Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Primal-dual regression approach for Markov decision processes with general state and action space

Denis Belomestny[1], John G. M. Schoenmakers[2]

submitted: August 30, 2022

[1] Duisburg-Essen University
Thea-Leymann-Str. 9
45127 Essen
Germany
E-Mail: denis.belomestny@uni-due.de

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: john.schoenmakers@wias-berlin.de

No. 2957
Berlin 2022

2020 *Mathematics Subject Classification.* 90C40, 65C05, 62G08.

Key words and phrases. Markov decision processes, dual representation, pseudo regression.

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

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Primal-dual regression approach for Markov decision processes with general state and action space

Denis Belomestny, John G. M. Schoenmakers

Abstract

We develop a regression based primal-dual martingale approach for solving finite time horizon MDPs with general state and action 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 tight error bounds for the estimated duality gap featuring polynomial dependence on the time horizon, and sublinear dependence on the cardinality/dimension of the possibly infinite state and action space. From a computational point of view the proposed method is efficient since, in contrast to usual duality-based methods for optimal control problems in the literature, the Monte Carlo procedures here involved 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 seen as MDPs and can, in principle, be solved via a dynamic programming approach. The objective usually is to find an optimal policy that maximizes the expected accumulated rewards (or minimizes the expected accumulated costs). These problems could be theoretically solved by the dynamic programming approach; however, in practice, this method 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 general cases, the possibility of beating the curse of the horizon remains an open issue.

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 RL algorithms like Q-learning Jin et al. [2018], optimistic value iteration Azar et al. [2017], 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. 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. Antos et al. [2007], Szepesvári [2010], Pires and Szepesvári [2016] 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 Policy Iteration (API, Bertsekas and Tsitsiklis [1996]) 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. Sutton and Barto [2018]), there is always an approximation error due to the choice of the family of policies that can be hardly quantified. Though the accuracy of a suboptimal policy is generally unknown, the lack of theoretical guarantees on a suboptimal policy can be potentially 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 Rogers [2002] and Haugh and Kogan [2004] 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 over a set of martingales or martingale type elements. In stylized 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

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

As such, in the dual approach one seeks for optimal minimizing martingales rather than optimal maximizing policies. Andersen and Broadie [2004] showed how to compute martingales using stopping rules via nested Monte Carlo simulations. In Rogers [2007] the dual representation for optimal stopping (hence American options) was generalized to Markovian control problems. Somewhat later Brown et al. [2010] presented a dual representation for quite general control problems in terms of Ãnformation relaxation and martingale penalties. On the other hand, the dual representation for optimal stopping was generalized to multiple stopping in Schoenmakers [2012] and Bender et al. [2015]. As a numerical approach to Rogers [2007], Belomestny et al. [2010] developed regression methods for such problems that can be seen, in a sense, as a generalization of Andersen and Broadie [2004]. However, it should be noted that in the convergence analysis of Belomestny et al. [2010] the primal value function estimates show exponential dependence on the finite horizon, and their dual algorithm is based on nested simulation while its convergence is 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 Zanger [2013].

In this paper, we propose a novel approach to 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). We thus avoid the delicate problem of inverting an empirical covariance matrix. We note that in the context of optimal stopping a similar primal procedure was proposed in Bayer et al. [2021], 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 Carlo simulation (as used in many dual-type methods proposed in the literature so far, see for instance the path-wise optimization approach for MDPs in Desai et al. [2012] and Brown et al. [2022] for an overview). Instead, for constructing the martingale elements, we propose to combine a point wise pseudo regression approach with a suitable interpolation method such that the martingale property is preserved. Furthermore we provide a rigorous convergence analysis showing that the stochastic error of approximating the true value function depends at most polynomially on the time horizon. Moreover, we show that the stochastic part of the error depends sublinearly on the dimension (or cardinality in the finite case) of the state and action spaces. Let us mention Zhu et al. [2017] for another approach to avoid nested simulations for estimating the conditional expectations, hence the martingale elements, inside the dual representation. However, Zhu et al. [2017] left the issue of bounding the duality gap in terms of the error bounds on the value functions as an open problem. From 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. 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 the discrete time finite horizon Markov Decision Process, given by the setup

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

made up by the following objects:

- a measurable state space $(\mathsf{S}, \mathcal{S})$, which may be finite or infinite;
- a measurable action space $(\mathsf{A}, \mathcal{A})$, which may be finite or infinite;
- an integer H which defines the horizon of the problem;
- for each $h \in]H]$, with $]H] := \{1, \ldots, H\}$[1], 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 $F : \mathsf{S} \to \mathbb{R}$.

Introduce a filtered probability space $\mathfrak{S} := \big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[H]}, \mathbb{P}\big)$ with

$$\Omega := (\mathsf{S} \times \mathsf{A})^{[H]}, \quad \mathcal{F} := (\mathcal{S} \otimes \mathcal{A})^{\otimes[H]}, \quad (\mathcal{F}_t)_{t\in[H]} := ((\mathcal{S} \otimes \mathcal{A})^{\otimes t})_{t\in[H]}.^2 \tag{2.1}$$

For a fixed policy $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{H-1})$ with $\pi_t : \mathsf{S} \to \mathcal{P}(\mathsf{A})$, we consider an adapted controlled process $\big(S_t, A_t\big)_{t=h,\ldots,H}$ on $\mathfrak{S}$ satisfying $S_0 = x \in \mathsf{S}$, $A_0 \sim \pi_0(x)$, and

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

The expected reward of this so called Markov Decision Process due to the chosen policy $\boldsymbol{\pi}$ is given by

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

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:

$$V_0^\star := \sup_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi},x}\left[\sum_{t=0}^{H-1} R_t(S_t, A_t) + F(S_H)\right] = \sup_{\boldsymbol{\pi}} V_0^{\boldsymbol{\pi}}(x_0). \tag{2.2}$$

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

[2] In order to avoid irrelevant measure theoretic technicalities it is assumed that our probability space is supported by discrete time processes, rather than Wiener processes for instance. Nonetheless, it is possible to involve larger probability spaces without essentially affecting the results in this paper.

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

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

be the optimal value function at $h \in [H]$. It is well known that under weak conditions there exists an optimal policy which depends on S_t in a deterministic way. In this case we will write $\boldsymbol{\pi}^\star = (\pi_t^\star(S_t))$ for some mappings $\pi_t^\star : \mathsf{S} \to \mathsf{A}$, where Dirac measures $\delta_{\{a\}}$ are identified with their supporting elements $a \in \mathsf{A}$. One has the following result, see Puterman [2014].

Theorem 2.1 *Let $x \in \mathsf{S}$ be fixed. It holds $V_H^\star(x) = F(x)$, and*

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

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

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

Then $Q_h^\star(x,a)$ is called the *optimal state-action value* 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 2.1,

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

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

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

$$S_0 = x_0, \quad S_{t+1} \equiv S_{t+1}(a_{<t+1}) \sim P_{t+1}(\cdot|S_t,a_t), \quad t = 0,\ldots,H-1.$$

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 random variables

$$\xi_t : (a_{<t},\omega) \in \mathsf{A}^t \times \Omega \to \mathbb{R}$$

satisfying

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

The next duality theorem, essentially due to Rogers [2007], may be seen as a generalization of the dual representation theorem for optimal stopping, developed independently in Rogers [2002] and Haugh and Kogan [2004], to Markov decision processes. For a more general dual representations in terms of information relaxation, see Brown et al. [2010]. Let us further mention dual representations in the context of multiple stopping developed in Schoenmakers [2012], Belomestny et al. [2009], and applications to flexible caps studied in Balder et al. [2013].

Theorem 3.1 *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*

$$V_0^{\mathrm{up}}(x;\boldsymbol{\xi}) := \mathbb{E}_{\boldsymbol{\pi},x}\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{<t}),a_t)-\xi_{t+1}(a_{<t+1})\right)+F(S_H(a_{<H}))\right)\right], \tag{3.1}$$

where as usual we suppress the dependence on ω for notational simplicity. Hence $V_0^{\mathrm{up}}(x;\boldsymbol{\xi})$ is an upper bound for $V_0^{\star}(x)$.

(ii) *If we set $\boldsymbol{\xi}^{\star} = (\xi_t^{\star},\, t\in[H]) \in \Xi$ with*

$$\xi_{t+1}^{\star}(a_{<t+1}) := V_{t+1}^{\star}(S_{t+1}(a_{<t+1})) - \mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t(a_{<t}),a_t)}\left[V_{t+1}^{\star}(S'_{t+1})\right], \tag{3.2}$$

for $t = 0,\ldots,H-1$, then, almost surely,

$$V_0^{\star}(x_0) = \sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\big(R_t(S_t(a_{<t}),a_t)-\xi_{t+1}^{\star}(a_{<t+1})\big)+F(S_H(a_{<H}))\right). \tag{3.3}$$

Remark 3.2 *In Theorem 3.1 and further below, supremum should be interpreted as essential supremum in case it concerns the supremum over an uncountable family of random variables.*

In principle Theorem 3.1 may be inferred from Rogers [2007] or Brown et al. [2010]. Nonetheless, also for the convenience of the reader, we here give a concise proof in terms of the present setup and terminology.

Proof. (i) Since for any $\boldsymbol{\xi} \in \Xi$ and policy $\boldsymbol{\pi}$ in (2.2) 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 (3.1) follows immediately.

(ii) We may write for any $a_{\geq h} \in \mathsf{A}^{H-j}$,

$$\begin{aligned}\sum_{t=0}^{H-1}&\big(R_t(S_t\left(a_{<t}\right),a_t)-\xi_{t+1}^{\star}(a_{\leq t})\big)+F(S_H(a_{<H}))\\ &= \sum_{t=0}^{H-1}R_t(S_t\left(a_{<t}\right),a_t)-\sum_{t=0}^{H-1}V_{t+1}^{\star}(S_{t+1}(a_{\leq t}))\\ &\quad+\sum_{t=0}^{H-1}\mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t(a_{<t}),a_t)}\left[V_{t+1}^{\star}(S'_{t+1})\right]+F(S_H(a_{<H})).\end{aligned}$$

Hence

$$\sum_{t=0}^{H-1}\left(R_t(S_t\left(a_{<t}\right),a_t)-\xi^{\star}_{t+1}(a_{\leq t})\right)+F(S_H(a_{<H}))=V^{\star}_0(x)+\Delta(x),$$

with

$$\Delta(x):=F(S_H(a_{<H}))-V^{\star}_H(S_H(a_{<H}))+\\ \sum_{t=0}^{H-1}\Big(R_t(S_t\left(a_{<t}\right),a_t)+\mathbb{E}_{S'_{t+1}\sim P_{t+1}(\cdot|S_t,a_t)}\left[V^{\star}_{t+1}(S'_{t+1})\right]-V^{\star}_t(S_t(a_{<t}))\Big)\geq 0$$

where the latter inequality follows from the Bellman principle, see Theorem 2.1. The statement (3.3) now follows by taking the (essential) supremum over $a_{\geq 0}\in \mathsf{A}^H$ on the left-hand-side, applying (3.1) and using the sandwich property. ■

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 3.1. However, these methods require as input a sequence of (approximate) value functions V_h, $h\in[H]$. Below we describe a regression-based regression algorithm for approximating the value functions $V^{\star}_h$, $h\in[H]$, backwardly in time. In fact, unlike the usual regression, the proposed algorithm is based on a kind of “pseudo” or “quasi” regression procedure with respect to some reference measure μ_h, which is assumed to be such that $P_h(\cdot|x,a)$ is absolutely continuous w.r.t. μ_h for any $h\in]H]$, $x\in\mathsf{S}$ and $a\in\mathsf{A}$. Furthermore, we consider a vector of basis functions

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

such that the matrix

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

is analytically known and invertible. The algorithm reads 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 now approximate $V^{\star}_h$ via simulating independent drwas $X_i\sim\mu_h$, $Y^a_i\sim P_{h+1}(\cdot|X_i,a)$, $a\in\mathsf{A}$, $i=1,\ldots,N$, and setting

$$V_{h,N}(x)=T_{h,N}V_{h+1,N}(x):=\sup_{a\in\mathsf{A}}(R_h(x,a)+\widetilde{P}^a_{h+1,N}V(x)),\tag{4.1}$$

where

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

Here $\widetilde{L}$ is a fixed positive constant which will be defined later,

$$\beta_{N,a}:=\frac{1}{N}\sum_{i=1}^{N}U^a_i,\quad U^a_i:=Z^a_i\Sigma^{-1}\boldsymbol{\gamma}_K(X_i),\quad Z^a_i:=V(Y^a_i),\quad i=1,\ldots,N.$$

Note that $\mathbb{E}\left[\beta_{N,a}\right]=\mathbb{E}\left[V(Y^a_1)\Sigma^{-1}\boldsymbol{\gamma}_K(X_1)\right]=:\beta_a$, where β_a solves the minimization problem

$$\inf_{\beta_a\in\mathbb{R}^K}\mathbb{E}\left[\left(V(Y^a_1)-\beta_a^{\top}\boldsymbol{\gamma}_K(X_1)\right)^2\right].$$

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

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

After H steps of the above procedure we obtain the estimates $V_{H,N}, \dots, V_{0,N}$.[3]

5 Dual regression algorithm

In this section we outline how to construct an upper biased estimate based on Theorem 3.1 from a given sequence of approximations V_t, $t \in [H]$ to $V^\star_t$, $t \in [H]$, obtained as described in Section 4, for example.

Theorem 3.1-(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

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

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 (3.2) 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

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

By generating a sample $Y^{x,a}_1, \dots, Y^{x,a}_N$ from $P_{t+1}(\cdot|x, a)$ we readily obtain a computable approximation of $m^\star_{t+1}(\cdot; x, a)$, that is, (5.3), by

$$\arg \inf_{m \in \mathcal{M}'_{t+1,x,a}} \left\{\frac{1}{N}\sum_{i=1}^{N}\left(V_{t+1}(Y^{x,a}_i) - m(Y^{x,a}_i)\right)^2\right\}, \tag{5.4}$$

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 $p_{t+1}(\cdot|x, a)$ doesn’t 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

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

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

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

[3] 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.

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 $\mathcal{M}'_{t+1,x,a} = \{m_{t+1,\phi}\left(\cdot;x,a\right) :\ \phi \in \Phi\}$ where Φ is the linear space of mappings $\mathbb{R}^d \to \mathbb{R}^d$, which are smooth, bounded, and with 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 $(m_{t+1,\varphi_k}\left(\cdot;x,a\right))_{k\in\mathbb{N}}$ 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, $(x_l,a_l)_{l=1,\ldots,L} \in \mathsf{S} \times \mathsf{A}$ say, yielding solutions $\phi_k(\cdot) := \phi(\cdot;x_k,a_k)$ and the corresponding martingale functions $m_{t+1,\phi_k}\left(\cdot;x_k,a_k\right)$. In order to obtain a martingale function $m_{t+1} \equiv m_{t+1}\left(\cdot;x,a\right)$ 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 $m_{t+1} = m_{t+1,\phi}\left(\cdot;x,a\right)$. 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

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

Another way of constructing $\boldsymbol{\xi} \in \Xi$ is to assume that the chain $(S_t(a_{<t}))$ can be constructed using the so-called *random iterative functions*:

$$S_t(a_{<t}) = \mathcal{K}_t(S_{t-1}(a_{<t-1}),a_{t-1},\varepsilon_t), \quad t \in]H], \tag{5.6}$$

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 i.i.d. sequence of E-valued random variables defined on a probability space $(\Omega,\mathcal{F},\mathrm{P})$. In this setup we may consider as the underlying probability space $\Omega := (\mathsf{E} \times \mathsf{A})^{[H]}$ instead of (2.1), with accordingly modified definitions of $\mathcal{F}$ and $(\mathcal{F}_t)$.

Let $\mathcal{P}_{\mathsf{E}}$ be the distribution of ε_1 on E, and assume that $(\psi_k,\ k \in \mathbb{N}_0)$ is a an orthonormal system in $L^2(\mathsf{E},\mathcal{P}_{\mathsf{E}})$ with $\psi_0 \equiv 1$, that is,

$$\int \psi_k(\varepsilon)\psi_{k'}(\varepsilon)\,d\mathcal{L}_{\mathsf{E}}(\varepsilon) = \delta_{kk'}.$$

By then letting

$$\eta_{t+1,K}(x,a) \equiv \eta_{t+1,K}(x,a,\varepsilon_{t+1}) = \sum_{k=1}^{K} c_k(x,a)\psi_k(\varepsilon_{t+1}) \tag{5.7}$$

for some natural $K > 0$ and “nice” functions $c_k : \mathsf{S} \times \mathsf{A} \to \mathbb{R}$, $k = 1,\ldots,K$, 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 consider the least-squares problem

$$\inf_{(c_1,\ldots,c_K)} \mathbb{E}\left[\left(V_{t+1}(Z^{x,a}) - \sum_{k=1}^{K} c_k\psi_k(\varepsilon_{t+1})\right)^2\right], \quad Z^{x,a} \equiv \mathcal{K}_{t+1}(x,a,\varepsilon_{t+1}), \tag{5.8}$$

for estimating the coefficients in (5.7). 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 $\boldsymbol{\psi}_K(\varepsilon) := [\psi_1(\varepsilon),\ldots,\psi_K(\varepsilon)]^{\top}$. The minimization problem (5.8) is then explicitly solved by

$$\bar{\mathbf{c}}_K(x,a) := \Sigma_{\mathsf{E},K}^{-1}\mathbb{E}\left[V_{t+1}(Z^{x,a})\boldsymbol{\psi}_K(\varepsilon)\right]. \tag{5.9}$$

In the sequel we assume that we know $\Sigma_{\mathsf{E},K}$. This assumption is not restrictive as we choose the basis $\boldsymbol{\psi}$ ourselves. In order to compute (5.9), 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 $\varepsilon_m \sim \mathcal{P}_\mathsf{E}$, $m = 1,\ldots,M$, and estimate its mean $\bar{\mathbf{c}}_K(x,a)$ by the empirical mean

$$\mathbf{c}_{K,M}(x,a) = [c_{1,M}(x,a),\ldots,c_{K,M}(x,a)]^\top := \frac{1}{M}\sum_{m=1}^{M} U_m(x,a). \tag{5.10}$$

We so obtain as martingale functions in (5.7),

$$\eta_{t+1,K,M} := \mathbf{c}_{K,M}^\top(x,a)\boldsymbol{\psi}_K(\varepsilon_{t+1}) = \sum_{k=1}^{K} c_{k,M}(x,a)\psi_k(\varepsilon_{t+1}). \tag{5.11}$$

Also the problem (5.8) may only numerically be solved on a grid in practice, and a suitable interpolation procedure is required to obtain (5.11) for generic $(x,a) \in \mathsf{S}\times\mathsf{A}$ (for details see Section 7). Finally, an upper biased upper bound for $V_0^\star(x)$, is obtained via an independent standard Monte Carlo estimate of the expectation

$$V_0^{\mathrm{up}}(x) = \mathbb{E}_{\boldsymbol{\pi},x}\left[\sup_{a_{\geq 0}\in\mathsf{A}^H}\left(\sum_{t=0}^{H-1}\left(R_t(S_t(a_{\geq t}),a_t) - \eta_{t+1,K,M}(S_t(a_{<t}),a_t)\right) + F(S_H)\right)\right]. \tag{5.12}$$

In Section 7 we will give a detailed convergence analysis of the dual estimator (5.12). 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 a convergence analysis of the primal algorithm designed in Section 4, under a few mild assumptions.

Assumption 6.1 *Assume that (5.6) holds, that is,*

$$Y^a = \mathcal{K}_h(X,a,\varepsilon_h), \quad h \in]H]. \tag{6.1}$$

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

$$|\mathcal{K}_h(x,a,\varepsilon) - \mathcal{K}_h(x',a',\varepsilon)| \leq L_\mathcal{K}\,\rho((x,a),(x',a')), \quad (x,a),(x',a')\in\mathsf{S}\times\mathsf{A}, \quad \varepsilon\in\mathsf{E}, \tag{6.2}$$

for some constant $L_\mathcal{K}$ *not depending on* h. *In (6.2), the metric* $\rho \equiv \rho_{\mathsf{S}\times\mathsf{A}}$ *on* $\mathsf{S}\times\mathsf{A}$ *is considered to be of the form*

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

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 6.2 *Assume that* $\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\{|R_h(x,a)| \vee |F(x)|\} \leq R_{\max}$ *and*

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

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

Assumption 6.3 *Assume that* $|\Sigma_{h,K}^{-1}\boldsymbol{\gamma}_K(x)|_\infty \leq \Lambda_K$ *for all* $x \in \mathsf{S}$, $h \in [H[$, *and*

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

for a constant $L_{\gamma,K} > 0$, *where* $|\cdot|$ *denotes the Euclidian norm.*

Note that under Assumptions 6.1, 6.2 and 6.3,

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

We now specify $\widetilde{L} := V^\star_{\max} - R_{\max} := HR_{\max}$, and 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

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

The next assumption concerns the measures $\mu_1,\ldots,\mu_H$.

Assumption 6.4 *Consider for any* $h < l$ *the Radon-Nikodym derivative*

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

where for a generic policy $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_H)$,

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

Assume that

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

The following theorem provides an upper bound for the difference between $V_{h,N}$ and $V^\star_h$.

Theorem 6.5 *Suppose that* $\mathbb{E}_{X\sim\mu_h}[|\boldsymbol{\gamma}_K(X)|^2] \leq \varrho^2_{\gamma,K}$ *for all* $h \in [H[$. *Then for* $h \in [H]$,

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

where $\lesssim$ *denotes* $\leq$ *up to a natural 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*

$$\mathcal{R}_{K,h} := \sup_{\boldsymbol{\zeta}\in\mathbb{R}^{K\times|\mathsf{A}|}}\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}\left(\beta^\top_{a,\boldsymbol{\zeta}}\boldsymbol{\gamma}_K(X) - P^a_{h+1}V_{h+1,\boldsymbol{\zeta}}(X)\right)^2\right]^{1/2},$$

where

$$\beta_{a,\boldsymbol{\zeta}} := \arg\min_{\beta\in\mathbb{R}^K}\mathbb{E}_{X\sim\mu_h}\left[\left(\beta^\top\boldsymbol{\gamma}_K(X) - P^a_{h+1}V_{h+1,\boldsymbol{\zeta}}(X)\right)^2\right]$$

and

$$V_{t,\boldsymbol{\zeta}}(x) := \sup_{a\in\mathsf{A}}\bigl(R_t(x,a) + \mathcal{T}_{\widetilde{L}}[\zeta^\top_a\boldsymbol{\gamma}_K(x)]\bigr) \text{ for } 0 \leq t < H, \quad V_{H,\boldsymbol{\zeta}}(x) := F(x).$$

Discussion

- The quantities $\mathcal{R}_{K,h}$ is the error of approximating the conditional expectation $P^a_{h+1}V_{h+1,\zeta}$ via a linear combination of $\gamma_1,\ldots,\gamma_K$ in a worst case scenario, that is, for the most unfavourable choice of $\boldsymbol{\zeta}$. Note that if $\gamma_1,\gamma_2,\ldots$ are bounded eigenfunction (corresponding to nonnegative eigenvalues) of the kernel P^a_{h+1} not depending on $a\in\mathsf{A}$, $F(x)=\beta^\top\boldsymbol{\gamma}_K(x)\geq 0$ for some $\beta\in\mathbb{R}^K$ and $R_t(x,a)=R_{1,t}(x)R_{2,t}(a)$ with $R_{1,t}(x)=c_t^\top\boldsymbol{\gamma}_K(x)\geq 0,$ then for $\widetilde{L}$ large enough, $\mathcal{R}_{K,h}=0$ (in this case we can take ζ_a independent of a in the definition of $V_{h+1,\zeta}$) and only the stochastic part of the error remains
$$\|V^\star_h-V_{h,N}\|_{L^2(\mu)}\lesssim H\mathfrak{R}^{\max}\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A})+V^\star_{\max})\sqrt{\frac{K}{N}}.$$
If, for example, $\mathsf{A}=[0,1]^{d_\mathsf{A}}$ for some $d_\mathsf{A}\in\mathbb{N}$, then $\mathsf{D}(\mathsf{A})=2\sqrt{d_\mathsf{A}}$ and $I_{\mathcal{D}}(\mathsf{A})\lesssim\sqrt{d_\mathsf{A}}$. This example shows that the bound depends sub-linearly in d_A.

- Let us remark on the assumption 6.4. 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,
$$P^{\pi_h}_{h+1}\cdot\ldots\cdot P^{\pi_{l-1}}_l(dy|x)=p^{\pi_h}_{h+1}\cdot\ldots\cdot p^{\pi_{l-1}}_l(y|x)\,dy.$$
Further assume that
$$\sup_{0\leq h<l<H,\boldsymbol{\pi}}p^{\pi_h}_{h+1}\ldots p^{\pi_{l-1}}_l(y|x)\leq Ce^{-\mathrm{c}|y-x|^2}\text{ for some }C,c>0,$$
and consider absolutely continuous reference measures $\mu_h(dx)=\mu_h(x)\,dx$. For the bound (6.4), we then have
$$\begin{aligned}(\mathfrak{R}^{\max})^2&=\sup_{0\leq h<l<H,\boldsymbol{\pi}}\int\int\frac{\mu_h(x)}{\mu_l(y)}\left(P^{\pi_h}_{h+1}\ldots P^{\pi_{l-1}}_l(y|x)\right)^2dxdy\\&\leq C^2\max_{0\leq h<l<H}\int\int\frac{\mu_h(x)}{\mu_l(x+u)}e^{-2\mathrm{c}|u|^2}dxdu.\end{aligned}$$
The latter expression can be easily bounded by choosing μ_h to be Gaussian with an appropriate covariance structure depending on h. For example, take $d=1$ and
$$\mu_h(x)=\sqrt{\frac{c}{\pi(h+1)}}e^{-\frac{c}{h+1}x^2},\quad h\in[H[,$$
then straightforward calculations yield
$$\mathfrak{R}^{\max}\leq C\sqrt{\max_{0\leq h<l<H}\frac{(l+1)\pi}{c\sqrt{2(l-h)-1}}}\leq C\sqrt{\frac{H\pi}{c}}.$$
In this case the bound of Theorem 6.5 may grow in H as $H^{3/2}$ (provided that all errors $\mathcal{R}_{K,h}$, $h\in[H]$, are bounded) as opposed to the 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.

Proof. *One-step analysis:* Suppose that after h steps of the algorithm the estimates $V_{H,N},\ldots,V_{h+1,N}$ of the value functions $V^\star_H,\ldots,V^\star_{h+1}$, respectively, are constructed using sampled data $\mathcal{D}_{N,h+1}$, such that $\|V_{t,N}\|_\infty\leq V^\star_{\max}$ a.s. for all $t=h+1,\ldots,H$. Denote for $a\in\mathsf{A}$,
$$\ell^a(\beta):=\mathbb{E}_{X\sim\mu_h}\left[(Z^a-\beta^\top\boldsymbol{\gamma}_K(X))^2\right],\quad Z^a\sim V_{h+1,N}(Y^{a,X}),\quad Y^{a,X}\sim P_{h+1}(\cdot|X,a).$$

The unique minimizer of $\ell^a(\beta)$ is given by $\beta_a := \mathbb{E}\left[Z^a\Sigma^{-1}\boldsymbol{\gamma}_K(X)\right]$ and we have

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

It then holds that

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

where according to Proposition B.1, (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 $k=1,\ldots,K$,

$$\mathbb{E}\left[\sup_{a\in\mathsf{A}}\left(\beta_{N,a,k}-\beta_{a,k}\right)^2\right] \lesssim \frac{(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A})+V^\star_{\max})^2\Lambda^2_K}{N}. \tag{6.5}$$

Due to the very structure of $V_{h+1,N}$ (see (4.1)), we have

$$\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta^\top_a\boldsymbol{\gamma}(X)-P^a_{h+1}V_{h+1,N}(X))^2\right] \le \mathcal{R}^2_{K,h}, \tag{6.6}$$

and we then get

$$\begin{aligned}\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\widetilde{P}^a_{h+1,N}V_{h+1,N}(X)-P^a_{h+1}V_{h+1,N}(X))^2\right]^{1/2} &\le \mathcal{R}_{K,h}+\\ &\varrho_{\gamma,K}\Lambda_K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{A})+V^\star_{\max})\sqrt{\frac{K}{N}}\end{aligned} \tag{6.7}$$

due to (6.5), (6.6), and the estimate

$$\begin{aligned}&\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\widetilde{P}^a_{h+1,N}V_{h+1,N}(X)-P^a_{h+1}V_{h+1,N}(X))^2\right]^{1/2} \le\\ &\qquad\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta^\top_{N,a}\boldsymbol{\gamma}_K(X)-P^a_{h+1}V_{h+1,N}(X))^2\right]^{1/2} \le\\ &\qquad\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta^\top_{N,a}\boldsymbol{\gamma}_K(X)-\beta^\top_a\boldsymbol{\gamma}_K(X))^2\right]^{1/2}\\ &\qquad+\mathbb{E}_{X\sim\mu_h}\left[\sup_{a\in\mathsf{A}}(\beta^\top_a\boldsymbol{\gamma}_K(X)-P^a_{h+1}V_{h+1,N}(X))^2\right]^{1/2}.\end{aligned}$$

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

$$\Delta^a_{h,N}(x) := \widetilde{P}^a_{h+1,N}V_{h+1,N}(x)-P^a_{h+1}V_{h+1,N}(x) \quad\text{and}\quad \Delta_h(x) := \sup_{a\in\mathsf{A}}|\Delta^a_{h,N}(x)|. \tag{6.8}$$

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

$$\begin{aligned}V_h^\star(x)-V_{h,N}(x)&=\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+P_{h+1}^aV_{h+1}^\star(x)\right\}-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}_{h+1,N}^aV_{h+1,N}(x)\right\}\\&=R_h(x,\pi_h^\star(x))+\int V_{h+1}^\star(x')P_{h+1}(dx'|x,\pi_h^\star(x))\\&-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}_{h+1,N}^aV_{h+1,N}(x)\right\}\\&\le\int\left(V_{h+1}^\star-V_{h+1,N}\right)(x')P_{h+1}(dx'|x,\pi_h^\star(x))\\&+\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+P_{h+1}^aV_{h+1,N}(x)\right\}-\sup_{a\in\mathsf{A}}\left\{R_h(x,a)+\widetilde{P}_{h+1,N}^aV_{h+1,N}(x)\right\}\\&\le P_{h+1}^{\pi_h^\star}\left(V_{h+1}^\star-V_{h+1,N}\right)(x)+\ \Delta_h(x)\end{aligned}\tag{6.9}$$

and analogously,

$$V_h^\star(x)-V_{h,N}(x)\ge P_{h+1}^{\pi_{h,N}}[V_{h+1}^\star-V_{h+1,N}](x)-\Delta_h(x).\tag{6.10}$$

By iterating (6.9) and (6.10) upwards, and using that $V_{H,N}=V_H^\star$, we obtain, respectively,

$$\begin{aligned}V_h^\star(x)-V_{h,N}(x)&\le\sum_{k=1}^{H-h-1}P_{h+1}^{\pi_h^\star}\dots P_{h+k}^{\pi_{h+k-1}^\star}[\Delta_{h+k}](x)+\Delta_h(x),\ \text{ and}\\V_h^\star(x)-V_{h,N}(x)&\ge-\sum_{k=1}^{H-h-1}P_{h+1}^{\pi_{h,N}}\dots P_{h+k}^{\pi_{h+k-1,N}}[\Delta_{h+k}](x)-\Delta_h(x).\end{aligned}$$

We thus have pointwise,

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

which implies

$$\|V_h^\star-V_{h,N}\|_{L^2(\mu_h)}\le2\sup_{\boldsymbol{\pi}}\sum_{k=1}^{H-h-1}\left\|P_{h+1}^{\pi_h}\dots P_{h+k}^{\pi_{h+k-1}}[\Delta_{h+k}]\right\|_{L^2(\mu_h)}+\|\Delta_h\|_{L^2(\mu_h)}\,.$$

Hence we have

$$\|V_h^\star-V_{h,N}\|_{L^2(\mu_h)}\le2\mathfrak{R}^{\max}\sum_{l=h}^{H-1}\|\Delta_l\|_{L^2(\mu_l)}$$

(note that $\mathfrak{R}^{\max}\ge1$), and then, by the definitions (6.8) and the estimate (6.7), the statement of the theorem follows. ■

7 Convergence analysis of the dual algorithm

7.1 Convergence of martingale functions

For the dual representation (5.12) we construct an H-tuple of martingale functions $\widetilde{\boldsymbol{\eta}}:=(\widetilde{\eta}_{t+1,K,M}(x,a),\ t\in[H[)$, see for instance (5.11), from a given pre-computed H-tuple of approximate value functions $(V_{t+1,N},\ t\in[H[)$, based on sampled data, denoted by $\mathcal{D}_N$, as outlined in Section 5.

Let us consider a fixed time $t \in [H[$ and suppress time subscripts where notationally convenient. We fix two (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. 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

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

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 Beliakov [2006]. The so constructed interpolant achieves the bound

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

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

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

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

$$\inf_{c_1,\ldots,c_K} \mathbb{E}_{\varepsilon \sim \mathcal{P}_{\mathsf{E}}} \left[\left(V_{t+1}^{\star}(\mathcal{K}_{t+1}(x,a,\varepsilon)) - \sum_{k=1}^{K} c_k \psi_k(\varepsilon) \right)^2 \right] \tag{7.2}$$

for any $(x,a) \in \mathsf{S} \times \mathsf{A}$, and define $\eta_{t+1,K}(x,a) := \mathbf{c}_K^\top(x,a)\boldsymbol{\psi}_K(\varepsilon)$. Let us make a few of assumptions.

Assumption 7.1 *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 of the function $V_{t+1}^{\star}(\mathcal{K}_{t+1}(x,a,\cdot))$ on $\mathrm{span}(\psi_1, \ldots, \psi_K)$ and its estimate $\widetilde{\eta}_{t+1,K,M}$.

Theorem 7.2 *Under Assumptions 6.1, 6.2, 6.3, and 7.1 it holds that*

$$\mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}|\eta_{t+1,K}(x,a)-\widetilde{\eta}_{t+1,K,M}(x,a)|^2\right]\lesssim$$
$$\varrho^2_{\psi,K}\frac{K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{S}\times\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{S}\times\mathsf{A})+V^\star_{\max})^2\Lambda^2_{\mathsf{E},K}}{M}$$
$$+K\Lambda^2_{\mathsf{E},K}\varrho^2_{\psi,K}\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_\infty\|V^\star_{t+1}-V_{t+1,N}\|^2_{L^2(\mu_{t+1})}$$
$$+K\varrho^2_{\psi,K}L^2_{V,K}L^2_{\mathcal{K}}\Lambda^2_{\mathsf{E},K}\rho^2_L(\mathsf{S},\mathsf{A}),$$

where $\lesssim$ denotes $\leq$ up to a natural constant, the constants $L_{V,K}$, $L_{\mathcal{K}}$, and the measure μ_{t+1} are defined in Section 6.

Let us now consider the approximation error

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

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

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

where $(\psi_k)_{k\in\mathbb{N}}$ is a an orthonormal system in $L^2(\mathsf{E},\mathcal{P}_{\mathsf{E}})$ with $\psi_0\equiv 1$. If $\|\psi_k\|_\infty\leq\psi^*_k$ for all $k\in\mathbb{N}$, then

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

If

$$\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\sum_{k=1}^{\infty}k^\beta|c^\star_{k,t+1}(x,a)|\psi^*_k\leq C<\infty \tag{7.3}$$

for some $\beta>0$, then $\mathcal{E}^2_{K,t}\leq C^2K^{-2\beta}$.

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 Reznikov and Saff [2016] it can be shown that

$$[\mathbb{E}\rho^p_L(\mathsf{S}\times\mathsf{A})]^{1/p}\lesssim\sqrt{d_{\mathsf{S}}}\left(\frac{p\log L}{L}\right)^{1/d_{\mathsf{S}}}+\sqrt{d_{\mathsf{A}}}\left(\frac{p\log L}{L}\right)^{1/d_{\mathsf{A}}}, \tag{7.4}$$

  where $\lesssim$ stands for inequality up to a constant not depending on L.

Proof. For the unique minimizer of (7.2) one has that

$$\mathbf{c}_K(x,a) := \Sigma_{\mathsf{E},K}^{-1}\,\mathbb{E}\left[V_{t+1}^{\star}(\mathcal{K}_{t+1}(x,a,\varepsilon))\boldsymbol{\psi}_K(\varepsilon)\right]. \tag{7.5}$$

Likewise, the unique minimizer of the problem

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

is given by

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

Now let $\mathbf{c}_{K,N}(x,a)$ be the Monte Carlo estimate of $\bar{\mathbf{c}}_K(x,a)$ as constructed in Section 5, see (5.9) and (5.10). We then have

$$\begin{aligned}\mathbb{E}&\left[\left|\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}(\mathbf{c}_{K,N}-\bar{\mathbf{c}}_K)^{\top}(x,a)\boldsymbol{\psi}_K(\varepsilon)\right|^2|\mathcal{D}_N\right]\\&\leq\mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|(\mathbf{c}_{N,K}-\bar{\mathbf{c}}_K)^{\top}(x,a)\right|^2|\mathcal{D}_N\right]\mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}\left[|\boldsymbol{\psi}_K(\varepsilon)|^2\right],\end{aligned} \tag{7.6}$$

where according to Proposition B.1 (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))

$$\mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}|(\mathbf{c}_{K,N}-\bar{\mathbf{c}}_K)(x,a)|^2\right]\leq\frac{K(L_{V,K}L_{\mathcal{K}}I_{\mathcal{D}}(\mathsf{S}\times\mathsf{A})+L_{V,K}L_{\mathcal{K}}\mathsf{D}(\mathsf{S}\times\mathsf{A})+V_{\max}^{\star})^2\Lambda_{\mathsf{E},K}^2}{M}. \tag{7.7}$$

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

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

we have

$$\begin{aligned}\mathbb{E}&\left[\left|\max_{(x,a)\in\mathsf{S}_L\times\mathsf{A}_L}(\mathbf{c}_K-\bar{\mathbf{c}}_K)^{\top}(x,a)\boldsymbol{\psi}_K(\varepsilon)\right|^2\right]\\&\leq\mathbb{E}\left[\max_{(x,a)\in\mathsf{S}_L\times\mathsf{A}_L}|(\mathbf{c}_K-\bar{\mathbf{c}}_K)(x,a)|^2\right]\mathbb{E}_{\varepsilon\sim\mathcal{P}_{\mathsf{E}}}\left[|\boldsymbol{\psi}_K(\varepsilon)|^2\right]\\&\leq K\varrho_{\psi,K}^2\Lambda_{\mathsf{E},K}^2\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left\|\frac{dP_{t+1}(\cdot|x,a)}{d\mu_{t+1}(\cdot)}\right\|_{\infty}\|V_{t+1}^{\star}-V_{t+1,N}\|_{L^2(\mu_{t+1})}^2.\end{aligned} \tag{7.8}$$

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

$$\begin{aligned}&|c_{k,M}(x,a)-c_{k,M}(x',a')|\\&\quad\leq\frac{1}{M}\sum_{m=1}^{M}|V_{t+1,N}(\mathcal{K}_{t+1}(x,a,\varepsilon_m))-V_{t+1,N}(\mathcal{K}_{t+1}(x',a',\varepsilon_m))||\Sigma_{\mathsf{E},K}^{-1}\boldsymbol{\psi}_K(\varepsilon_m)|_{\infty}\\&\quad\leq L_{V,K}L_{\mathcal{K}}\Lambda_{\mathsf{E},K}\rho((x,a),(x',a'))\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta_{t+1,K,N}(x,a)-\widetilde{\eta}_{t+1,K,N}(x,a)\right|^{2}\right]\\
&=\mathbb{E}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\left(\mathbf{c}_{K,N}-I\left[\mathbf{c}_{K,N}\right]\right)^{\top}(x,a)\boldsymbol{\psi}_{K}(\varepsilon_{t+1})\right|^{2}\right]\\
&\leq\varrho_{\psi,K}^{2}\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\left(\mathbf{c}_{N,K}-I\left[\mathbf{c}_{N,K}\right]\right)(x,a)\right|^{2}\\
&=\varrho_{\psi,K}^{2}\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\sum_{k=1}^{K}\left(c_{k,N}-I\left[c_{k,N}\right]\right)^{2}(x,a)\\
&\leq K\varrho_{\psi,K}^{2}L_{V,K}^{2}L_{\mathcal{K}}^{2}\Lambda_{\mathsf{E},K}^{2}\rho_{L}^{2}(\mathsf{S},\mathsf{A}),
\end{aligned}\tag{7.9}$$

using (7.1). Finally note that

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

and then the result follows by the triangle inequality and gathering (7.6)–(7.9). ■

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 $\eta^{\star}=(\eta_{t}^{\star}(x,a),\ t\in]H])$ are constructed based on the sampled data $\mathcal{D}_{M,N}$ such that Theorem 7.2 holds. Consider for $\widetilde{\boldsymbol{\xi}}:=(\widetilde{\eta}_{t+1}(S_{t}(a_{<t}),a_{t}),\ t\in[H[)\in\Xi$, the upper bias

$$\begin{aligned}
V_{0}^{\mathrm{up}}(x;\widetilde{\boldsymbol{\xi}})-V_{0}^{\star}(x)&=\mathbb{E}_{x}\left[\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left(\sum_{t=0}^{H-1}\left(R_{t}(S_{t}(a_{<t}),a_{t})-\widetilde{\eta}_{t+1}(S_{t}(a_{<t}),a_{t})\right)+F(S_{H})\right)\right]\\
&\quad-\mathbb{E}_{x}\left[\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left(\sum_{t=0}^{H-1}\left(R_{t}(S_{t}(a_{<t}),a_{t})-\eta_{t+1}^{\star}(S_{t}(a_{<t}),a_{t})\right)+F(S_{H})\right)\right]\\
&\leq\mathbb{E}_{x}\left[\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left|\sum_{t=0}^{H-1}\eta_{t+1}^{\star}(S_{t}(a_{<t}),a_{t})-\sum_{t=0}^{H-1}\widetilde{\eta}_{t+1}(S_{t}(a_{<t}),a_{t})\right|\right]\\
&\leq\sum_{t=0}^{H-1}\mathbb{E}_{x}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta_{t+1}^{\star}(x,a)-\widetilde{\eta}_{t+1}(x,a)\right|\right]\\
&\leq\sum_{t=0}^{H-1}\mathbb{E}_{x}\left[\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta_{t+1}^{\star}(x,a)-\widetilde{\eta}_{t+1}(x,a)\right|^{2}\right]^{1/2}.
\end{aligned}$$

Furthermore, similarly,

$$\begin{aligned}
&\operatorname{Var}\left[\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left(\sum_{t=0}^{H-1}\left(R_{t}(S_{t}(a_{<t}),a_{t})-\widetilde{\eta}_{t+1}(S_{t}(a_{<t}),a_{t})\right)+F(S_{H})\right)\right]\\
&\quad=\operatorname{Var}\left[\begin{array}{c}\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left(\sum_{t=0}^{H-1}\left(R_{t}(S_{t}(a_{<t}),a_{t})-\widetilde{\eta}_{t+1}(S_{t}(a_{<t}),a_{t})\right)+F(S_{H})\right)\\-\sup_{a_{\geq0}\in\mathsf{A}^{H}}\left(\sum_{t=0}^{H-1}\left(R_{t}(S_{t}(a_{<t}),a_{t})-\eta_{t+1}^{\star}(S_{t}(a_{<t}),a_{t})\right)+F(S_{H})\right)\end{array}\right]\\
&\qquad\qquad\leq\mathbb{E}\left[\left(\sum_{t=0}^{H-1}\sup_{(x,a)\in\mathsf{S}\times\mathsf{A}}\left|\eta_{t+1}^{\star}(x,a)-\widetilde{\eta}_{t+1}(x,a)\right|\right)^{2}\right].
\end{aligned}$$

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

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

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

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

with

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

we obtain

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

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 $\|\eta\|_{\varphi,2}:=\inf\{t>0:\mathbb{E}\left[\exp\left(\eta^2/t^2\right)\right]\leq 2\}$. We say that η is *sub-Gaussian* if $\|\eta\|_{\varphi,2}<\infty$. In particular, this implies that for some constants $C,c>0$,

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

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 $K\geq 0$ such that

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

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

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

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

B Estimation of mean uniformly in parameter

The following proposition holds.

Proposition B.1 *Let f be a function on $\mathsf{X} \times \Xi$ such that*

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

with some constant $L > 0$. Furthermore assume that $\|f\|_\infty \leq F < \infty$ for some $F > 0$. Let ξ_n, $n = 1,\ldots,N$, 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 $\leq$ 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 $n = 1,\ldots,N$. 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 [Vershynin, 2018, 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,

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

By the Dudley integral inequality, e.g. [Vershynin, 2018, 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. [Vershynin, 2018, Theorem 2.6.2.], for any $\delta \in (0,1)$,

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

■

References

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

A. Antos, R. Munos, and C. Szepesvári. Fitted q-iteration in continuous action-space mdps. 2007.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. 70:263–272, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/azar17a.html.

S. Balder, A. Mahayni, and J. Schoenmakers. Primal-dual linear Monte Carlo algorithm for multiple stopping—an application to flexible caps. *Quant. Finance*, 13(7):1003–1013, 2013. ISSN 1469-7688. doi: 10.1080/14697688.2013.775476. URL http://dx.doi.org/10.1080/14697688.2013.775476.

C. Bayer, M. Redmann, and J. Schoenmakers. Dynamic programming for optimal stopping via pseudo-regression. *Quant. Finance*, 21(1):29–44, 2021.

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

D. Belomestny, C. Bender, and J. Schoenmakers. True upper bounds for bermudan products via non-nested Monte Carlo. *Math. Finance*, 19(1):53–71, 2009. ISSN 0960-1627. doi: 10.1111/j.1467-9965.2008.00357.x. URL citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.404.8521.

D. Belomestny, A. Kolodko, and J. Schoenmakers. Regression methods for stochastic control problems and their convergence analysis. *SIAM J. Control Optim.*, 48(5):3562–3588, 2010.

C. Bender, J. Schoenmakers, and J. Zhang. Dual representations for general multiple stopping problems. *Math. Finance*, 25(2):339–370, 2015. ISSN 0960-1627. doi: 10.1111/mafi.12030. URL citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.757.1137.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. ISBN Athena Scientific. URL http://www.athenasc.com/ndpbook.html.

D. B. Brown, J. E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Oper. Res.*, 58(4, part 1):785–801, 2010.

D. B. Brown, J. 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.

V. V. Desai, V. F. Farias, and C. C. Moallemi. Bounds for markov decision processes. *Reinforcement learning and approximate dynamic programming for feedback control*, pages 452–473, 2012.

M. Haugh and L. Kogan. Pricing American options: A duality approach. *Oper. Res.*, 52(2):258–270, 2004.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. 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. URL https://proceedings.neurips.cc/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

B. Á. Pires and C. Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151. PMLR, 2016.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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. ISSN 1073-7928. doi: 10.1093/imrn/rnv342. URL https://doi.org/10.1093/imrn/rnv342.

L. Rogers. Pathwise stochastic optimal control. *SIAM J. Control and Optimization*, 46:1116–1132, 01 2007. doi: 10.1137/050642885.

L. C. G. Rogers. Monte Carlo valuation of American options. *Mathematical Finance*, 12(3):271–286, 2002.

J. Schoenmakers. A pure martingale dual for multiple stopping. *Finance Stoch.*, 16:319–334, 2012.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

C. Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

R. Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596. An introduction with applications in data science, With a foreword by Sara van de Geer.

D. Z. Zanger. Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing. *Finance and Stochastics*, 17(3):503–534, 2013. ISSN 0949-2984. doi: 10.1007/s00780-013-0204-9. URL https://doi.org/10.1007/s00780-013-0204-9.

H. Zhu, F. Ye, and E. Zhou. Solving the dual problems of dynamic programs via regression. *IEEE Transactions on Automatic Control*, 63(5):1340–1355, 2017.