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

Preprint ISSN 2198-5855

Weighted mesh algorithms for general Markov decision processes: Convergence and tractability

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

submitted: January 25, 2024

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

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

No. 3088
Berlin 2024

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

Key words and phrases. Markov decision processes, mesh method, tractability.

D.B. and V.Z. gratefully acknowledge financial support from the German science foundation (DFG), grant Nr.497300407. J.S. 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/

Weighted mesh algorithms for general Markov decision processes: Convergence and tractability

Denis Belomestny, John G. M. Schoenmakers, Veronika Zorina

Abstract

We introduce a mesh-type approach for tackling discrete-time, finite-horizon Markov Decision Processes (MDPs) characterized by state and action spaces that are general, encompassing both finite and infinite (yet suitably regular) subsets of Euclidean space. In particular, for bounded state and action spaces, our algorithm achieves a computational complexity that is tractable in the sense of Novak & Woźniakowski [9], and is polynomial in the time horizon. For unbounded state space the algorithm is "semi-tractable" in the sense that the complexity is proportional to ϵ^{-c} with some dimension independent $c \geq 2$, for achieving an accuracy ϵ, and polynomial in the time horizon with degree linear in the underlying dimension. As such the proposed approach has some flavor of the randomization method by Rust [11] which deals with infinite horizon MDPs and uniform sampling in compact state space. However, the present approach is essentially different due to the finite horizon and a simulation procedure due to general transition distributions, and more general in the sense that it encompasses unbounded state space. To demonstrate the effectiveness of our algorithm, we provide illustrations based on Linear-Quadratic Gaussian (LQG) control problems.

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 [10], and for MDPs with application to finance, see [1]. 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 an approximate dynamic programming approach, see [10]; however, in practice, this approach suffers from the so-called "curse of dimensionality" and the "curse of horizon" meaning that the complexity (running time) of the program increases exponentially in the dimension of the problem (dimensions of the state and action spaces) and the horizon (or effective horizon for discounted infinite horizon MDPs). Traditional dynamic programming (DP) algorithms, such as value- or policy-iteration, exhibit exponential scaling with MDP size, even when coupled with advanced multigrid algorithms, see [4] for recent review of various approximative DP algorithms for general state and action spaces. Furthermore, the curse of dimensionality can be considered as a lower bound on the complexity of any MDP, not confined to any specific algorithm, as evidenced by Chow and Tsitsiklis [6].

The problem of solving MDPs without curse of dimensionality attracted a lot of attention in the literature. The first work in this direction was Rust [11] where the author proposed in an infinite horizon setting a weighted mesh algorithm with complexity proportional to ϵ^{-4} for a target accuracy ϵ and a polynomial in the underlying dimension. Another approach based on Monte Carlo tree search and sparse sampling was suggested in Kearns et al. [8]. In particular, the authors in [8] demonstrated that a specific online tree-building algorithm successfully circumvents the curse of dimensionality in discounted MDPs. This achievement has been further extended to partially observable MDPs (POMDPs) by the same authors in [7]. The bounds established

DOI 10.20347/WIAS.PREPRINT.3088

in these two papers remain independent of the dimension of the state space but exhibit exponential scaling with $1/(1-\gamma)$, representing the effective horizon-time, where γ is the discount factor of the MDP. Moreover, the complexity depends on the number of actions polynomially with power again proportional to the effective horizon-time. A recent work [2] proposed a nonlinear Multilevel Monte Carlo approach to solve infinite horizon MDPs without curse of dimensionality. Note that the complexity estimates in [2] is of order ϵ^{-c} with c depending on the effective horizon-time. Moreover, the number of actions is assumed to be finite. Let us stress that the setting of finite horizon MDPs is essentially different from the infinite horizon one where we need to solve a fixed point problem. Finite horizon setting requires a backward dynamic programming procedure and simulation of the paths of the underlying Markov process instead of the one step transitions as in the infinite horizon MDPs. As a result, the convergence analysis of the algorithms in finite horizon MDPs becomes much more intricate.

In this paper, we present a novel approach for addressing high-dimensional finite horizon Markov Decision Processes (MDPs) using a weighted mesh approach. This methodology shares conceptual similarities with the approach proposed by Rust [11] (see also [12]). However, it's essential to note that the work by Rust focuses on infinite horizon discounted MDPs, introducing a crucial distinction between the two settings. Unlike Rust, who can independently sample at each step of the iteration procedure, our approach involves drawing trajectories of the underlying state process and proceeding backwardly. This results in a more intricate structure of weights and their dependence on controls. Additionally, Rust's work imposes rather restrictive assumptions on the underlying MDP, assuming, for instance, a compact state space, finite action space and transition densities uniformly bounded away from zero. These assumptions exclude consideration of many interesting cases, such as Gaussian processes with non-compact supports (refer to [5] for a discussion on this and related issues). In our work, we allow for non-compact state spaces, continuous action spaces and general classes of transition densities.

This paper's primary contribution is the introduction of a new weighted mesh algorithm designed for a broad range of finite horizon Markov Decision Processes (MDPs). We have also conducted a thorough complexity analysis of this algorithm. Our findings reveal that this algorithm is capable of efficiently solving a wide spectrum of finite horizon MDPs, including those with non-compact state/action spaces that are subsets of $\mathbb{R}^d$ and feature general transition densities. Significantly, the computational complexity of our algorithm, denoted as $\mathcal{C}(\epsilon, d)$, demonstrates a polynomial dependence on the horizon length, ensuring ϵ-accuracy in approximating the corresponding value functions at a given point. Moreover, it holds

$$\lim_{d\to\infty}\lim_{\epsilon\searrow 0}\frac{\log\mathcal{C}\left(\epsilon,d\right)}{f(d)\log(1/\epsilon)}=0$$

for any f with arbitrary slow convergence to infinity as $d\to\infty$. This type of dependency on d and ϵ can be characterized as "semi-tractable" or indicative of a "weak curse of dimensionality." To our knowledge, this marks the first instance in the literature where a general finite horizon Markov Decision Process (MDP) is approximated with an algorithm that exhibits at most a "semi-tractable" level of complexity.

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. 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 (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 $(\mathsf{S}, \mathcal{S}, \rho_{\mathsf{S}})$;
- a measurable action space $(\mathsf{A}, \mathcal{A}, \rho_{\mathsf{A}})$;
- 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 $\mathfrak{S} := \big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [H]}, \mathbb{P}\big)$. 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 $S_0 \in \mathsf{S}$, $A_0 \sim \pi_0(S_0)$, and

$$S_{t+1} \sim P_{t+1}(\cdot | \, S_t, A_t), \quad A_t \sim \pi_t(S_t), \quad t = 0, \ldots, H-1. \tag{2.1}$$

Assumption 2.1 *In the sequel, we shall assume that chain $(S_t(a_{<t}))$ comes from the system of so-called random iterative functions:*

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

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 i.i.d. sequence of E-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$.

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

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

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:

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

where Π is a set of all measurable mappings $(\mathsf{S} \to \mathcal{P}(\mathsf{A}))^{\otimes H}$. 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

$$V_h^{\star}(x) := \sup_{\boldsymbol{\pi}} V_h^{\boldsymbol{\pi}}(x) \tag{2.3}$$

be the optimal value function at $h \in [H]$. It is well known that under weak conditions, there exists an optimal policy solving (2.3) 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 [10].

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

Theorem 2.2 *Let $x \in \mathsf{S}$ be fixed. It holds $V_H^\star(x) = F(x)$, and*

$$V_h^\star(x) = \sup_{a\in A} \left(R_h(x,a) + \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V_{h+1}^\star(S_{h+1})\right]\right), \quad h = H-1,\ldots,0. \tag{2.4}$$

Furthermore, if R_h is continuous and the action space is compact, the supremum in (2.4) 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

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

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 2.2,

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

where $T_h V(x) := \sup_{a\in A}\left(R_h(x,a) + P_{h+1}^a V(x)\right)$ with $P_{h+1}^a V(x) := \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V(S_{h+1})\right]$. Let us denote by $a_{<h}$ the deterministic vector of actions $a_{<h} = (a_0,\ldots,a_{h-1}) \in \mathsf{A}^h$, similarly $a_{\le h}$ etc., and denote with $S_h \equiv (S_h(a_{<h}))_{h\in\{0,\ldots,H\}}$ the process defined (in distribution) via

$$S_0 = x_0, \quad S_{h+1} \equiv S_{h+1}(a_{<h+1}) \sim P_{h+1}(\cdot|S_h,a_h), \quad h = 0,\ldots,H-1. \tag{2.5}$$

Suppose that the distribution $P_{h+1}(\cdot|S_h,a_h)$ possesses a Lebesgue density $p_{h+1}^{a_h}(\cdot|S_h)$, then the (unconditional) density of S_h denoted by $p_h^{a_{<h}}$ fulfills

$$p_0^{a_{<0}}(y) = \delta_{x_0}(y), \quad p_{h+1}^{a_{<h+1}}(y) = \int_{\mathsf{S}} p_h^{a_{<h}}(z) p_{h+1}^{a_h}(y|z)\,dz, \quad h\in[H[.$$

3 Algorithm

Fix some "representative" controls $b_0,\ldots,b_{H-1} \in \mathsf{A}$ and simulate independently for $n = 1,\ldots,N$, the chains $\left(S_h^{(n)} = S_h^{(n)}(b_{<h})\right)_{h\in[H]}$ according to (2.5) all starting from a fixed point $x_0 \in \mathsf{S}$. Fix some bounded function f on S and consider the following approximation

$$\mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[f(S_{h+1})\right] \approx \mathcal{E}_{h,N}(x,a;f) := \sum_{n=1}^N f(S_{h+1}^{(n)})\, w_{h,n,N}(x,a) \quad \text{with}$$

$$w_{h,n,N}(x,a) := \frac{p_{h+1}^a(S_{h+1}^{(n)}|x)}{\sum_{k=1,\,k\neq n}^N p_{h+1}^{b_h}(S_{h+1}^{(n)}|S_h^{(k)})} \Bigg/ \sum_{n'=1}^N \frac{p_{h+1}^a(S_{h+1}^{(n')}|x)}{\sum_{k'=1,\,k'\neq n'}^N p_{h+1}^{b_h}(S_{h+1}^{(n')}|S_h^{(k')})} \tag{3.1}$$

for any $(x,a) \in (\mathsf{S}\times\mathsf{A})$ where by definition $0/0 = 0$. Note that $w_{l,n,N}$ satisfy $w_{l,n,N} \ge 0$ and $\sum_{n=1}^N w_{l,n,N} = 1$. We thus propose the following (pseudo) weighted stochastic mesh algorithm.

Algorithm 3.1

- *Initialization:* $\overline{V}_H(S_H^{(n)}) = F(S_H^{(n)}),\, n = 1,\ldots,N.$
- *Backward step: Suppose that for* $h+1 \leq H$, $\overline{V}_{h+1}(S_{h+1}^{(n)})$ *is constructed for* $n = 1,\ldots,N$. *Then we define*

  $$\overline{V}_h(S_h^{(r)}) = \sup_{a\in\mathsf{A}} \Bigl(R_h(S_h^{(r)}, a) + \mathcal{E}_{h,N}(S_h^{(r)}, a; \overline{V}_{h+1})\Bigr) \tag{3.2}$$

  for $r = 1,\ldots,N.$
- *Output:* $\overline{V}_0(x_0)$.

Note that the above algorithm depends on the choice of controls $b_0,\ldots,b_{H-1} \in \mathsf{A}$. However, as we show in the next section, this choice of controls doesn't influence the convergence rates of the algorithm under proper assumptions.

4 Convergence analysis

In this section we study the convergence of the stochastic mesh algorithm presented in Section 3. First we assume that the state/action space is compact and then extend our results to a noncompact case. Throughout this section we make the following assumption.

Assumption 4.1 *Assume that* $\mathsf{S} \subset \mathbb{R}^{d_\mathsf{S}}$ *and* $\mathsf{A} \subset \mathbb{R}^{d_\mathsf{A}}$ *for some natural numbers* d_S *and* d_A. *Moreover,* S *and* A *are compact with (finite) diameters* $\mathrm{diam}(\mathsf{S})$ *and* $\mathrm{diam}(\mathsf{A})$, *respectively.*

Assumption 4.2 *There exist constants* $\delta > 0$, $\Lambda > 0$ *and a function* $\mathcal{L} : \mathsf{S} \to \mathbb{R}_+$ *such that the one-step transition densities* $(p_h^a(y|x),\ h \in [H])$ *satisfy*

$$0 < \delta \leq p_h^a(y|x) \leq \Lambda, \quad |p_h^{a_1}(y|x_1) - p_h^{a_2}(y|x_2)| \leq \mathcal{L}(y)(|x_1 - x_2| + \rho_\mathsf{A}(a_1, a_2))$$

for all $x, x_1, x_2, y \in \mathsf{S}$, $a, a_1, a_2 \in \mathsf{A}$ *and* $h = 1,\ldots,H$, *where* $\max\{\|\mathcal{L}\|_{L^1(\mathsf{S})}, \|\mathcal{L}\|_{L^\infty(\mathsf{S})}\} \leq L$. *Moreover*

$$\max\{|R_h(s,a)|, |F(s)|\} \leq G, \quad (s,a) \in \mathsf{S}\times\mathsf{A}, \quad h \in [H[.$$

Under these assumptions, we can prove the following bound.

Theorem 4.3 *It holds that*

$$\mathbb{E}\left[|\overline{V}_0(x_0) - V_0^\star(x_0)|\right] \lesssim \frac{H^2 G}{\sqrt{N}} \left(\frac{L\,\mathrm{DI}(\mathsf{S}\times\mathsf{A}) + L\,\mathrm{diam}(\mathsf{S})\mathrm{diam}(\mathsf{A}) + \Lambda}{\delta} + \frac{\Lambda^2}{\delta^2}\right)$$

for all $N > N_0$ *with* N_0 *large enough and* $\lesssim$ *denoting* $\leq$ *up to some (absolute) proportionality constant.*

Proof. For $r = 1, \ldots, N$, one has

$$\begin{aligned}
&\left|\overline{V}_h(S_h^{(r)}) - V_h^\star(S_h^{(r)})\right| \\
&\leq \sup_{a\in\mathsf{A}} \left|\sum_{n=1}^N \overline{V}_{h+1}(S_{h+1}^{(n)}) w_{h,n,N}(S_h^{(r)}, a) - \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|S_h^{(r)},a)}\left[V_{h+1}^\star(S_{h+1})\right]\right| \\
&\leq \sup_{a\in\mathsf{A}} \sum_{n=1}^N \left|\overline{V}_{h+1}(S_{h+1}^{(n)}) - V_{h+1}^\star(S_{h+1}^{(n)})\right| w_{h,n,N}(S_h^{(r)}, a) \\
&+ \sup_{a\in\mathsf{A}} \left|\sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) w_{h,n,N}(S_h^{(r)}, a) - \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|S_h^{(r)},a)}\left[V_{h+1}^\star(S_{h+1})\right]\right| \\
&\leq \|\overline{V}_{l+1} - V_{l+1}^\star\|_N + \mathcal{R}_{l+1},
\end{aligned}$$

where

$$\|\overline{V}_h - V_h^\star\|_N := \max_{1\leq r\leq N}\left|\overline{V}_h(S_h^{(r)}) - V_h^\star(S_h^{(r)})\right|$$

and

$$\mathcal{R}_{h+1} := \sup_{\substack{a\in\mathsf{A}\\1\leq r\leq N}} \left|\sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) w_{h,n,N}(S_h^{(r)}, a) - \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|S_h^{(r)},a)}\left[V_{h+1}^\star(S_{h+1})\right]\right|.$$

Hence, since $\overline{V}_H - V_H^\star = 0$,

$$\|\overline{V}_h - V_h^\star\|_N \leq \|\overline{V}_{h+1} - V_{h+1}^\star\|_N + \mathcal{R}_{h+1} \leq \sum_{k=h}^{H-1}\mathcal{R}_{k+1}. \tag{4.1}$$

We now proceed with the estimation of $\mathbb{E}\left[\mathcal{R}_{h+1}\right]$, $h = 0, \ldots, H-1$, and write

$$\sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) w_{h,n,N}(S_h^{(r)}, a) - \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|S_h^{(r)},a)}\left[V_{h+1}^\star(S_{h+1})\right] = T_{1,r}(a) + T_{2,r}(a) + T_{3,r}(a)$$

where

$$\begin{aligned}
T_{1,r}(a) &:= \sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) \frac{p_{h+1}^a(S_{h+1}^{(n)}|S_h^{(r)})}{\sum_{k=1,\,k\neq n}^N p_{h+1}^{b_h}(S_{h+1}^{(n)}|S_h^{(k)})}\left(\left[\sum_{n'=1}^N \frac{p_{h+1}^a(S_{h+1}^{(n')}|S_h^{(r)})}{\sum_{k'=1,\,k'\neq n'}^N p_{h+1}^{b_h}(S_{h+1}^{(n')}|S_h^{(k')})}\right]^{-1} - 1\right), \\
T_{2,r}(a) &:= \sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) \left(\frac{p_{h+1}^a(S_{h+1}^{(n)}|S_h^{(r)})}{\sum_{k=1,\,k\neq n}^N p_{h+1}^{b_h}(S_{h+1}^{(n)}|S_h^{(k)})} - \frac{p_{h+1}^a(S_{h+1}^{(n)}|S_h^{(r)})}{N p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}\right), \\
T_{3,r}(a) &:= \sum_{n=1}^N V_{h+1}^\star(S_{h+1}^{(n)}) \frac{p_{h+1}^a(S_{h+1}^{(n)}|S_h^{(r)})}{N p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})} - \mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|S_h^{(r)},a)}\left[V_{h+1}(S_{h+1})\right].
\end{aligned}$$

We have

$$\begin{aligned}
|T_{1,r}(a)| &\leq HG \sup_{x\in\mathsf{S},\,a\in\mathsf{A}}\left|\sum_{n'=1}^N \frac{p_{h+1}^a(S_{h+1}^{(n')}|x)}{\sum_{k'=1,\,k'\neq n'}^N p_{h+1}^{b_h}(S_{h+1}^{(n')}|S_h^{(k')})} - 1\right| \\
&\leq HG \sup_{x\in\mathsf{S},\,a\in\mathsf{A}}\left|1 - \frac{1}{N}\sum_{n=1}^N \frac{p_{h+1}^a(S_{h+1}^{(n)}|x)}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}\right| \\
&+ \frac{HG\Lambda}{N}\sum_{n=1}^N\left|\frac{1}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})} - \frac{1}{\frac{1}{N}\sum_{k=1,\,k\neq n}^N p_{h+1}^{b_h}(S_{h+1}^{(n)}|S_l^{(k)})}\right|.
\end{aligned}$$

Note that

$$\mathbb{E}\left[\frac{p_{h+1}^{a}(S_{h+1}^{(n)}|x)}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}\right]=\int_{\mathsf{S}}p_{h+1}^{a}(z|x)\,dz=1,$$

and

$$p_{h+1}^{a_{<h+1}}(y)\;=\int_{\mathsf{S}}p_{h}^{a_{<h}}(z)p_{h+1}^{a_h}(y|z)\,dz\geq\delta\int_{\mathsf{S}}p_{h}^{a_{<h}}(z)\,dz=\delta$$

for all $y\in\mathsf{S}$, $h\in[H[$, $a_{<h+1}\in\mathsf{A}^{l+1}$. It then follows by Proposition B.1, that

$$\mathbb{E}\sup_{x\in\mathsf{S},a\in\mathsf{A}}\left|1-\frac{1}{N}\sum_{n=1}^{N}\frac{p_{h+1}^{a}(S_{h+1}^{(n)}|x)}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}\right|\lesssim\frac{L\,\mathrm{DI}(\mathsf{S}\times\mathsf{A})+L\,\mathrm{diam}(\mathsf{S}\times\mathsf{A})+\Lambda}{\delta\sqrt{N}}.$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}_{S_{h+1}^{(1)}}\left[\left|\frac{1}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})}-\frac{1}{\frac{1}{N}\sum_{k=1,\,k\neq1}^{N}p_{h+1}^{b_h}(S_{h+1}^{(1)}|S_{h}^{(k)})}\right|\right]\\
&\leq\frac{N}{N-1}\delta^{-2}\mathbb{E}_{S_{h+1}^{(1)}}\left[\left|\frac{1}{N}\sum_{k=1,\,k\neq1}^{N}p_{h+1}^{b_h}(S_{h+1}^{(1)}|S_{h}^{(k)})-p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})\right|\right]\\
&\leq\delta^{-2}\mathbb{E}_{S_{h+1}^{(1)}}\left[\left|\frac{1}{N-1}\sum_{k=1,\,k\neq1}^{N}\left(p_{h+1}^{b_h}(S_{h+1}^{(1)}|S_{h}^{(k)})-p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})\right)\right|\right]\\
&+\frac{\delta^{-2}}{N-1}p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})\\
&\leq\frac{\delta^{-2}}{\sqrt{N-1}}\sqrt{\int_{\mathsf{S}}p_{h+1}^{b_h}(S_{h+1}^{(1)}|z)^2p_{h}^{b_{<h}}(z)\,dz}+\frac{\delta^{-2}}{N-1}p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)}),
\end{aligned}$$

and so for each $n=1,\ldots,N$ we have by symmetry and Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}-\frac{1}{\frac{1}{N}\sum_{k=1,\,k\neq n}^{N}p_{h+1}^{b_h}(S_{h+1}^{(n)}|S_{h}^{(k)})}\right|\right]\\
&\qquad=\mathbb{E}\left[\left|\frac{1}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})}-\frac{1}{\frac{1}{N}\sum_{k=1,\,k\neq1}^{N}p_{h+1}^{b_h}(S_{h+1}^{(1)}|S_{h}^{(k)})}\right|\right]\\
&\qquad\qquad\lesssim\frac{\delta^{-2}}{\sqrt{N}}\sqrt{\int_{\mathsf{S}\times\mathsf{S}}p_{h+1}^{b_h}(z'|z)^2p_{h}^{b_{<h}}(z)p_{h+1}^{b_{<h+1}}(z')\,dz\,dz'}\lesssim\frac{\Lambda}{\delta^2\sqrt{N}}
\end{aligned}$$

for $N>N_0$. Analogously, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{r\in[N],\,a\in\mathsf{A}}|T_{2,r}(a)|\right]&\leq HG\Lambda\mathbb{E}\left[\left|\frac{1}{\frac{1}{N}\sum_{k=1,\,k\neq1}^{N}p_{h+1}^{b_h}(S_{h+1}^{(1)}|S_{h}^{(k)})}-\frac{1}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(1)})}\right|\right]\\
&\leq\frac{HG\Lambda^2}{\delta^2\sqrt{N}}
\end{aligned}$$

for $N>N_0$. We next consider $T_{3,r}$. For each fixed $x\in\mathsf{S}$ and $a\in\mathsf{A}$, we have

$$\mathbb{E}\left[V_{h+1}^{\star}(S_{h+1}^{(n)})\frac{p_{h+1}^{a}(S_{h+1}^{(n)}|x)}{p_{h+1}^{b_{<h+1}}(S_{h+1}^{(n)})}\right]=\int_{\mathsf{S}}V_{h+1}^{\star}(z)p_{h+1}^{a}(z|x)\,dz=\mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V_{h+1}^{\star}(x)\right].$$

Then, by Proposition B.1 again, it follows that

$$\mathbb{E}\left[\max_{r\in[N],\,a\in\mathsf{A}}|T_{3,r}(a)|\right]\leq$$
$$\mathbb{E}\left[\sup_{x\in\mathsf{S},\,a\in\mathsf{A}}\left|\frac{1}{N}\sum_{n=1}^{N}V^{\star}_{h+1}(S^{(n)}_{h+1})\frac{p^{a}_{h+1}(S^{(n)}_{h+1}|x)}{p^{b<h+1}_{h+1}(S^{(n)}_{h+1})}-\mathbb{E}_{S_{h+1}\sim P_{h+1}(\cdot|x,a)}\left[V^{\star}_{h+1}(S_{h+1})\right]\right|\right]$$
$$\lesssim\frac{HG}{\delta}\frac{L\,\mathrm{DI}(\mathsf{S}\times\mathsf{A})+L\,\mathrm{diam}(\mathsf{S}\times\mathsf{A})+\Lambda}{\sqrt{N}}.$$

Finally, we apply (4.1) for $h=0$. ■

4.1 Non-compact case

If S is not compact subset of $\mathbb{R}^{d}$ we consider its approximation by compact susbsets. Let $\mathcal{D}$ be compact subset of S and let $(S^{(h,x),\mathcal{D}}_{k}(\boldsymbol{\pi}),\ k=h,\ldots,H)$ be a process obtained by reflection of the chain $(S^{(h,x)}_{k}(\boldsymbol{\pi}),\ k=h,\ldots,H)$ reflected in $\mathcal{D}$ as described in Appendix C. For a fixed policy $\boldsymbol{\pi}\in\Pi$ and $x\in\mathcal{D}$, consider the exit (stopping) times,

$$\tau^{x,\mathcal{D}}_{h}:=\min\left\{k\geq h:S^{(h,x)}_{k}\notin\mathcal{D}\right\}.$$

where $(S^{(h,x)}_{k}=S^{(h,x)}_{k}(\boldsymbol{\pi}),\ k=h,\ldots,H)$ stands for the chain (2.5) following the policy $\boldsymbol{\pi}$ and starting in x at time h. Hence

$$S^{(h,x),\mathcal{D}}_{k}1_{\left\{\tau^{x,\mathcal{D}}_{h}>k\right\}}\overset{\mathrm{Law}}{=}S^{(h,x)}_{k}1_{\left\{\tau^{x,\mathcal{D}}_{h}>k\right\}}.$$

We now consider the MDP in the compact domain $\mathcal{D}$,

$$V^{\mathcal{D}}_{h}(x):=\sup_{\boldsymbol{\pi}\in\Pi}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{k=h}^{H-1}R_{k}(S^{(h,x),\mathcal{D}}_{k},A_{k})+F_{H}(S^{(h,x),\mathcal{D}}_{H})\right]\tag{4.2}$$

as an approximation to $V_{h}(x),\ h\in[H-1]$. It is not difficult to see that

$$\left|V^{\mathcal{D}}_{h}(x)-V^{\star}_{h}(x)\right|\lesssim HG\sup_{\boldsymbol{\pi}\in\Pi}\mathbb{P}_{\boldsymbol{\pi}}(\tau^{x,\mathcal{D}}_{h}\leq H)$$
$$\lesssim HG\sum_{l=h}^{H}\sup_{\boldsymbol{\pi}\in\Pi}\mathbb{P}_{\boldsymbol{\pi}}\big(S^{(x,h)}_{l}\notin\mathcal{D}\big).$$

Furthermore, the one-step transition density $p^{a,\mathcal{D}}_{h}$ of the process $(S^{\mathcal{D}}_{h})$ is given by (see Appendix C)

$$p^{a,\mathcal{D}}_{h}(y|x)=p^{a}_{h}(y|x)+\frac{1}{\lambda(\mathcal{D})}\int_{\mathbb{R}^{d}\setminus\mathcal{D}}p^{a}_{h}(z|x)dz,\quad x,y\in\mathcal{D}.$$

Instead of Assumption 4.2 we now consider the following weaker assumption on $\mathcal{D}$.

Assumption 4.4 *For any compact subset $\mathcal{D}$ of S there exist some constants $\delta_{\mathcal{D}}>0,\Lambda>0,L_{\mathcal{D}}>0$, and a function $\mathcal{L}_{\mathcal{D}}:\mathsf{S}\to\mathbb{R}_{+}$ such that the one-step transition density p^{a}_{h} satisfies*

$$0<\delta_{\mathcal{D}}\leq p^{a}_{h}(z|x)\leq\Lambda,\quad|p^{a_1}_{h}(y|x_1)-p^{a_2}_{h}(y|x_2)|\leq\mathcal{L}_{\mathcal{D}}(y)(|x_1-x_2|+\rho_{\mathsf{A}}(a_1,a_2))$$

for all $x, z, x_1, x_2 \in \mathcal{D}$, $y \in \mathbb{R}^d$, $a, a_1, a_2 \in \mathsf{A}$ *and* $h = 1, \ldots, H$, *where*

$$\|\mathcal{L}_{\mathcal{D}}\|_{L^\infty(\mathcal{D})} + \frac{1}{\lambda\left(\mathcal{D}\right)}\|\mathcal{L}_{\mathcal{D}}\|_{L^1(\mathbb{R}^d\setminus\mathcal{D})} \le L_{\mathcal{D}}.$$

Moreover

$$\max\{|R_h(s,a)|, |F(s)|\} \le G, \quad (s,a) \in \mathsf{S} \times \mathsf{A}, \quad h \in [H[.$$

Hence $p_h^{a,\mathcal{D}}(y|x) \ge \delta_{\mathcal{D}}$ for $x, y \in \mathcal{D}$ and $a \in \mathsf{A}$, and furthermore,

$$\begin{aligned} |p_h^{a_1,\mathcal{D}}(y|x_1) - p_h^{a_2,\mathcal{D}}(y|x_2)| &\le |p_h^{a_1}(y|x_1) - p_h^{a_2}(y|x_2)| + \frac{1}{\lambda(\mathcal{D})}\int_{\mathbb{R}^d\setminus\mathcal{D}} |p_h^{a_1}(z|x_1) - p_h^{a_2}(z|x_2)|\, dz \\ &\le L_{\mathcal{D}}(|x_1 - x_2| + \rho_{\mathsf{A}}(a_1, a_2)). \end{aligned}$$

Theorem 4.5 *Fix some* x_0 *then under Assumption 4.4, it holds that*

$$\begin{aligned} \mathbb{E}\left[\left|\overline{V}_0^{\mathcal{D}}(x_0) - V_0^\star(x_0)\right|\right] \lesssim \frac{H^2 G}{\sqrt{N}} &\left(\frac{L_{\mathcal{D}}\,\mathrm{DI}(\mathcal{D}\times\mathsf{A}) + L_{\mathcal{D}}\,\mathrm{diam}(\mathcal{D})\mathrm{diam}(\mathsf{A}) + \Lambda_{\mathcal{D}}}{\delta_{\mathcal{D}}} + \frac{\Lambda^2}{\delta_{\mathcal{D}}^2}\right) \\ &+ HG\sum_{l=0}^{H}\sup_{\pi\in\Pi}\mathbb{P}_\pi\left(S_l^{(x_0,0)} \notin \mathcal{D}\right) \end{aligned}$$

for all $N > N_0$ *with* N_0 *large enough and* $\lesssim$ *denoting* $\le$ *up to some absolute proportionality constant.*

5 Complexity

In this section, we estimate the computational budget, that is, the complexity, needed for computing $V^\star(x_0)$ (in L^1) with a given accuracy $\epsilon > 0$, by the algorithm presented in Section 3. For simplicity we disregard the cost of the optimization step and identify the overall cost with HN^2, that is, the costs of computing all weights (3.1) for N trajectories.

5.1 Complexity for compact S and A

In order to explicitly incorporate the dimension of the state and action space in the complexity estimation, we consider a sequence of MDPs for running $d = 1, 2, \ldots$ Without much loss of generality we assume that $\mathsf{S}_d = B_{R_d}^d \subset \mathbb{R}^d$, $\mathsf{A}_d = B_{A_d}^d \subset \mathbb{R}^d$ for some $R_d > 0$ and $A_d > 0$ with B_R^d being the Euclidean ball in $\mathbb{R}^d$ of radius R. We further assume that in dimension d, the transition probabilities are given by $p_{d,h}^a(y|x)$. Furthermore it is assumed that the bound G in Assumption 4.2 holds uniformly in d. Obviously, if R_d, A_d, and $p_{d,h}^a(y|x)$ are such that L_d, Λ_d, and δ_d^{-1} due to Assumption 4.2 can taken to be polynomially bounded in d, then Theorem 4.3 implies that

$$\mathcal{C}\left(\epsilon, d\right) \lesssim \frac{H^9 G^4}{\epsilon^4}\text{polynomial}\left(d\right)$$

In this case, the mesh algorithm is tractable in the sense of [9], that is,

$$\lim_{d+\epsilon^{-1}\to\infty}\frac{\log\mathcal{C}\left(\epsilon, d\right)}{d + \epsilon^{-1}} = 0.$$

This result can be seen as an extension of [11] to the case of finite horizon MDPs with more general state and action spaces.

5.2 Complexity for noncompact S and compact A. In particular, A can be infinite.

Let us now consider the noncompact case with $\mathsf{S} = \mathbb{R}^d$, $A \subset \mathbb{R}^d$ in the setup of Section 5.1. We then have the following result.

Proposition 5.1 *Suppose that, for a generic d, there is a sequence of compact sets $\mathcal{D}_{d,N}$, $N \in \mathbb{N}$, such that*

$$\frac{L_{\mathcal{D}_{d,N}}\,\mathrm{DI}(\mathcal{D}_{d,N} \times \mathsf{A}_d) + L_{\mathcal{D}_{d,N}}\,\mathrm{diam}(\mathcal{D}_{d,N})\mathrm{diam}(\mathsf{A}_d)}{\delta_{\mathcal{D}_{d,N}}} + \frac{\Lambda_d^2}{\delta_{\mathcal{D}_{d,N}}^2} \leq C_1(H, \log N, d)N^{\alpha}$$

and

$$\sum_{h=1}^{H} \sup_{\boldsymbol{\pi}\in\Pi} \mathbb{P}_{\boldsymbol{\pi}}\big(S_h^{(x_0,0)} \notin \mathcal{D}_{d,N}\big) \leq C_2(H, \log N, d)N^{-\beta}$$

for $N > N_0$, where C_1 and C_2 are functions on $\mathbb{N} \times \mathbb{R} \times \mathbb{N}$ such that

$$0 \leq C_{1,2}(x, y, d) \leq c_d\,|xy|^{q_d} \quad \textit{for all } x, y \geq 1,$$

and the parameters $\alpha \in [0, 1/2)$, $\beta > 0$ do not depend on N and d. Here both $c_d > 0$ and $q_d \in \mathbb{R}_+$ are independent of H and ϵ. Then the complexity $\mathcal{C}(\epsilon, d)$ of our algorithm can be bounded as

$$\mathcal{C}(\epsilon, d) \lesssim H \max\left(2GcH^2\left(\frac{2H}{1-2\alpha}\right)^{q_d}, 2GcH\left(\frac{H}{\beta}\right)^{q_d}, 1\right)^{2\max(1/\beta,2/(1-2\alpha))} \times \frac{\log^{2q_d\max(1/\beta,2/(1-2\alpha))}(1/\epsilon)}{\epsilon^{2\max(1/\beta,2/(1-2\alpha))}}. \tag{5.1}$$

Corollary 5.2 *If one has in addition that $q_d \leq \eta d$ and $c_d \leq c_0 \exp(\lambda d)$ for some universal constants $c_0, \eta, \lambda > 0$, one obtains*

$$\mathcal{C}(\epsilon, d) \lesssim H \max\left(2Gc_0e^{\lambda d}H^{2+\eta d}\frac{2^{\eta d}}{(1-2\alpha)^{\eta d}}, 2Gc_0e^{\lambda d}H^{1+\eta d}\frac{1}{(\beta\wedge 1)^{\eta d}}, 1\right)^{2\max(1/\beta,2/(1-2\alpha))} \times \frac{\log^{2\eta d\max(1/\beta,2/(1-2\alpha))}(1/\epsilon)}{\epsilon^{2\max(1/\beta,2/(1-2\alpha))}},$$

which implies

$$\log \mathcal{C}(\epsilon, d) = r_1 \log H + \left(r_2 + r_3 \log H + r_4 \log\log\frac{1}{\epsilon}\right) d + r_5 \log\frac{1}{\epsilon}$$

for certain constants $r_1, \ldots, r_5 > 0$. From this it is easy to see that the problem is not tractable in the sense of [9], but, since

$$\lim_{d\to\infty}\lim_{\epsilon\searrow 0}\frac{\log \mathcal{C}(\epsilon, d)}{f(d)\log(1/\epsilon)} = 0 \quad \textit{for any } f \textit{ with } f(d) \to \infty \textit{ as } d \to \infty,$$

the problem is semi-tractable in the sense of [3] and we have a kind of "weak curse of dimensionality".

The next section provides an example where Corollary 5.2 applies.

5.3 Example: Gaussian transition densities

Let us consider the case of Gaussian transition probabilities of the form

$$p_h^a(y|x) \equiv p_{d,h}^a(y|x) = \frac{1}{(2\pi\sigma_h^2)^{d/2}} \exp(-|x-y-a|^2/(2\sigma_h^2)), \quad x, y \in \mathbb{R}^d \tag{5.2}$$

where the (scalar) variances $\sigma_h,\ h \in]H]$ are all bounded from above and below, that is, $0 < \sigma_{\min} \leq \sigma_h \leq \sigma_{\max} < \infty$. Let $\mathsf{S} = \mathbb{R}^d$, $\mathsf{A}_d = B_A^d \subset \mathbb{R}^d$ for some $A > 0$ and $\mathcal{D}_{d,N} = B_{R_N}^d$ with B_R^d being the Euclidean ball in $\mathbb{R}^d$ of radius R. Such densities naturally appear as transition densities of discretized (e.g. via Euler scheme) diffusion processes, see 6 for numerical illustrations. Let us check now the assumptions of Proposition 5.1 and Corollary 5.2. In what follows, we do not always denote dependence on d explicitly, for notational convenience. Choosing

$$R_N = \sqrt{\gamma\sigma_{\min}^2 \log(N)/4} \ \text{ for some } \ \gamma \in (0, 1/4), \tag{5.3}$$

we see that

$$p_h^a(y|x) \geq \frac{1}{(2\pi\sigma_{\max}^2)^{d/2}} \exp(-A^2/\sigma_{\min}^2) N^{-\gamma} =: \delta_{\mathcal{D}_N} \tag{5.4}$$

for all $a \in \mathsf{A}$, $x, y \in \mathcal{D}_N$. Furthermore, we have

$$p_h^a(y|x) \leq (2\pi\sigma_{\min}^2)^{-d/2} =: \Lambda \ \text{ for all } \ h \in [H]. \tag{5.5}$$

Note that for all $x_1, x_2 \in \mathcal{D}_N$, $x, y \in \mathbb{R}^d$, $a, a_1, a_2 \in \mathsf{A}$, $h \in]H]$,

$$\begin{aligned} |\nabla_x p_h^a(y|x)| &= |\nabla_a p_h^a(y|x)| \\ &= \frac{1}{\sigma_h^2 (2\pi\sigma_h^2)^{d/2}} |x-a-y| \exp(-|x-y-a|^2/(2\sigma_h^2)). \end{aligned}$$

Hence

$$\begin{aligned} &|p_h^{a_1}(y|x_1) - p_h^{a_2}(y|x_2)| \\ &\leq \frac{\sqrt{2}}{\sigma_{\min}^{d+2}(2\pi)^{d/2}} \sup_{x \in \mathcal{D}_N, a \in \mathsf{A}} \left\{ |x-a-y| \exp(-|x-y-a|^2/(2\sigma_{\max}^2)) \right\} \\ &\times (|x_1 - x_2| + |a_1 - a_2|) \\ &=: \mathcal{L}_{\mathcal{D}_N}(y) \left(|x_1 - x_2| + |a_1 - a_2| \right). \end{aligned}$$

So on the one hand we have

$$\begin{aligned} \|\mathcal{L}_{\mathcal{D}_N}\|_{L^\infty(\mathcal{D}_N)} &\leq \frac{\sqrt{2}}{\sigma_{\min}^{d+2}(2\pi)^{d/2}} (2R_N + A) \\ &\simeq \frac{R_N}{\sigma_{\min}^{d+2} 2^{(d-3)/2} \pi^{d/2}}, \quad N \to \infty. \end{aligned}$$

On the other hand, for $R_N > A$ and $|y| \geq 2R_N$ it holds that

$$0 \leq \mathcal{L}_{\mathcal{D}_N}(y) \leq \frac{\sqrt{2}}{\sigma_{\min}^{d+2}(2\pi)^{d/2}} 3|y| \exp(-\left(|y| - 2R_N\right)^2 / (2\sigma_{\max}^2)),$$

from which we see that $\|\mathcal{L}_{\mathcal{D}_N}\|_{L^1\left(\mathbb{R}^d\right)} < \infty$ for any N, and moreover

$$\begin{aligned}\|\mathcal{L}_{\mathcal{D}_N}\|_{L^1\left(\mathbb{R}^d\right)} &\leq \|\mathcal{L}_{\mathcal{D}_N}\|_{L^\infty\left(B_{R_{2N}}\right)} \operatorname{Vol}\left(B_{R_{2N}}\right) \\ &\quad + \frac{\sqrt{2}}{\sigma_{\min}^{d+2}(2\pi)^{d/2}} \int_{|y|\geq 2R_N} 3\,|y| \exp(-\left(|y|-2R_N\right)^2/(2\sigma_{\max}^2))dy \\ &= \frac{2^{(5+d)/2}}{\sigma_{\min}^{d+2}\Gamma(d/2+1)} R_N^{d+1} \\ &\quad + \frac{3\sigma_{\max}}{\sigma_{\min}^{d+2}2^{d/2-1}\Gamma(d/2)} \int_0^\infty \left(\sigma_{\max}\sqrt{2t}+2R_N\right)^d t^{-1/2}\exp(-t)dt \\ &\equiv \mathsf{Term1}_N + \mathsf{Term2}_N,\end{aligned}$$

where some standard estimates show that $\mathsf{Term2}_N \lesssim_d R_N^d$, and so is asymptotically dominated by $\mathsf{Term1}_N$. Then similar calculations show that

$$\|\mathcal{L}_{\mathcal{D}_N}\|_{L^\infty(\mathcal{D}_N)} + \frac{\|\mathcal{L}_{\mathcal{D}_N}\|_{L^1\left(\mathbb{R}^d\setminus\mathcal{D}_N\right)}}{\lambda\left(\mathcal{D}_N\right)} \leq 2\frac{1+2^{(d+3)/2}}{\sigma_{\min}^{d+2}\pi^{d/2}} R_N =: L_{\mathcal{D}_N}. \tag{5.6}$$

By taking into account that $d_{\mathsf{S}} = d_{\mathsf{A}} = d$, $\mathrm{DI}(\mathcal{D}_{d,N}\times\mathsf{A}_d) \lesssim (A+R_N)\sqrt{d}$, we then have by (5.3), (5.4), (5.5), (5.6) that

$$\begin{aligned}\left(L_{\mathcal{D}_N}\mathrm{DI}(\mathcal{D}_{d,N}\times\mathsf{A}_d) + L_{\mathcal{D}_N}\operatorname{diam}(\mathcal{D}_N)\operatorname{diam}(\mathsf{A}) + \Lambda\right)/\delta_{\mathcal{D}_N} &\lesssim L_{\mathcal{D}_N}\operatorname{diam}(\mathcal{D}_N)\operatorname{diam}(\mathsf{A})/\delta_{\mathcal{D}_N} \\ &\lesssim 8A\frac{1+2^{(d+3)/2}}{\sigma_{\min}^{d+2}\pi^{d/2}}\frac{R_N^2}{\delta_{\mathcal{D}_N}} \\ &\simeq 2\gamma A\left(1+2^{(2d+3)/2}\right) \\ &\quad\times\left(\sigma_{\max}/\sigma_{\min}\right)^d \exp(A^2/\sigma_{\min}^2)\log N\cdot N^\gamma\end{aligned} \tag{5.7}$$

for $N\to\infty$. Further we have

$$\frac{\Lambda}{\delta_{\mathcal{D}_N}^2} \leq \left(\sigma_{\max}/\sigma_{\min}\right)^{2d}\exp(2A^2/\sigma_{\min}^2)N^{2\gamma}, \tag{5.8}$$

which dominates (5.7). That is,

$$C_1(H,\log N,d) = \left(\sigma_{\max}/\sigma_{\min}\right)^{2d}\exp(2A^2/\sigma_{\min}^2)$$

and thus $\alpha := 2\gamma < 1/2$ as required. Next we bound

$$\mathbb{P}_{\boldsymbol{\pi}}\left(S_h^{(x_0,0)}\notin\mathcal{D}_N\right) \leq \sup_{a_{<h}\in\mathsf{A}^h}\int_{\mathbb{R}^d\setminus B_{R_N}} p_h^{a_{<h}}(y)\,dy.$$

Note that

$$p_h^{a_{<h}}(y) = \frac{1}{\left(2\pi\bar{\sigma}_h^2\right)^{d/2}}\exp\left(-|x_0-y-\bar{a}_{<h}|^2/\left(2\bar{\sigma}_h^2\right)\right)$$

where $\bar{\sigma}_h^2 = \sum_{l=1}^h\sigma_l^2$ and $\bar{a}_{<h} = \sum_{l=0}^{h-1}a_l$. Suppose that N is large enough such that $|x_0|\leq R_N/4$ and $HA\leq R_N/4$ then

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}\left(S_h^{(x_0,0)}\notin\mathcal{D}_N\right) &\leq \sup_{a_{<h}\in\mathsf{A}^h}\frac{1}{\left(2\pi\bar{\sigma}_h^2\right)^{d/2}}\int_{\mathbb{R}^d\setminus B_{R_N}}\exp\left(-|x_0-y-\bar{a}_{<h}|^2/\left(2\bar{\sigma}_h^2\right)\right)\,dy \\ &\leq \frac{1}{\left(2\pi\right)^{d/2}}\int_{|z|>R_N/(2\bar{\sigma}_h)}\exp\left(-|z|^2/2\right)\,dz \\ &= \frac{\Gamma\left(d/2,R_N^2/(8\bar{\sigma}_h^2)\right)}{\Gamma\left(d/2\right)}\end{aligned}$$

where $\Gamma(s,x)$ denotes the incomplete Gamma function, which has asymptotics $\Gamma(s,x) \simeq x^{s-1}e^{-x}$ for $x \to \infty$. By plugging in the choice for R_N we get for $N \to \infty$,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}\Big(S_h^{(x_0,0)} \notin \mathcal{D}_N\Big) &\simeq \frac{8}{2^{3d/2}\Gamma(d/2)}(R_N/\bar{\sigma}_h)^{d-2}\exp\left(-R_N^2/(8\bar{\sigma}_h^2)\right) \\ &= \frac{32}{2^{5d/2}\Gamma(d/2)}(\frac{\sigma_{\min}}{\bar{\sigma}_h})^{(d-2)}\gamma^{\frac{d}{2}-1}N^{-\gamma\sigma_{\min}^2/\left(32\bar{\sigma}_h^2\right)}\log^{\frac{d}{2}-1}N \\ &\le \frac{32}{2^{5d/2}\Gamma(d/2)}\left(\sqrt{H}\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{(2-d)_+}\gamma^{\frac{d}{2}-1}N^{-\gamma\sigma_{\min}^2/\left(32\bar{\sigma}_H^2\right)}\log^{\frac{d}{2}-1}N\end{aligned}$$

with $(2-d)_+ := \max(2-d,0)$, uniform in $h \in]H]$. We so may take $\beta = \gamma\sigma_{\min}^2/(32\bar{\sigma}_H^2)$ and

$$C_2(H,\log N,d) = \frac{32H}{2^{5d/2}\Gamma(d/2)}\left(\sqrt{H}\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{(2-d)_+}\gamma^{\frac{d}{2}-1}\log^{\frac{d}{2}-1}N.$$

Thus, the conditions of Corollary 5.2 are satisfied with $\alpha = 2\gamma$, $\beta = \gamma\sigma_{\min}^2/(32\bar{\sigma}_H^2)$, $\eta = 1/2$, and $\lambda = 2\log(\sigma_{\max}/\sigma_{\min})$, where $\gamma \in (0,1/4)$ can be further chosen to ensure that $\beta \le 1/2 - \alpha$, leading to a complexity bound $\epsilon^{-2/\beta}\log^{d/\beta}(1/\epsilon) \times$ polynomial(H,d).

6 Linear-quadratic Gaussian (LQG) control problems

Let us consider a classical stochastic linear-quadratic-Gaussian (LQG) control problem for controlled d-dimensional diffusion process of the form

$$dX_t = 2\sqrt{\lambda}\,m_t\,dt + \sqrt{2}\,dW_t \tag{6.1}$$

with $t \in [0,T]$, $X_0 = x_0 \in \mathbb{R}^d$, and with the objective functional

$$J_0^m(x_0) = \mathbb{E}_{m,x_0}\big[-\int_0^T \|m_t\|^2\,dt + F(X_T)\big].$$

Here $(m_t)_{t\in[0,T]}$ with $m_t \in \mathbb{R}^d$ is the adapted control process and F is the terminal reward if $F \ge 0$ or terminal costs if $F < 0$. Further, λ is a positive constant representing the “strength” of the control, and $(W_t)_{t\in[0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. Our goal is to maximize the functional $J_0^m(x_0)$ over a class of control processes $(m_t)_{0\le t\le T}$. The HJB equation for the problem at a generic time $t \in [0,T]$ and $x \in \mathbb{R}^d$, that is

$$J_t^\star(x) := \sup_m \mathbb{E}_{m,x}\big[-\int_t^T \|m_s\|^2\,ds + F(X_T)\big] = \sup_m J_t^m(x),$$

is given by

$$\begin{aligned}\frac{\partial}{\partial t}J_t^\star(x) + \Delta J_t^\star(x) + \lambda\|\nabla J_t^\star(x)\|^2 = 0,\\ J_T^\star(x) = F(x)\end{aligned} \tag{6.2}$$

(see e.g., Yong & Zhou [14, Chapter 4]) where $J_t^\star(x)$ of (6.2) at $t = 0$ is the “optimal negative cost” when the state starts from x. Using the Cole-Hopf transformation $x \to \exp(\lambda J_t^\star(x))$ one transforms the nonlinear PDE (6.2) to the backward heat equation. As a result, the solution of (6.2) admits the explicit formula

$$J_t^\star(x) = \frac{1}{\lambda}\log\Big(\mathbb{E}\Big[\exp\Big(\lambda F(x+\sqrt{2}W_{T-t})\Big)\Big]\Big). \tag{6.3}$$

Table 1: Results for $F(x) = -\log((1+\|x\|^2)/2)$ and $d = 1$. The explicit formula (6.3) gives 0.4542.

mean	bias	standard deviation	number of trajectories
0.464	0.034	0.044	10
0.459	0.007	0.009	100
0.451	0.007	0.009	200
0.451	0.004	0.004	500

This can be used to test the accuracy of the proposed algorithm.

In our implementation, we first discretize the equation (6.1) using the Euler scheme with time step Δ,

$$S_{h+1} = S_h + 2\sqrt{\lambda} m_{h\Delta}\Delta + \sqrt{\Delta}\,\varepsilon_{h+1}, \quad h \in [H[$$

with $H = [T/\Delta]$, $\varepsilon_{h+1} \sim \mathcal{N}(0, I_d)$ and $S_0 = x_0$. We then consider the discrete time controlled Markov chain

$$S_{h+1} = S_h + a_h + \sqrt{\Delta}\,\varepsilon_{h+1}, \quad h \in [H[, \tag{6.4}$$

by taking as control at time h,

$$a_h := 2\sqrt{\lambda} m_{h\Delta}\Delta \in [-A, A]^d \text{ for some } A > 0.$$

As such the conditional density of the Markov chain (6.4) is Gaussian and of the form (5.2) with $\sigma_h^2 = \Delta$ for every h. Thus the objective is to maximize the functional

$$V_0^{\boldsymbol{\pi}}(x_0) = \mathbb{E}_{\boldsymbol{\pi},x_0}\left[-\frac{1}{4\lambda\Delta}\sum_{k=0}^{H-1}\|\pi_k(S_k)\|^2 + F(S_H)\right]$$

over all policies $\boldsymbol{\pi} = (\pi_k(S_k))_{k\in[H[}$, where $\pi_k : \mathbb{R}^d \to [-A, A]^d$. The optimal value of the objective as seen from a generic time h with starting point $S_h = x \in \mathbb{R}^d$ is given by

$$V_h^\star(x) = \sup_{\pi_h,\ldots\pi_{H-1}} \mathbb{E}_{\boldsymbol{\pi},x}\left[-\frac{1}{4\lambda\Delta}\sum_{k=h}^{H-1}\|\pi_k(S_k)\|^2 + F(S_H)\middle| S_h = x\right],$$

and satisfies the backward dynamic program

$$V_h^\star(x) = \max_{a\in[-A,A]^d}\Big(-\frac{\|a\|^2}{4\lambda\Delta} + \mathbb{E}\left[V_{h+1}^\star(x + a + \sqrt{\Delta}\,\varepsilon_{h+1})\right]\Big), \quad h = H-1, \ldots, 0,$$

with $V_H^\star(x) = F(x)$.

In our numerical experiments, we take

$$F(x) = \pm\log((1+\|x\|^2)/2),$$

$T = 0.2$ and $\Delta = 0.01$, hence $H = 20$. Actions are sampled uniformly on $[-1, 1]^d$ and the optimization is performed over the resulting grid. The representative controls $b_0, \ldots, b_{H-1}$ are all taken to be zero. The results for dimension $d = 1$ are presented in Table 1 and Table 2. They are obtained using a grid of 50 actions. The value of the explicit formula (6.3) is approximated using MC with 10000 samples. The results for dimension $d = 5$ are presented in Table 3 and Table 4, and obtained using a grid of 400 actions. Let us further study the performance of our algorithm for different values of the parameter λ. Figure 1 shows the estimates of $V_0^\star(0)$ (red line) together with the values obtained from (6.3) using MC with 10000 paths (green line) for the case $F(x) = \log((1+\|x\|^2)/2)$ and $d = 1$.

Table 2: Results for $F(x) = \log((1 + \|x\|^2)/2)$ and $d = 1$. The explicit formula (6.3) gives -0.357.

mean	bias	standard deviation	number of trajectories
-0.416	0.096	0.077	10
-0.377	0.021	0.016	100
-0.373	0.014	0.011	200
-0.369	0.008	0.006	500

Table 3: Results for $F(x) = -\log((1 + \|x\|^2)/2)$ and $d = 5$. The explicit formula (6.3) gives -0.2474.

mean	bias	standard deviation	number of trajectories
-0.19	0.12	0.15	10
-0.22	0.026	0.034	100
-0.21	0.011	0.015	200
-0.23	0.011	0.013	500

Table 4: Results for $F(x) = \log((1 + \|x\|^2)/2)$ and $d = 5$. The explicit formula (6.3) gives 0.4054.

mean	bias	standard deviation	number of trajectories
0.378	0.13	0.17	10
0.343	0.029	0.038	100
0.337	0.014	0.018	200
0.337	0.013	0.016	500

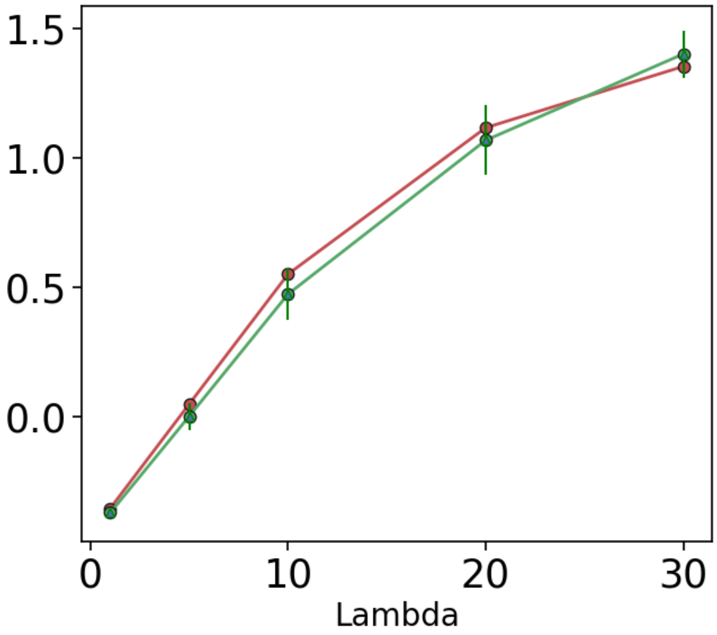

Figure 1: Results for different values of the parameter λ and $F(x) = \log((1 + |x|)^2/2)$.

7 Proofs

7.1 Proof of Proposition 5.1

Let us start with a simple observation. If $x = x(t)$ satisfies the equation

$$\frac{x}{\log^b x} = t, \quad t > 0, \; b > 0, \tag{7.1}$$

one then has for $t \uparrow \infty$,

$$x(t) = (1 + o(1)) t \log^b t. \tag{7.2}$$

This is easily seen as follows. Clearly $x \uparrow \infty$ when $t \uparrow \infty$. Then, by setting $z = \log x$ and $s = \log t$ we get $e^z / z^b = e^s$, and we so may write

$$z \left(1 - b \frac{\log z}{z} \right) = s$$

where $z \uparrow \infty$ when $s \uparrow \infty$. Since $z^{-1} \log z = o(1)$ for $s \uparrow \infty$ we conclude that $z = (1 + o(1))s$, for $s \uparrow \infty$, and hence

$$\log x = (1 + o(1)) \log t, \quad t \uparrow \infty. \tag{7.3}$$

Next, substituting (7.3) in (7.1) yields (7.2).

Now suppose we need to achieve an accuracy $\epsilon > 0$. The assumptions imply that it is sufficient to choose N large enough such that

$$GcH^{2+q} N^{\alpha - \frac{1}{2}} \log^q N \le \frac{\epsilon}{2} \quad \text{and} \quad GcH^{1+q} N^{-\beta} \log^q N \le \frac{\epsilon}{2}, \quad \text{or}$$
$$\frac{N}{\log^{2q/(1-2\alpha)} N} \ge \left(\frac{2GcH^{2+q}}{\epsilon} \right)^{2/(1-2\alpha)} \quad \text{and} \quad \frac{N}{\log^{q/\beta} N} \ge \left(\frac{2GcH^{1+q}}{\epsilon} \right)^{1/\beta},$$

where $c \equiv c_d$ and $q \equiv q_d$. By considering equalities instead of inequalities we get equations of the form (7.1) with asymptotic solutions due to (7.3),

$$N = (1 + o(1)) \left(\frac{2^{q+1} GcH^{2+q}}{(1 - 2\alpha)^q} \right)^{2/(1-2\alpha)} \frac{\log^{2q/(1-2\alpha)} \frac{1}{\epsilon}}{\epsilon^{2/(1-2\alpha)}} \quad \text{and}$$
$$N = (1 + o(1)) \left(\frac{2GcH^{1+q}}{\beta^q} \right)^{1/\beta} \frac{\log^{q/\beta} \frac{1}{\epsilon}}{\epsilon^{1/\beta}}, \quad \text{for} \quad \epsilon \downarrow 0,$$

respectively. We thus end up with a complexity $\mathcal{C}(\epsilon) = HN^2$ which is bounded by (5.1).

A Some auxiliary notions

The Orlicz 2-norm of a real valued random variable η with respect to the function $\psi_2(x) = e^{x^2} - 1$, $x \in \mathbb{R}$, is defined by $\|\eta\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\left[\exp\left(\eta^2/t^2\right)\right] \leq 2\}$. We say that η is *sub-Gaussian* if $\|\eta\|_{\psi_2} < \infty$. In particular, this implies that for some constants $C, c > 0$,

$$\mathbb{P}(|\eta| \geq t) \leq 2\exp\left(-\frac{ct^2}{\|\eta\|_{\psi_2}^2}\right) \quad \text{and} \quad \mathbb{E}[|\eta|^p]^{1/p} \leq C\sqrt{p}\|\eta\|_{\psi_2} \quad \text{for all} \quad p \geq 1.$$

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

$$\|X_t - X_s\|_{\psi_2} \leq K\mathsf{d}(t,s), \quad \forall t, s \in \mathcal{T}.$$

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 (or Dudley integral) of X and

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

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

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

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

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

Proof. Denote

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

with $M_f(x) = \mathbb{E}[f(x,\xi)]$, 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 sub-Gaussian 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 [13, 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. [13, Theorem 8.1.6], for any $\delta \in (0, 1)$,

$$\sup_{x,x' \in \mathsf{X}} |Z(x) - Z(x')| \lesssim L\bigl[\mathrm{DI}(\mathsf{X}) + \mathrm{diam}(\mathsf{X})\sqrt{\log(2/\delta)}\bigr]$$

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. [13, Theorem 2.6.2.], for any $\delta \in (0, 1)$,

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

■

C Reflection of Markov chains

Let $(S_h)_{h=0,\ldots,H}$ be a Markov chain in $\mathbb{R}^d$ with one-step transition density $p_{h+1}(y|x)$ for $0 \leq h < H$. Let further $\mathcal{D} \subset \mathbb{R}^d$ be a compact Borel subset and $q(dy) = \lambda(dy)/\lambda(\mathcal{D})$ with λ being Lebesgue measure on $\mathbb{R}^d$. We then construct a Markov chain $\left(S_h^{\mathcal{D}}\right)_{h=0,\ldots,H}$ in $\mathcal{D}$ as follows: Suppose $S_h^{\mathcal{D}} = x \in \mathcal{D}$. Let $Y \in \mathbb{R}^d$ be a random variable with density $p_{h+1}(y|x)$ and $Q \in \mathcal{D}$ be a random variable independent of Y with density $\lambda^{-1}(\mathcal{D})$, hence Q is uniformly distributed on $\mathcal{D}$. We then define

$$S_{h+1}^{\mathcal{D}} := \left\{ \begin{array}{ll} Y & \text{if} \quad Y \in \mathcal{D} \\ Q & \text{if} \quad Y \notin \mathcal{D} \end{array} \right. .$$

For any non-negative Borel function f in $\mathbb{R}^d$ with support $\mathcal{D}$ one thus has

$$\begin{aligned} \mathbb{E}\left[f(S_{h+1}^{\mathcal{D}})\right] &= \mathbb{E}\left[f(Y)1_{\{Y \in \mathcal{D}\}}\right] + \mathbb{E}\left[f(Q)1_{\{Y \notin \mathcal{D}\}}\right] \\ &= \int_{\mathcal{D}} f(y)p_{h+1}(y|x)dy + \frac{1}{\lambda(\mathcal{D})}\int_{\mathcal{D}} f(y)dy \int_{\mathbb{R}^d \setminus \mathcal{D}} p_{h+1}(z|x)dz. \end{aligned}$$

Hence, $S_{h+1}^{\mathcal{D}}$ is governed by the one-step transition density $p_{h+1}^{\mathcal{D}}$ on $\mathcal{D} \times \mathcal{D}$ given by

$$p_{h+1}^{\mathcal{D}}(y|x) = p_{h+1}(y|x) + \frac{1}{\lambda(\mathcal{D})}\int_{\mathbb{R}^d \setminus \mathcal{D}} p_{h+1}(z|x)dz, \quad x, y \in \mathcal{D}.$$

Consider furthermore the stopping time

$$\tau^{\mathcal{D}} := \min\{h : S_h \notin \mathcal{D}\} \quad \text{with} \quad S_0 = x_0 \in \mathcal{D}.$$

It is not difficult to see that one has that

$$\left(S_h : 0 \leq h < \tau^{\mathcal{D}}\right) \overset{\mathcal{L}}{=} \left(S_h^{\mathcal{D}} : 0 \leq h < \tau^{\mathcal{D}}\right), \tag{C.1}$$

Loosely speaking, $S_h^{\mathcal{D}}$ behaves like S_h before the first exit of the set $\mathcal{D}$. We will say that $S_h^{\mathcal{D}}$ is the *reflected Markov chain* obtained from reflecting the process S_h in $\mathcal{D}$.

References

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

[2] Christian Beck, Arnulf Jentzen, Konrad Kleinberg, and Thomas Kruse. Nonlinear monte carlo methods with polynomial runtime for bellman equations of discrete time high-dimensional stochastic optimal control problems. arXiv 2303.03390, 2023.

[3] 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.

[4] Denis Belomestny and John Schoenmakers. Primal-dual regression approach for markov decision processes with general state and action space. arXiv 2210.00258, 2022.

[5] Robert L Bray. A comment on “Using randomization to break the curse of dimensionality”. *Econometrica*, 90(4):1915–1929, 2022.

[6] Chef-Seng Chow and John N Tsitsiklis. The complexity of dynamic programming. *Journal of complexity*, 5(4):466–488, 1989.

[7] Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

[8] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine learning*, 49:193–208, 2002.

[9] Erich Novak and Henryk Woźniakowski. *Tractability of multivariate problems. Vol. 1: Linear information*. European Mathematical Society, 2008.

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

[11] John Rust. Using randomization to break the curse of dimensionality. *Econometrica: Journal of the Econometric Society*, pages 487–516, 1997.

[12] Csaba Szepesvári. Efficient approximate planning in continuous space markovian decision problems. *AI Communications*, 14(3):163–176, 2001.

[13] 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.

[14] Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer Science & Business Media, 1999.