\left[V_{t+1}^\star(\mathcal{K}_{t+1}(x,a,\varepsilon))\boldsymbol{\psi}_K(\varepsilon)\right]. \tag{9.8}$$

Likewise, the unique minimizer of the problem

$$\inf_{\mathbf{c}\in\mathbb{R}^K} \mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[\left(V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon)) - \mathbf{c}^\top \boldsymbol{\psi}_K(\varepsilon)\right)^2 |\mathcal{D}_{t+1}^N\right]$$

is given by

$$\bar{\mathbf{c}}_K(x,a) := \Sigma_{\mathsf{E},K}^{-1}\, \mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon))\boldsymbol{\psi}_K(\varepsilon)|\mathcal{D}_{t+1}^N\right].$$

Now let $\mathbf{c}_{K,M}(x,a)$ be the Monte Carlo estimate of $\bar{\mathbf{c}}_K(x,a)$ as constructed in Section 5, see (5.8) and (5.9). We then have

$$\mathbb{E}_{\mathcal{P}_\mathsf{E}\otimes\mathsf{P}}\left[\left|\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} (\mathbf{c}_{K,M} - \bar{\mathbf{c}}_K)^\top(x,a)\boldsymbol{\psi}_K(\varepsilon)\right|^2 |\mathcal{D}_{t+1}^N\right] \le \mathbb{E}_\mathsf{P}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} \left|(\mathbf{c}_{K,M} - \bar{\mathbf{c}}_K)^\top(x,a)\right|^2 |\mathcal{D}_{t+1}^N\right] \mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[|\boldsymbol{\psi}_K(\varepsilon)|^2\right], \tag{9.9}$$

where according to Proposition 7 (applied componentwise with $p=2$ to the vector function $f(x,a,\varepsilon) = V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon))\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon)$, see (6.3))

$$\mathbb{E}_\mathsf{P}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} |(\mathbf{c}_{K,M} - \bar{\mathbf{c}}_K)(x,a)|^2 |\mathcal{D}_{t+1}^N\right] \le \frac{K(L_{V,K_{\rm pr}} L_\mathcal{K} I_\mathcal{D}(\mathsf{S}\times\mathsf{A}) + L_{V,K_{\rm pr}} L_\mathcal{K} \mathsf{D}(\mathsf{S}\times\mathsf{A}) + V_{\max}^\star)^2 \Lambda_{\mathsf{E},K}^2}{M}. \tag{9.10}$$

Since for any pair $(x,a)\in\mathsf{S}\times\mathsf{A}$,

$$\begin{aligned}
|(\mathbf{c}_K - \bar{\mathbf{c}}_K)(x,a)|^2 &= \left|\mathbb{E}_{\varepsilon\sim\mathcal{P}_\mathsf{E}}\left[\left(V_{t+1}^\star(\mathcal{K}_{t+1}(x,a,\varepsilon)) - V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon))\right)\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon)|\mathcal{D}_{t+1}^N\right]\right|^2 \\
&\le \int \left|V_{t+1}^\star(\mathcal{K}_{t+1}(x,a,\varepsilon)) - V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon))\right|^2 d\mathcal{P}_\mathsf{E}(\varepsilon) \int \left|\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon)\right|^2 d\mathcal{P}_\mathsf{E}(\varepsilon) \\
&\le K\Lambda_{\mathsf{E},K}^2 \sup_{(x,a)\in\mathsf{S}\times\mathsf{A}} \left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_\infty \int \left|V_{t+1}^\star(y) - V_{t+1,N}(y)\right|^2 \mu_{t+1}(dy),
\end{aligned}$$

we have

$$\mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}\left[\left|\max_{(x,a)\in\mathsf{S}_L\times\mathsf{A}_L}(\mathbf{c}_K-\bar{\mathbf{c}}_K)^\top(x,a)\boldsymbol{\psi}_K(\varepsilon)\right|^2|\mathcal{D}_{t+1}^N\right] \tag{9.11}$$
$$\leq \max_{(x,a)\in\mathsf{S}_L\times\mathsf{A}_L}|(\mathbf{c}_K-\bar{\mathbf{c}}_K)(x,a)|^2\,\mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}\left[|\boldsymbol{\psi}_K(\varepsilon)|^2\right]$$
$$\leq K\varrho_{\psi,K}^2\Lambda_{\mathsf{E},K}^2\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_\infty\|V_{t+1}^\star-V_{t+1,N}\|_{L^2(\mu_{t+1})}^2.$$

Next due to (6.3), we derive for any $k\in[K]$,

$$|c_{k,M}(x,a)-c_{k,M}(x',a')|$$
$$\leq\frac{1}{M}\sum_{m=1}^M|V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\widetilde{\varepsilon}_m))-V_{t+1,N}(\mathcal{K}_{t+1}(x',a',\widetilde{\varepsilon}_m))||\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\widetilde{\varepsilon}_m)|_\infty$$
$$\leq L_{V,K_{\mathrm{pr}}}L_{\mathcal{K}}\Lambda_{\mathsf{E},K}\rho((x,a),(x',a'))$$

and so with $I[\mathbf{c}_{K,M}]:=(I[c_{1,M}],\ldots,I[c_{K,M}])^\top$ we further have

$$\mathbb{E}_{\mathcal{P}_{\mathsf{E}}\otimes\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}|\eta_{t+1,K,M}(x,a)-\widetilde{\eta}_{t+1,K,M}(x,a)|^2|\mathcal{D}_{t+1}^N\right]$$
$$=\mathbb{E}_{\mathcal{P}_{\mathsf{E}}\otimes\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|(\mathbf{c}_{K,M}-I[\mathbf{c}_{K,M}])^\top(x,a)\boldsymbol{\psi}_K(\varepsilon_{t+1})\right|^2|\mathcal{D}_{t+1}^N\right]$$
$$\leq\varrho_{\psi,K}^2\mathbb{E}_{\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}|(\mathbf{c}_{K,M}-I[\mathbf{c}_{K,M}])(x,a)|^2|\mathcal{D}_{t+1}^N\right]$$
$$\leq\varrho_{\psi,K}^2\sum_{k=1}^K\mathbb{E}_{\mathsf{P}}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}(c_{k,M}-I[c_{k,M}])^2(x,a)|\mathcal{D}_{t+1}^N\right]$$
$$\leq K\varrho_{\psi,K}^2L_{V,K_{\mathrm{pr}}}^2L_{\mathcal{K}}^2\Lambda_{\mathsf{E},K}^2\rho_L^2(\mathsf{S},\mathsf{A}), \tag{9.12}$$

using (7.1). Finally note that

$$\eta_{t+1,K}-\widetilde{\eta}_{t+1,K,M}=(\mathbf{c}_K-\bar{\mathbf{c}}_K)^\top\boldsymbol{\psi}_K+(\bar{\mathbf{c}}_K-\mathbf{c}_{K,M})^\top\boldsymbol{\psi}_K+\eta_{t+1,K,M}-\widetilde{\eta}_{t+1,K,M}$$

and then the result follows by the triangle inequality, gathering (9.9)–(9.12), and finally taking the unconditional expectation $\mathbb{E}_{\mathcal{P}_{\mathsf{E}}\otimes\mathsf{P}}$.

Appendix A. Some auxiliary notions

The Orlicz 2-norm of a real valued random variable η with respect to the function $\varphi(x) = e^{x^2} - 1$, $x \in \mathbb{R}$, is defined by $\|\eta\|_{\varphi,2} := \inf\{t > 0 : \mathbb{E}\left[\exp\left(\eta^2/t^2\right)\right] \le 2\}$. We say that η is *sub-Gaussian* if $\|\eta\|_{\varphi,2} < \infty$. In particular, this implies that for some constants $C, c > 0$,

$$\mathsf{P}(|\eta| \ge t) \le 2\exp\left(-\frac{ct^2}{\|\eta\|^2_{\varphi,2}}\right) \quad \text{and} \quad \mathbb{E}[|\eta|^p]^{1/p} \le C\sqrt{p}\|\eta\|_{\varphi,2} \quad \text{for all} \quad p \ge 1.$$

Consider a real valued random process $(X_t)_{t\in\mathcal{T}}$ on a metric parameter space $(\mathcal{T}, \mathsf{d})$. We say that the process has *sub-Gaussian increments* if there exists $K \ge 0$ such that

$$\|X_t - X_s\|_{\varphi,2} \le K\mathsf{d}(t,s), \quad \forall t, s \in \mathcal{T}.$$

Let (Y, ρ) be a metric space and $\mathsf{X} \subseteq \mathsf{Y}$. For $\varepsilon > 0$, we denote by $\mathcal{N}(\mathsf{X}, \rho, \varepsilon)$ the covering number of the set X with respect to the metric ρ, that is, the smallest cardinality of a set (or net) of ε-balls in the metric ρ that covers X. Then $\log\mathcal{N}(\mathsf{X}, \rho, \varepsilon)$ is called the metric entropy of X and

$$I_{\mathcal{D}}(\mathsf{X}) := \int_0^{\mathsf{D}(\mathsf{X})} \sqrt{\log\mathcal{N}\big(\mathsf{X}, \rho, u\big)}\, du$$

with $\mathsf{D}(\mathsf{X}) := \operatorname{diam}(\mathsf{X}) := \max_{x,x'\in\mathsf{X}} \rho(x, x')$, is called the Dudley integral. For example, if $|\mathsf{X}| < \infty$ and $\rho(x, x') = 1_{\{x\neq x'\}}$ we get $I_{\mathcal{D}}(\mathsf{X}) = \sqrt{\log|\mathsf{X}|}$.

Appendix B. Estimation of mean uniformly in parameter

The following proposition holds.

Proposition 7. *Let f be a function on $\mathsf{X}\times\Xi$ such that*

$$\big|f(x,\xi) - f(x',\xi)\big| \le L\rho(x,x') \tag{B.1}$$

with some constant $L > 0$. Furthermore assume that $\|f\|_\infty \le F < \infty$ for some $F > 0$. Let ξ_n, $n = 1,\dots,N$, be i.i.d. sample from a distribution on Ξ. Then we have

$$\mathbb{E}^{1/p}\left[\sup_{x\in\mathsf{X}}\left|\frac{1}{N}\sum_{n=1}^N (f(x,\xi_n) - \mathbb{E}f(x,\xi_n))\right|^p\right] \lesssim \frac{LI_{\mathcal{D}} + (L\mathsf{D} + F)\sqrt{p}}{\sqrt{N}},$$

where $\lesssim$ may be interpreted as $\le$ up to a natural constant.

Proof. Denote

$$Z(x) := \frac{1}{\sqrt{N}}\sum_{n=1}^N (f(x,\xi_n) - M_f(x))$$

with $M_f(x) = \mathsf{E}[f(x,\xi)]$, that is, $Z(x)$ is a centered random process on the metric space (X, ρ). Below we show that the process $Z(x)$ has sub-Gaussian increments. In order to show it, let us introduce

$$Z_n = f(x,\xi_n) - M_f(x) - f(x',\xi_n) + M_f(x').$$

Under our assumptions we get

$$\|Z_n\|_{\psi_2} \lesssim L\rho(x,x'),$$

that is, Z_n is subgaussian for any $n = 1,\dots,N$. Since

$$Z(x) - Z(x') = N^{-1/2}\sum_{n=1}^N Z_n,$$

is a sum of independent sub-Gaussian r.v, we may apply [30, Proposition 2.6.1 and Eq. (2.16)]) to obtain that $Z(x)$ has sub-Gaussian increments with parameter $K \asymp L$. Fix some $x_0 \in \mathsf{X}$. By the triangular inequality,

$$\sup_{x\in\mathsf{X}}|Z(x)| \le \sup_{x,x'\in\mathsf{X}}|Z(x) - Z(x')| + |Z(x_0)|\,.$$

By the Dudley integral inequality, e.g. [30, Theorem 8.1.6], for any $\delta \in (0,1)$,

$$\sup_{x,x' \in \mathsf{X}} |Z(x) - Z(x')| \lesssim L\big[I_{\mathcal{D}} + \mathsf{D}\sqrt{\log(2/\delta)}\big]$$

holds with probability at least $1-\delta$. Again, under our assumptions, $Z(x_0)$ is a sum of i.i.d. bounded centered random variables with ψ_2-norm bounded by F. Hence, applying Hoeffding's inequality, e.g. [30, Theorem 2.6.2.], for any $\delta \in (0,1)$,

$$|Z(x_0)| \lesssim F\sqrt{\log(1/\delta)}.$$

□

References

[1] Leif Andersen and Mark Broadie. A Primal-Dual Simulation Algorithm for Pricing Multi-Dimensional American Options. *Management Science*, 50(9):1222–1234, 2004.

[2] András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[3] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017.

[4] Sven Balder, Antje Mahayni, and John Schoenmakers. Primal-dual linear Monte Carlo algorithm for multiple stopping—an application to flexible caps. *Quant. Finance*, 13(7):1003–1013, 2013.

[5] Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Universitext. Berlin: Springer, 2011.

[6] Christian Bayer, Martin Redmann, and John Schoenmakers. Dynamic programming for optimal stopping via pseudo-regression. *Quant. Finance*, 21(1):29–44, 2021.

[7] Gleb Beliakov. Interpolation of Lipschitz functions. *Journal of computational and applied mathematics*, 196(1):20–44, 2006.

[8] Denis Belomestny, Christian Bender, and John Schoenmakers. True upper bounds for bermudan products via non-nested Monte Carlo. *Math. Finance*, 19(1):53–71, 2009.

[9] Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Math. Finance*, 30(4):1591–1616, 2020.

[10] Denis Belomestny, Anastasia Kolodko, and John Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM J. Control Optim.*, 48(5):3562–3588, 2010.

[11] Christian Bender, John Schoenmakers, and Jianing Zhang. Dual representations for general multiple stopping problems. *Math. Finance*, 25(2):339–370, 2015.

[12] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[13] David B Brown, James E Smith, et al. Information relaxations and duality in stochastic dynamic programs: A review and tutorial. *Foundations and Trends® in Optimization*, 5(3):246–339, 2022.

[14] David B. Brown, James E. Smith, and Peng Sun. Information relaxations and duality in stochastic dynamic programs. *Oper. Res.*, 58(4, part 1):785–801, 2010.

[15] Vijay V Desai, Vivek F Farias, and Ciamac C Moallemi. Bounds for Markov decision processes. *Reinforcement learning and approximate dynamic programming for feedback control*, pages 452–473, 2012.

[16] Paul Glasserman, Bin Yu, et al. Number of paths versus number of basis functions in American option pricing. *The Annals of Applied Probability*, 14(4):2090–2119, 2004.

[17] E. Gobet, J. G. López-Salas, P. Turkedjiev, and C. Vázquez. Stratified regression Monte-Carlo scheme for semilinear pdes and BSDEs with large scale parallelization on GPUs. *SIAM Journal on Scientific Computing*, 38(6):C652–C677, 2016.

[18] Emmanuel Gobet and Plamen Turkedjiev. Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions. *Mathematics of Computation*, 85(299):1359–1391, 2016.

[19] Martin Haugh and Leonid Kogan. Pricing American options: A duality approach. *Oper. Res.*, 52(2):258–270, 2004.

[20] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[21] Bernardo Ávila Pires and Csaba Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151. PMLR, 2016.

[22] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley Ser. Probab. Math. Stat. New York, NY: John Wiley & Sons, Inc., 1994.

[23] A. Reznikov and E. B. Saff. The covering radius of randomly distributed points on a manifold. *Int. Math. Res. Not. IMRN*, (19):6065–6094, 2016.

[24] L. Rogers. Pathwise stochastic optimal control. *SIAM J. Control and Optimization*, 46:1116–1132, 01 2007.

[25] Leonard C. G. Rogers. Monte Carlo valuation of American options. *Mathematical Finance*, 12(3):271–286, 2002.
[26] J. Schoenmakers. A pure martingale dual for multiple stopping. *Finance Stoch.*, 16:319–334, 2012.
[27] Lars Stentoft. Convergence of the least squares Monte Carlo approach to American option valuation. *Management Science*, 50(9):1193–1203, 2004.
[28] R. S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.
[29] Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.
[30] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.
[31] Daniel Z. Zanger. Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing. *Finance and Stochastics*, 17(3):503–534, 2013.
[32] Daniel Z Zanger. General error estimates for the Longstaff–Schwartz least-squares Monte Carlo algorithm. *Mathematics of Operations Research*, 45(3):923–946, 2020.
[33] Helin Zhu, Fan Ye, and Enlu Zhou. Solving the dual problems of dynamic programs via regression. *IEEE Transactions on Automatic Control*, 63(5):1340–1355, 2017.

[1]Faculty of Mathematics, Duisburg-Essen University, Thea-Leymann-Str. 9, D-45127 Essen, Germany
Email address: denis.belomestny@uni-due.de

[2]Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany
Email address: schoenma@wias-berlin.de