Max-Plus Matching Pursuit for Deterministic Markov Decision Processes

Francis Bach

Max-Plus Matching Pursuit for Deterministic Markov Decision Processes

Francis Bach
Inria
Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University, Paris, France
francis.bach@ens.fr

June 19, 2019

Abstract

We consider deterministic Markov decision processes (MDPs) and apply *max-plus* algebra tools to approximate the value iteration algorithm by a smaller-dimensional iteration based on a representation on dictionaries of value functions. The set-up naturally leads to novel theoretical results which are simply formulated due to the max-plus algebra structure. For example, when considering a fixed (non adaptive) finite basis, the computational complexity of approximating the optimal value function is not directly related to the number of states, but to notions of covering numbers of the state space. In order to break the curse of dimensionality in factored state-spaces, we consider adaptive basis that can adapt to particular problems leading to an algorithm similar to matching pursuit from signal processing. They currently come with no theoretical guarantees but work empirically well on simple deterministic MDPs derived from low-dimensional continuous control problems. We focus primarily on deterministic MDPs but note that the framework can be applied to all MDPs by considering measure-based formulations.

1 Introduction

Function approximation for Markov decision processes (MDPs) is an important problem in reinforcement learning. Simply extending classical representations from supervised learning is not straightforward because of the specific non-linear structure of MDPs; for example the linear parametrization of the value function is not totally adapted to the algebraic structure of the Bellman operator which is central in their analysis and involves "max" operations.

Following [24, 1] which applied similar concepts to problems in optimal control, we consider a different semi-ring than the usual ring $(\mathbb{R}, +, x)$, namely the *max-plus* semi-ring $(\mathbb{R} \cup \{-\infty\}, \oplus, \otimes) =$

$(\mathbb{R} \cup \{-\infty\}, \max, +)$. The new resulting algebra is natural for MDPs, as for example for deterministic discounted MDPs, the Bellman operator happens to be additive and positively homogeneous for the max-plus algebra.

In this paper, we explore classical concepts in linear representations in machine learning and signal processing, namely approximations from a finite basis and sparse approximations through greedy algorithms such as matching pursuit [23, 22, 32, 12], explore them for the max-plus algebra and apply it to deterministic MDPs with known dynamics (where the goal is to estimate the optimal value function). We make the following contributions, after briefly reviewing MDPs in Section 2:

- In Section 3, we apply *max-plus* algebra tools to approximate the value iteration algorithm by a smaller-dimensional iteration based on a representation on dictionaries of value functions.
- As shown in Section 4, the set-up naturally leads to novel theoretical results which are simply formulated due to the max-plus algebra structure. For example, when considering a fixed (non adaptive) finite basis, the computational complexity of approximating the optimal value function is not directly related to the number of states, but to notions of covering numbers.
- In Section 5, in order to circumvent the curse of dimensionality in factored state-spaces, we consider adaptive basis that can adapt to particular problems leading to an algorithm similar to matching pursuit. It currently comes with no theoretical guarantees but works empirically well in Section 6 on simple deterministic MDPs derived from low-dimensional control problems.

2 Markov Decision Processes

We consider a standard MDP [28, 31], defined by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, transition probabilities $p(\cdot|\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a discount factor $\gamma \in [0, 1)$. In this paper, we focus on the goal of finding (or approximating) the optimal value function $V_* : \mathcal{S} \to \mathbb{R}$, from which the optimal policy $\pi_* : \mathcal{S} \to \mathcal{A}$, that leads to the optimal sum of discounted rewards, can be obtained as $\pi_*(s) \in \arg\max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V_*(s')$. We assume that the transition probabilities and the reward function are known.

We denote by T the Bellman operator from $\mathbb{R}^{\mathcal{S}}$ to $\mathbb{R}^{\mathcal{S}}$ defined as, for a function $V : \mathcal{S} \to \mathbb{R}$,

$$TV(s) = \max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s').$$

The optimal value function V_* is the unique fixed point of T. In order to find an approximation of V_*, we simply need to find V such that $\|TV - V\|_\infty$ is small, as $\|V_* - V\|_\infty \leqslant (1-\gamma)^{-1}\|V - TV\|_\infty$ (see proof in App. A.1 taken from [4, Prop. 2.1]).

Value iteration algorithm. The usual value iteration algorithm considers the recursion $V_t = TV_{t-1}$, which converges exponentially fast to V_* if $\gamma < 1$ [31]. More precisely, if range(r) is defined as range$(r) = \max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} r(s,a) - \min_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} r(s,a)$, and we initialize at $V_0 = 0$,

we reach precision range$(r)\varepsilon(1-\gamma)^{-1}$ after at most $t = \log(1/\varepsilon)(\log(1/\gamma))^{-1} \leqslant \log(1/\varepsilon)(1-\gamma)^{-1}$ iterations [31]. In this paper, we consider discount factors which are close to 1, and the term dependent on $(\log(1/\gamma))^{-1}$ or $(1-\gamma)^{-1}$ will always be the leading one—this thus excludes from consideration sampling-based algorithms with a better complexity in terms of $|\mathcal{S}|$ and $|\mathcal{A}|$ but worse dependence on the discount factor γ (see, e.g., [30, 17] and references therein). Throughout this paper, we are going to refer to $\tau = (1-\gamma)^{-1}$ as the *horizon* of the MDP (this is the expectation of a random variable with geometric distribution proportional to powers of γ), which characterizes the expected number of steps in the future that need to be taken into account for computing rewards.

Deterministic MDPs. In this paper, we consider *deterministic MDPs*, i.e., MDPs for which given a state s and an action a, a deterministic state s' is reached. For these MDPs, choosing an action is equivalent to choosing the reachable state s'. Thus, the transition behavior if fully characterized by an edge set $\mathcal{E} \subset \mathcal{S} \times \mathcal{S}$, and we obtain the resulting reward function $\bar{r} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, where $\bar{r}(s, s')$ is the maximal reward from actions leading from state s to state s', which is defined to be $-\infty$ if $(s, s') \notin \mathcal{E}$. The Bellman operator then takes the form

$$TV(s) = \max_{s' \in \mathcal{S}} \bar{r}(s, s') + \gamma V(s').$$

We focus primarily on deterministic MDPs but note in Section 7 that the framework can be applied to all MDPs by considering a measure-based formulation [16, 19].

Factored state-spaces. We also consider large state spaces (typically coming from the discretization of control problems). A classical example will be factored state-spaces where $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_d$, but we do not assume in general factorized dependences of the reward function on certain variables like in factored MDPs [15].

3 Max-plus Algebra applied to Deterministic MDPs

Many works consider a regular linear parameterization of the value function (see, e.g., [14] and references therein), as $V(s) = \sum_{w \in \mathcal{W}} \alpha(w) w(s)$ where $\mathcal{W}$ is a set of basis functions $w : \mathcal{S} \to \mathbb{R}$. Following [1], we consider max-plus linear combinations. For more general properties of max-plus algebra, see, e.g., [2, 8, 10].

3.1 Max-plus-linear operators and max-plus-linear combinations

In this section, we consider algebraic properties of our problem within the max-plus-algebra. For deterministic MDPs, the key property is that the Bellman operator is additive and max-plus-homogeneous, that is, (a) $T(V + c) = TV + \gamma c$ for any constant c, which can be rewritten as $T(c \otimes V) = c^{\otimes\gamma} TV$ (where the equality $\gamma c = c^{\otimes\gamma}$ for $\gamma \in \mathbb{R}_+$, is an extension of the relationship $2c = c \otimes c$), and (b) $T(\max\{V, V'\}) = \max\{TV, TV'\}$, which can be rewritten as $T(V \oplus V') = TV \oplus TV'$, where all operations are taken element-wise for all $s \in \mathcal{S}$.

We now explore various max-plus approximation properties [10]. First, regular linear combinations become: $V(s) = \bigoplus_{w \in \mathcal{W}} \alpha(w) \otimes w(s) = \max_{w \in \mathcal{W}} \alpha(w) + w(s)$. We introduce the notation

$$\mathcal{W}\alpha(s) = \max_{w \in \mathcal{W}} \alpha(w) + w(s), \tag{1}$$

so that the equation above may be rewritten as $V = \mathcal{W}\alpha$. Inverses of max-plus-linear operator are typically not defined, but a weaker notion of pseudo-inverse (often called "residuation" [2]) can be defined due to the idempotence of $\otimes$. We thus consider the operator $\mathcal{W}^+ : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{W}}$ defined as

$$\mathcal{W}^+ V(w) = \min_{s \in \mathcal{S}} V(s) - w(s). \tag{2}$$

We have, for the pointwise partial order on $\mathbb{R}^{\mathcal{S}}$, $\mathcal{W}\alpha \leqslant V \Leftrightarrow \alpha \leqslant \mathcal{W}^+ V$, that is:

$$\forall s \in \mathcal{S},\ \mathcal{W}\alpha(s) \leqslant V(s) \Leftrightarrow \forall (s,w) \in \mathcal{S} \times \mathcal{W},\ \alpha(w) + w(s) \leqslant V(s) \Leftrightarrow \forall w \in \mathcal{W},\ \alpha(w) \leqslant \mathcal{W}^+ V(w).$$

Moreover, as shown in [10, 1], $\mathcal{W}\mathcal{W}^+\mathcal{W} = \mathcal{W}$ and $\mathcal{W}^+\mathcal{W}\mathcal{W}^+ = \mathcal{W}^+$, and thus $\mathcal{W}^+$ plays a role of pseudo-inverse, and $\mathcal{W}\mathcal{W}^+$ a role of projection on the image of $\mathcal{W}$; moreover, $\mathcal{W}\mathcal{W}^+ V \leqslant V$ for all V, and $\mathcal{W}\mathcal{W}^+$ is idempotent and non-expansive for the ℓ_∞-norm.

One approximation algorithm would be to replace $V_{t+1} = TV_t$ by $V_{t+1} = \mathcal{W}\mathcal{W}^+ TV_t$, that is, if we consider V_t of the form $V_t = \mathcal{W}\alpha_t$, then V_{t+1} would be of the form $\mathcal{W}\alpha_{t+1}$ where

$$\alpha_{t+1}(w) = \mathcal{W}^+ T \mathcal{W} \alpha_t(w) = \min_{s \in \mathcal{S}} \max_{w' \in \mathcal{W}} \gamma \alpha_t(w') + Tw'(s) - w(s),$$

which requires to solve at each iteration an infimum problem over $\mathcal{S}$, which is computionally expensive as $O(|\mathcal{S}| \cdot |\mathcal{W}|)$, which is typically worse than $O(|\mathcal{E}|)$ for classical value iteration. We are thus looking for an extra approximation that will lead to a decomposition where after some compilation, the iteration complexity is only dependent on the number of basis functions.

3.2 Max-plus transposition

An important idea from [1] is to first project onto a different low-dimensional different image, with a projection which is efficient. In the regular functional linear setting, this would be equivalent to imposing equality only for certain efficient measurements. We thus define, given a set $\mathcal{Z}$ of functions z from $\mathcal{S}$ to $\mathbb{R}$ (which can be equal to $\mathcal{W}$ or not),

$$\mathcal{Z}^\top V(z) = \max_{s \in \mathcal{S}} V(s) + z(s). \tag{3}$$

The notation $\mathcal{Z}^\top$ comes from the following definition of dot-product; we define the max-plus dot-product between functions V and z from $\mathcal{S}$ to $\mathbb{R}$ (and more generally for all functions defined on $\mathcal{W}$ or $\mathcal{Z}$) as $\langle z | V \rangle = \bigoplus_{s \in \mathcal{S}} z(s) \otimes V(s) = \max_{s \in \mathcal{S}} z(s) + V(s)$. We then have for any $\beta \in \mathbb{R}^{\mathcal{Z}}$, $\langle \mathcal{Z}^\top V | \beta \rangle = \langle V | \mathcal{Z}\beta \rangle$, hence the transpose notation. We can also define the residuation as:

$$\mathcal{Z}^{\top +} \beta(s) = \min_{z \in \mathcal{Z}} \beta(z) - z(s), \tag{4}$$

so that $\mathcal{Z}^{\top +}\mathcal{Z}^\top$ goes from $\mathbb{R}^{\mathcal{S}}$ to $\mathbb{R}^{\mathcal{S}}$. The operator $\mathcal{Z}^{\top +}\mathcal{Z}^\top$ on functions from $\mathcal{S}$ to $\mathbb{R}$ is also the projection on the image of $\mathcal{Z}^{\top +}$; moreover, $\mathcal{Z}^{\top +}\mathcal{Z}^\top V \geqslant V$ for all V and $\mathcal{Z}^{\top +}\mathcal{Z}^\top$ is idempotent and non-expansive for the ℓ_∞-norm. Note that $\mathcal{Z}^{\top +}\beta = -\mathcal{Z}(-\beta)$ so that properties of $\mathcal{Z}^{\top +}\beta$ can be inferred from the ones of $\mathcal{Z}$. Similarly $\mathcal{Z}^\top V = -\mathcal{Z}^+(-V)$ and, $\mathcal{Z}^{\top +}\mathcal{Z}^\top V = -\mathcal{Z}\mathcal{Z}^+(-V)$.

3.3 Reduced value iteration

Extending [1] to MDPs, we consider of $V_{t+1} = TV_t$, the iteration $V_{t+1} = \mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top+}\mathcal{Z}^\top TV_t$. If V_t is represented as $\mathcal{W}\alpha_t$, then V_{t+1} is represented as $\mathcal{W}\alpha_{t+1}$ where $\alpha_{t+1} = \mathcal{W}^+\mathcal{Z}^{\top+}\mathcal{Z}^\top T\mathcal{W}\alpha_t$, which we can decompose as $\beta_{t+1} = \mathcal{Z}^\top T\mathcal{W}\alpha_t$ and $\alpha_{t+1} = \mathcal{W}^+\mathcal{Z}^{\top+}\beta_{t+1} = (\mathcal{Z}^\top\mathcal{W})^+\beta_{t+1}$.

The key point is then that the two operators $\mathcal{Z}^\top T\mathcal{W} : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{Z}}$ and $\mathcal{W}^+\mathcal{Z}^{\top+} : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{W}}$, can be pre-computed at a cost that will be independent of the discount factor γ. Indeed, given $\langle z|w\rangle = \max_{s\in\mathcal{S}} z(s) + w(s)$ and $\langle z|Tw\rangle = \max_{s\in\mathcal{S}} z(s) + Tw(s)$, we have:

$$\mathcal{Z}^\top T\mathcal{W}\alpha(z) = \max_{s\in\mathcal{S}} T\mathcal{W}\alpha(s) + z(s) = \max_{s\in\mathcal{S}} \max_{w\in\mathcal{W}} \gamma\alpha(w) + Tw(s) + z(s) = \max_{w\in\mathcal{W}} \gamma\alpha(w) + \langle z|Tw\rangle$$
$$\mathcal{W}^+\mathcal{Z}^{\top+}\beta(w) = \min_{s\in\mathcal{S}} \mathcal{Z}^{\top+}\beta(s) - w(s) = \min_{s\in\mathcal{S}} \min_{z\in\mathcal{Z}} \beta(z) - z(s) - w(s) = \min_{z\in\mathcal{Z}} \beta(z) - \langle z|w\rangle.$$

Therefore, the computational complexity of the iteration is $O(|\mathcal{W}| \cdot |\mathcal{Z}|)$, once the $|\mathcal{W}| \cdot |\mathcal{Z}|$ values $\langle z|w\rangle = \max_{s\in\mathcal{S}} z(s)+w(s)$, and $\langle z|Tw\rangle = \max_{s\in\mathcal{S}} z(s)+Tw(s)$ have been computed. This requires some compilation time which is *independent* of γ and the final required precision. Note that if these values are computed up to some precision ε, then we get an overall extra approximation factor of $\varepsilon/(1-\gamma)$. The iterations then become

$$\beta_{t+1}(z) = \mathcal{Z}^\top T\mathcal{W}\alpha(z) = \max_{w\in\mathcal{W}} \gamma\alpha_t(w) + \langle z|Tw\rangle \quad (5)$$
$$\alpha_{t+1}(w) = \mathcal{W}^+\mathcal{Z}^{\top+}\beta_{t+1}(w) = \min_{z\in\mathcal{Z}} \beta_{t+1}(z) - \langle z|w\rangle. \quad (6)$$

As seen below, they correspond to γ-contractant operators and are thus converging exponentially fast. In the worst case (full graph), the complexity is $O(|\mathcal{W}| \cdot |\mathcal{Z}|)$. Moreover, as presented below, a good approximation of V_* by $\mathcal{W}$ and $\mathcal{Z}^\top$ leads to an approximation guarantee (see proof in App. A.2).

Proposition 1 *(a) The operator $\hat{T} = \mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top+}\mathcal{Z}^\top T$ is γ-contractive and has a unique fixed point V_∞. (b) If $\|\mathcal{W}\mathcal{W}^+V_* - V_*\|_\infty \leqslant \eta$ and $\|\mathcal{Z}^{\top+}\mathcal{Z}^\top V_* - V_*\|_\infty \leqslant \eta$, then $\|V_\infty - V_*\|_\infty \leqslant \frac{2\eta}{1-\gamma}$.*

Therefore, an approximation guarantee for V_* translates to an approximation error multiplied by the horizon $\tau = (1-\gamma)^{-1}$. Thus large horizons τ (i.e., large γ) will degrade performance (see examples in Figure 2). If we consider a discount factor of γ^ρ (corresponding to the operator T^ρ instead of T, for $\rho > 1$, see below), in the result above, τ is replaced by $(1 + \tau/\rho)$ (see proof in App. A.2).

Algorithmic complexity. After t steps of the approximate algorithms in Eq. (5) and Eq. (6), starting from zero, we get V_t such that $\|\hat{T}V_t - V_t\|_\infty \leqslant \gamma^t\|\hat{T}V_0 - V_0\|_\infty \leqslant \gamma^t\text{range}(r)$, and thus such that $\|V_t - V_\infty\|_\infty \leqslant \gamma^t\text{range}(r)(1-\gamma)^{-1}$, with the overall bound $\|V_t - V_*\|_\infty \leqslant (2\eta + \gamma^t\text{range}(r))(1-\gamma)^{-1}$ if the assumption (b) in Prop. 1 is satisfied. Thus to get an error of $\varepsilon\text{range}(r)\tau$ for a fixed ε, it is sufficient to have an approximation error $\eta = \text{range}(r)\varepsilon/4$ and a number of iterations $t = \log(2/\varepsilon)(1-\gamma)^{-1} = \log(2/\varepsilon)\tau$. Thus, the overall complexity will be proportional to $|\mathcal{W}| \cdot |\mathcal{Z}| \cdot \log(2/\varepsilon) \cdot \tau$ in addition to the compilation time required to compute the $|\mathcal{W}| \cdot |\mathcal{Z}|$ values $\langle z|w\rangle$ and $\langle z|Tw\rangle$ (which is independent of γ).

Extensions. We can apply the reasoning above to T^ρ and replace γ by γ^ρ, with then ρ fewer iterations in the leading terms, but a more expensive compilation time. The horizon τ is then equal to $\tau_\rho = (1-\gamma^\rho)^{-1}$ which is equivalent to τ/ρ for $\rho = o(\tau)$. Moreover, when $\rho = O(\tau)$, the approximation error is reduced and we only need to have $\eta = \rho \cdot \text{range}(r)\varepsilon/4$. This will also be helpul within matching pursuit (see Section 5).

4 Approximation by Max-plus Operators

In this section, we consider several classes of functions $\mathcal{W}$ and $\mathcal{Z}$, with their associated approximation properties, that are needed for Prop. 1 to provide interesting guarantees. Some of these sets are already present in [1] (distance and squared distance functions), whereas others are new (piecewise constant approximations and Bregman divergences). We present examples of such approximations in Figure 1 and Figure 2, where we note a difference between the approximation of some function V by $\mathcal{W}\mathcal{W}^+$ or $\mathcal{Z}^{\top+}\mathcal{Z}^\top V$ and their approximation with the same basis functions *within an MDP*, as obtained from Prop. 1 (with ρ very large, it would be equivalent, but not for small ρ).

4.1 Indicator functions and piecewise constant approximations

We consider functions $w : \mathcal{S} \to \mathbb{R}$ of the form $w(s) = 0$ if $s \in A(w)$ and $-\infty$ otherwise, where $A(w)$ is a subset of $\mathcal{S}$, with $(A(w))_{w\in\mathcal{W}}$ forming a partition of $\mathcal{S}$. For simplicity, we assume here that $\mathcal{Z} = \mathcal{W}$. The image of $\mathcal{W}$ is the set of piecewise constant functions with respect to this partition. Given a function V, $\mathcal{W}\mathcal{W}^+V$ is the lower approximation of V by a piecewise constant function, while $\mathcal{Z}^{\top+}\mathcal{Z}^\top V$ is the upper approximation of V by a piecewise constant function (see Figure 1).

Reduced iteration. Since the image of $\mathcal{Z}^{\top+}$ and $\mathcal{W}$ are the same, the iteration reduces to $V_t = \mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top+}\mathcal{Z}^\top T V_{t-1} = \mathcal{Z}^{\top+}\mathcal{Z}^\top T V_{t-1}$. Thus, representing V_t as $\mathcal{Z}^{\top+}\alpha_t$ with $\alpha_t \in \mathbb{R}^{\mathcal{Z}}$, we get

$$\alpha_{t+1}(w) = \max_{w'\in\mathcal{W},(w,w')\in\mathcal{E}(A)} R(w,w') + \gamma\alpha_t(w'), \tag{7}$$

where $R(w,w') = \max_{s\in A(w)} \max_{s'\in A(w')} \bar{r}(s,s')$ and $\mathcal{E}(A)$ is the set of (w,w') such that $R(w,w')$ is finite. This is is exactly a deterministic MDPs on the clusters. The complexity of the algorithm depends on some *compilation* time, in order to compute the reduced matrix R (or the one corresponding to T^ρ), and some running-time *per iteration*. These depends whether we consider a dense graph or a sparse d-dimensional graphs (corresponding to a d-dimensional grid). The complexities are presented below (with all constants dependent on d removed), and proved in App. C.

	Compilation time	Iteration time
Value iteration (dense)	$\lvert\mathcal{S}\rvert^3 \log\rho$	$\lvert\mathcal{S}\rvert^2$
Value iteration (sparse)	$\lvert\mathcal{S}\rvert\rho^{2d}$	$\lvert\mathcal{S}\rvert\rho^d$
Reduced MDP (dense)	$\lvert\mathcal{S}\rvert^2 + \min\{\lvert\mathcal{S}\rvert^3\log\rho, \lvert\mathcal{W}\rvert\lvert\mathcal{S}\rvert^2\rho\}$	$\lvert\mathcal{W}\rvert^2$
Reduced MDP (sparse)	$\lvert\mathcal{S}\rvert + \min\{\lvert\mathcal{S}\rvert\rho^{2d}, \lvert\mathcal{W}\rvert\lvert\mathcal{S}\rvert\rho\}$	$\lvert\mathcal{W}\rvert\rho^d$

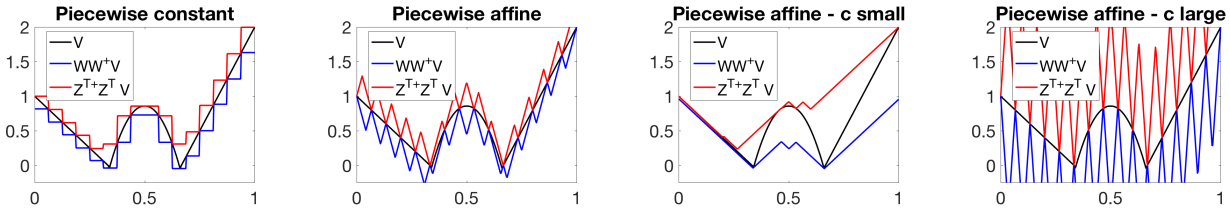

Figure 1: Approximation of a function V with different finite basis with 16 elements. One-dimensional case. From left to right: piece-wise constant basis function, piecewise affine basis functions with well chosen value of c, then too small and too large.

For all cases above, the optimization error (to approximate V_∞ in Prop. 1) is $\mathrm{range}(r)\tau e^{-\rho t/\tau}$; thus, in the sparse case, the number of iterations to achieve precision $\mathrm{range}(r)\tau\varepsilon$ is proportional to $(\tau/\rho)\log(1/\varepsilon)$. Below, for simplicity, we are only considering trade-offs for the sparse graph cases. For plain value iteration, having $\rho > 1$ larger than one could be beneficial only when $d = 1$ (otherwise, the compilation time scales as $|\mathcal{S}|\rho^{2d}$ and the iteration running time as $|\mathcal{S}|\rho^{d-1}$, which are both increasing in ρ). When using the clustered representation, it seems that the situation is the same, but as shown later the approximation error comes into play and larger values of ρ could be useful (both for running time with fixed dictionaries and for stability of matching pursuit).

Approximation properties. See the proof of the following proposition in App. B.1, which provided approximation guarantees for Prop. 1. We assume that $\mathcal{W}$ is composed of piecewise constant functions.

Proposition 2 *If we denote by $\eta(n, \mathcal{S})$ the smallest radius of a cover of $\mathcal{S}$ by balls of a given radius, by considering the Voronoi partition associated with the n ball centers, we get:*

$$\min_{\alpha\in\mathbb{R}^w} \|V - \mathcal{W}\alpha\|_\infty \leqslant \|V - \mathcal{W}\mathcal{W}^+V\|_\infty \leqslant \mathrm{Lip}_p(V)[2\eta(n, \mathcal{S})]^p, \tag{8}$$

where $\mathrm{Lip}_p(V)$ is the p-th order Hölder continuity constant of V (i.e., so that for all $s, s' \in \mathcal{S}$, $|V(s) - V(s')| \leqslant \mathrm{Lip}_p(V)d(s, s')^p$).

While for factored spaces and with no assumptions on the optimal value function V_* beyond continuity, $\eta(n, \mathcal{S})$ may not scale well, the approximation could be much better in special cases (e.g., when V_* depends only on a subset of variables). For factored spaces $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_d$, with a ℓ_∞-metric, then $\eta_n(\mathcal{S}) \leqslant \max_{i\in\{1,\dots,d\}} \eta(n^{1/d}, \mathcal{S}_i)$. If each factor $\mathcal{S}_i$ is simple (e.g., a chain graph), then $\eta(n, \mathcal{S}_i) = O(1/n)$, and we get $\eta_n(\mathcal{S}) = O(1/n^{1/d})$. Here, we do not escape the curse of dimensionality.

Going beyond exponential complexity. In order to avoid the rate $O(1/n^{1/d})$ above, we can consider further assumptions: if V_* is well approximated by a piecewise constant function with respect to a specific partition (dedicated to V_*), the bound in Eq. (8) can be greatly reduced.

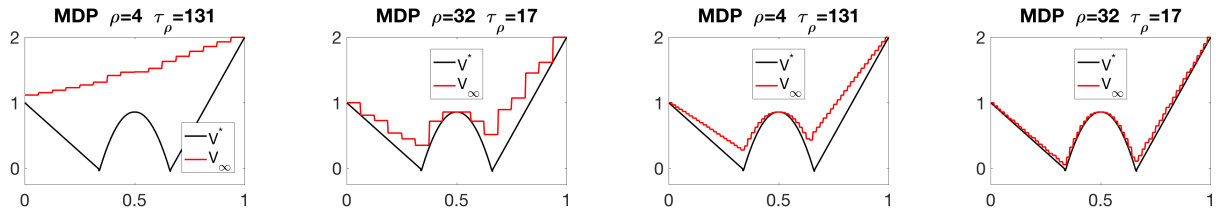

Figure 2: Approximation of a function V with different finite basis with 16 or 64 elements, *within the MDP* (i.e., leading to V_∞ in Prop. 1), and with values of ρ that are 4 and 32. One-dimensional case. From left to right: $(n = 16, \rho = 4)$, $(n = 16, \rho = 32)$, $(n = 64, \rho = 4)$, $(n = 16, \rho = 32)$.

Note that the approximation with a small number of basis functions here is the same for $\mathcal{W}\mathcal{W}^\top$ and ${\mathcal{Z}^\top}^+\mathcal{Z}^\top$ when $\mathcal{Z} = \mathcal{W}$ (this will not be the case in Section 4.2 below).

We can get a reduced set of basis functions, if for example V_* depends only on k variables, then there is a partition for which the error in Eq. (8) is of the order $O(1/n^{1/k})$, and we can then escape the curse of dimensionality if k is small and we can find these k relevant variables. This will be done algorithmically in Section 5, by considering a greedy algorithm in $[0,1]^d$ with sets $\prod_{i=1}^d \left[\frac{j_i}{2^{k_i}}, \frac{j_i+1}{2^{k_i}}\right]$ and a split according to a single dimension at every iteration. This variable selection does not need to be global and the covering number can benefit from local independences.

Optimal choices of hyperparameters. The approximation error from Prop. 2 is of order $(1+\tau/\rho)\mathrm{Lip}(V_*)|\mathcal{W}|^{-1/k}$, and thus, for $\rho = O(\tau)$, in order to achieve a final approximation error of $\mathrm{range}(r)\tau\varepsilon$, we need $|\mathcal{W}|$ to be of the order of $[\mathrm{Lip}(V_*)\mathrm{range}(r)^{-1}\varepsilon^{-1}\rho^{-1}]^k$. Without compilation time this leads to a complexity proportional to $\left[\mathrm{Lip}(V_*)\mathrm{range}(r)^{-1}\varepsilon^{-1}\right]^k \tau\rho^{d-k-1}\log(1/\varepsilon)$; thus, it is advantageous to have large values of ρ when $k = d$ (full dependence). This has to be mitigated by the compilation time that is less than $|\mathcal{S}|\rho^{2d}$, which grows with ρ and d. We will see in Section 6 that larger values are also better for a good selection of dictionary elements in matching pursuit.

4.2 Distance functions and piecewise-affine functions

Following [1], we consider functions of the form $w(s) = -c \cdot d(s, w)$ for $w \in \mathcal{S}$, and d a distance on $\mathcal{S}$. Thus $\mathcal{W}$ may be identified to a subset of $\mathcal{S}$. When $\mathcal{S}$ is a subset of $\mathbb{R}^d$ and with ℓ_1 or ℓ_∞ metrics, we get piecewise affine functions (see Figure 1). We then have (see proof in App. B.2):

Proposition 3 *(a) If $\mathcal{W} = \mathcal{S}$ and $c \geqslant \mathrm{Lip}(V)$ (Lipschitz-constant of V), then $\mathcal{W}\mathcal{W}^+V = V$. (b) If $c \geqslant \mathrm{Lip}_1(V)$, then $\|V - \mathcal{W}\mathcal{W}^+V\|_\infty \leqslant 2c \cdot \max_{s\in\mathcal{S}} \min_{w\in\mathcal{W}} d(w, s)$.*

We thus get an approximation guarantee for Prop. 1 from a good covering of $\mathcal{S}$. The approximation also applies to $\mathcal{Z}^\top$. In terms of approximation guarantees, then we need to cover $\mathcal{S}$ with sufficiently many elements of $\mathcal{W}$, and thus with n points in $\mathcal{W}$ we get an approximation of exactly the same order than for clustered partitions (but with the need to know the Lipschitz constant $\mathrm{Lip}_1(V)$ to set the extra parameter).

In terms of approximation by a finite basis, a problem here is that being well approximated by some $\mathcal{W}\alpha$, for $|\mathcal{W}|$ small, does not mean that one can be well approximated by $\mathcal{Z}\beta$, with $|\mathcal{Z}|$ small (it is in one dimension, not in higher dimensions). Moreover, these functions are not local so computing $\langle z|w\rangle$ and $\langle z|Tw\rangle$ could be harder. Indeed, the compile time is $O((|\mathcal{E}| + |\mathcal{Z}| \cdot |\mathcal{S}|) \cdot |\mathcal{W}|)$ while each iteration of the reduced algorithm is $O(|\mathcal{W}| \cdot |\mathcal{Z}|)$.

Distance functions for variable selection. To allow variable selection, we can consider a more general family of distance functions, namely, function of the form $w(s) = -\max_{i\in A} d(s_i, w_i)$ or $w(s) = -\sum_{i\in A} d(s_i, w_i)$ where $A \subset \{1, \ldots, d\}$ for factored spaces. We consider the second option in our experiments in Section 6.

4.3 Extensions

Smooth functions. As outlined by [1], in the Euclidean continuous case, we can consider square distance funtions, which we generalize to Bregman divergences in Appendix B.3. Note that these will approximate smooth functions with more favorable approximation guarantees (but note that in MDPs obtained from the discretization of a continuous control problem, optimal value functions are non-smooth in general [11]). Finally, other functions could be considered as well, such as linear functions restricted to a subset, or ridge functions of the form $w(s) = \sigma(A^\top s + b)$.

Random functions. If we select a random set of points from a Poisson process with fixed intensity, then the maximal diameter of the associated random Voronoi partition has a known scaling [9], similar to the covering number up to logarithmic terms. Thus, we can use distance functions from Section 4.2 (where we can sample the constant c from an exponential distribution) or clusters from Section 4.1, or also squared distances like in Section B.3.

5 Greedy Selection by Matching Pursuit

The sets of functions proposed in Section 4 still suffer from the curse of dimensionality, that is, the cardinalities $|\mathcal{W}|$ and $|\mathcal{Z}|$ should still scale (at most linearly) with $|\mathcal{S}|$, and thus exponentially in dimension if $|\mathcal{S}|$ is a factored state-space. We consider here greedily selecting new functions, to use dictionaries adapted to a given MDP, mimicking the similar approach of sparse decompositions in signal processing and unsupervised learning [23, 22].

We assume that we are given $\mathcal{W}$ and $\mathcal{Z}$, and we want to test new sets $\mathcal{W}_{\text{new}}$ and $\mathcal{Z}_{\text{new}}$, which are close to $\mathcal{W}$ and $\mathcal{Z}$ (typically one function w split in two, that is, $w = w_1 \otimes w_2 = \max\{w_1, w_2\}$ or one additional function w_{new}). Note that pruning [13] could also be considered as well.

We assume that we have the optimal $\alpha \in \mathbb{R}^{\mathcal{W}}$ and $\beta \in \mathbb{R}^{\mathcal{Z}}$ (after convergence of the iterations defined in Eq. (5) and Eq. (6)), such that $\beta = \mathcal{Z}^\top T\mathcal{W}\alpha$ and $\alpha = \mathcal{W}^+\mathcal{Z}^{\top+}\beta$. We denote by $V = \mathcal{W}\alpha$ and $U = \mathcal{Z}^{\top+}\beta = \mathcal{Z}^{\top+}\mathcal{Z}^\top TV$. The criterion will be based on considering the best improvement of the new projections $\mathcal{W}_{\text{new}}\mathcal{W}^+_{\text{new}}$ and $\mathcal{Z}^{\top+}_{\text{new}}\mathcal{Z}^\top{}_{\text{new}}$ to the relevant function.

Criterion for $\mathcal{W}_{\rm new}$. Given the current difference between U and its projection $U - \mathcal{W}\mathcal{W}^+U = U - V \geqslant 0$, the criterion to minimize is $\|U - \mathcal{W}_{\rm new}\mathcal{W}_{\rm new}^+U\|$, for a given norm $\|\cdot\|$. If $\mathcal{W}_{\rm new} = \mathcal{W} \cup \{w_{\rm new}\}$, we have $\mathcal{W}_{\rm new}\mathcal{W}_{\rm new}^+U(s) = \max\big\{w_{\rm new}(s) - \langle w_{\rm new}| - U\rangle, V(s)\big\}$, and our criterion thus becomes $\|U - \mathcal{W}_{\rm new}\mathcal{W}_{\rm new}^+U\|_\infty = \max_{s\in\mathcal{S}} \min\big\{U(s) - V(s), U(s) - w_{\rm new}(s) + \langle w_{\rm new}| - U\rangle\big\}$ for the ℓ_∞-norm.

Criterion for $\mathcal{Z}_{\rm new}$. Given the current difference between TV and its projection $\mathcal{Z}^{\top+}\mathcal{Z}^\top TV - TV = U - TV \geqslant 0$, the criterion to minimize is $\|\mathcal{Z}_{\rm new}^{\top+}\mathcal{Z}^\top{}_{\rm new}TV - TV\|_\infty$. If $\mathcal{Z}_{\rm new} = \mathcal{Z} \cup \{z_{\rm new}\}$, we have $\mathcal{Z}_{\rm new}^{\top+}\mathcal{Z}^\top{}_{\rm new}TV(s) = \min\big\{ - z_{\rm new}(s) + \langle TV|z_{\rm new}\rangle, U(s)\big\}$, and and our criterion becomes $\|\mathcal{Z}_{\rm new}^{\top+}\mathcal{Z}^\top{}_{\rm new}TV - TV\|_\infty = \max_{s\in\mathcal{S}} \min\big\{U(s) - TV(s), -TV(s) - z_{\rm new}(s) + \langle TV|z_{\rm new}\rangle\big\}$.

Like in regular matching pursuit for classical linear approximation, allowing full flexibility for $z_{\rm new}$ or $w_{\rm new}$ would lead to trivial choices, here $z_{\rm new} = -TV$ and $w_{\rm new} = U$: in our MDP situation, we essentially end up with few changes compared to plain value iteration as the new atoms are just obtained by using our own approximation U or applying the transition operator to another approximation (i.e., TV). Thus, within matching pursuit, in order to benefit from a reduction in global approximation error, ensuring the diversity of the dictionary of atoms which we select from is crucial. To go beyond selection from a finite set (and learn a few parameters), we propose a convex relaxation of estimating $z_{\rm new}$ or $w_{\rm new}$ in App. D.

Special case of partitions. In this case , we have $\mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top+}\mathcal{Z}^\top = \mathcal{Z}^{\top+}\mathcal{Z}^\top$, and thus we only need to consider $\|\mathcal{Z}_{\rm new}^{\top+}\mathcal{Z}^\top{}_{\rm new}TV - TV\|_\infty$ above. A simplification there is to consider the current set $A(w)$ which contains s attaining the ℓ_∞ bound above, and only try to split this cluster.

Related work. Variable selection within an MDP could be seen as a special case of factored MDPs [15] where the reward function depends on a subset of variables, it would be interesting to study theoretically more complex dependences. Beyond factored MDPs, variable selection and more generally function approximation within MDPs has been considered in several works (see, e.g., [5, 27, 18, 21, 7, 20]), but do not provide the theoretical analysis that we provide here or consider linear function approximations, while we consider max-plus combinations; [26] considers variable resolution within the discretization of an optimal control problem, but not for a generic Markov decision process.

6 Experiments

All experiments can be exactly reproduced with the Matlab code which can be obtained from the author's webpage.[1]

Simulations on discretizations of control problems on $[0,1]$. We consider the discretization $\mathcal{S} = \big\{(i-1)(S-1)^{-1},\ i \in \{1,\dots,S\}\big\}$ of $[0,1]$, which we can identify to $\{1,\dots,S\}$. Given

[1] `www.di.ens.fr/~fbach/maxplus.zip`

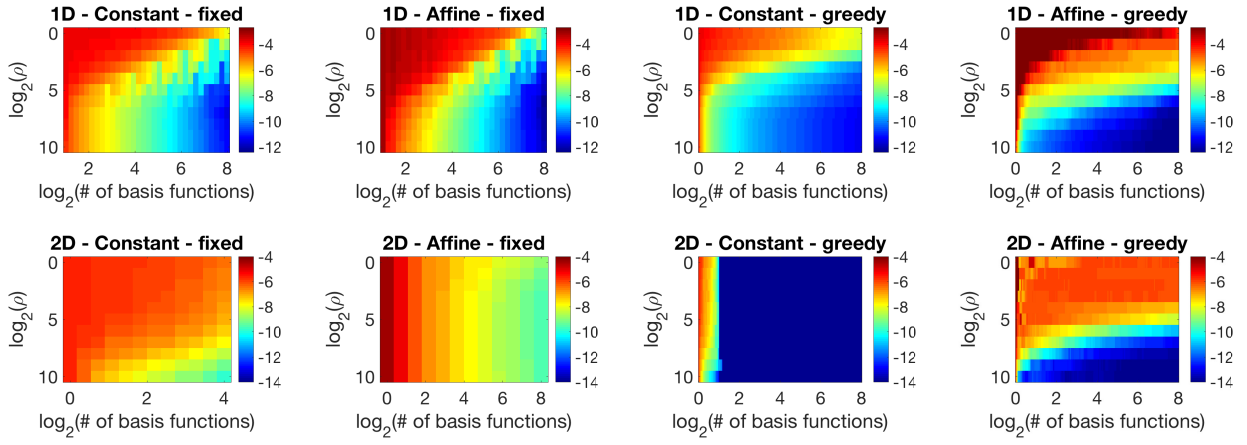

Figure 3: Approximation error $\|V - V_*\|_1$ of a function V as a function of ρ and the number of basis functions, for piecewise constant or affine functions. Top: One-dimensional case, bottom: two-dimensional case. Left: fixed non adaptive basis, right: greedy (matching pursuit).

the set of actions $\mathcal{A} = \{-1, 1\}$, following [25], we consider the dynamical systems $dx/dt = a$ (going left or right). The goal of the control problem is to maximize $\int_0^u \eta^t b(x(t))dt + \eta^t B(x(u))$ where u is the exit time of x from $[0, 1]$ (i.e., reaching the boundary). We then define the value function $V(x)$ as the supremum over controls $a(\cdot)$ starting from $x(0) = x$.

Then $V(x)$ satisfies the Hamilton-Jacobi-Bellman (HJB) equation [11, 25]: $V(x)\log\eta + |V'(x)| + b(x) = 0$, with the boundary condition $V(x) \geqslant B(x)$ for $x \in \{0, 1\}$. With the discretization above, with a step $\delta = 1/(S-1)$, we have the absorbing states 1 and S. We thus need to construct the reward from s to $s+1$, and from s to $s-1$, for any $s \in \{2, \dots, S-1\}$ equal to the reweighted function $\delta b(x)$ for the reached state, and a discount factor equal to $\gamma = \eta^\delta$. From states 1 and S, no move can be made and the reward $r(S, S)$ is equal to $(1-\gamma)B(1)$ and $r(1, 1) = (1-\gamma)B(0)$. When δ goes to zero, then the MDP solution converges to the solution of the optimal control problem.

In our example in Figure 1 and Figure 2, we consider pairs (b, V) that satisfy the HJB equation. In Figure 3 (top), for the same problem, we consider the performance of our greedy method (matching pursuit with an ℓ_1-norm criterion) or fixed basis for piecewise constant functions from Section 4.1 and piecewise affine functions from Section 4.2, when the horizon τ_ρ varies (from values of ρ) and the number of basis functions varies. We can see (a) the benefits of piecewise affine over piecewise constant functions in the sets $\mathcal{W}$ and $\mathcal{S}$ (that is, better approximation properties with the same number of basis functions), (b) the beneficial effect of larger ρ (i.e., using T^ρ instead of T), in particular for greedy techniques where the selection of good atoms does require a larger ρ.

Simulations on discretizations of control problems on $[0, 1]^d$. We consider two-dimensional extensions where the optimal value function only depends on a single variable (see details and more experiments in Appendix E.2, with a full dependence on the two variables) and we show performance plots in Figure 3 (bottom). Because of the sparsity assumption, the benefits of

matching pursuit are greater than for the one-dimensional case (empirically, only relevant atoms are selected, and thus $\|V - V_*\|_1$ converges to zero faster as the number of basis functions increases).

7 Conclusion

In this paper, we have presented a max-plus framework for value function approximation with a greedy matching pursuit algorithm to select atoms from a large dictionary. While our current framework and experiments deal with low-dimensional deterministic MDPs, there are many avenues for further algorithmic and theoretical developments.

First, for non-deterministic MDPs, the Bellman operator is unfortunately not max-plus additive; however, there are natural extensions to general MDPs using measure-based formulations on probability measures on $\mathcal{S}$ [16, 19], where using a policy $\pi(\cdot|\cdot)$, one goes from a measure μ to $\mu'(s') = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s)\mu(s)$, with a reward going to μ and μ' equal to $\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r(s,a)\pi(a|s)\mu(s)$. The difficulty here is the increased dimensionality of the problem due a new problem defined on probability measures. Also, going beyond model-based reinforcement learning could be done by estimation of model parameters; the max-plus formalism is naturally compatible with dealing with confidence intervals. Finally, it would be worth exploring multi-resolution techniques which are common in signal processing [23], to deal with higher-dimensional problems, where short-term and long-term interactions could be partially decoupled.

Acknowledgements

We acknowledge support the European Research Council (grant SEQUOIA 724063).

References

[1] Marianne Akian, Stéphane Gaubert, and Asma Lakhoua. The max-plus finite element method for solving deterministic optimal control problems: basic properties and convergence analysis. *SIAM Journal on Control and Optimization*, 47(2):817–848, 2008.

[2] François Baccelli, Guy Cohen, Geert Jan Olsder, and Jean-Pierre Quadrat. *Synchronization and Linearity: an Algebra for Discrete Event Systems*. John Wiley & Sons, 1992.

[3] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[4] Dimitri P. Bertsekas. Weighted sup-norm contractions in dynamic programming: A review and some new applications. Technical Report LIDS-P-2884, Dept. Elect. Eng. Comput. Sci., Massachusetts Institute of Technology, 2012.

[5] Dimitri P. Bertsekas and David A. Castanon. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE Transactions on Automatic Control*, 34(6):589–598, 1989.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] Lucian Busoniu, Damien Ernst, Bart De Schutter, and Robert Babuska. Cross-entropy optimization of control policies with adaptive basis functions. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(1):196–209, 2011.

[8] Peter Butkovič. *Max-linear Systems: Theory and Algorithms*. Springer Science & Business Media, 2010.

[9] Pierre Calka and Nicolas Chenavier. Extreme values for characteristic radii of a Poisson-Voronoi tessellation. *Extremes*, 17(3):359–385, 2014.

[10] Guy Cohen, Stéphane Gaubert, and Jean-Pierre Quadrat. Duality and separation theorems in idempotent semimodules. *Linear Algebra and its Applications*, 379(1):395–422, 2004.

[11] Michael G. Crandall and Pierre-Louis Lions. Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 277(1):1–42, 1983.

[12] Bogdan Dumitrescu and Paul Irofti. *Dictionary Learning Algorithms and Applications*. Springer, 2018.

[13] Stéphane Gaubert, William McEneaney, and Zheng Qu. Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms. In *Conference on Decision and Control and European Control Conference*, pages 1054–1061, 2011.

[14] Matthieu Geist and Olivier Pietquin. A brief survey of parametric value function approximation. Technical report, Supélec, 2010.

[15] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

[16] Onésimo Hernández-Lerma and Jean-Bernard Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, volume 30. Springer Science & Business Media, 2012.

[17] Sham Kakade, Mengdi Wang, and Lin F. Yang. Variance reduction methods for sublinear reinforcement learning. Technical Report 1802.09184, arXiv, 2018.

[18] Philipp W. Keller, Shie Mannor, and Doina Precup. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

[19] Jean-Bernard Lasserre, Didier Henrion, Christophe Prieur, and Emmanuel Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control and Optimization*, 47(4):1643–1666, 2008.

[20] De-Rong Liu, Hong-Liang Li, and Ding Wang. Feature selection and feature learning for high-dimensional batch reinforcement learning: A survey. *International Journal of Automation and Computing*, 12(3):229–242, 2015.

[21] Sridhar Mahadevan and Mauro Maggioni. Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 8(Oct):2169–2231, 2007.

[22] Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[23] Stéphane Mallat. *A Wavelet Tour of Signal Processing: the Sparse Way*. Academic Press, 2008.

[24] William M. McEneaney. Error analysis of a max-plus algorithm for a first-order HJB equation. In *Stochastic Theory and Control*, pages 335–351. Springer, 2002.

[25] Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40(3):265–299, 2000.

[26] Rémi Munos and Andrew Moore. Variable resolution discretization in optimal control. *Machine Learning*, 49(2-3):291–323, 2002.

[27] Relu Patrascu, Pascal Poupart, Dale Schuurmans, Craig Boutilier, and Carlos Guestrin. Greedy linear value-approximation for factored Markov decision processes. In *Proc. AAAI*, 2002.

[28] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[29] Joan Serra-Sagristà. Enumeration of lattice points in ℓ_1-norm. *Information Processing Letters*, 76(1-2):39–44, 2000.

[30] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, 2018.

[31] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[32] Vladimir Temlyakov. *Sparse Approximation with Bases*. Springer, 2015.

A Value iteration and reduced value-iteration convergence

In this appendix, we provide short proofs for the lemmas and propositions of the main paper related to (reduced) value iteration.

A.1 Approximation of the value function through approximate fixed points

This is taken from [4, Prop. 2.1] and presented because the proof structure is used later.

Lemma 1 ([4]) *If* $\|V - TV\|_\infty \leqslant \varepsilon$, *then* $\|V_* - V\|_\infty \leqslant \varepsilon(1-\gamma)^{-1}$.

Proof Consider $\varepsilon_t = \|V - T^t V\|_\infty$. We have $\varepsilon_t \leqslant \|TT^{t-1}V - TV\|_\infty + \|TV - V\|_\infty \leqslant \gamma\varepsilon_{t-1} + \varepsilon$, because T is γ-contractive. This leads to $\varepsilon_t \leqslant \sum_{u=0}^{t-1} \gamma^u \varepsilon \leqslant \frac{\varepsilon}{1-\gamma}$, thus $\|V - V_*\|_\infty \leqslant \frac{\varepsilon}{1-\gamma}$ by letting t tend to infinity. ■

A.2 Proof Prop. 1

(a) This is consequence of the non-expansiveness of $\mathcal{W}\mathcal{W}^+$ and $\mathcal{Z}^{\top +}\mathcal{Z}^\top$, and the γ-contractiveness of T.
(b) We have: $\|\hat{T}V_* - V_*\|_\infty = \|\mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top +}\mathcal{Z}^\top V_* - V_*\|_\infty \leqslant \|\mathcal{W}\mathcal{W}^+\mathcal{Z}^{\top +}\mathcal{Z}^\top V_* - \mathcal{W}\mathcal{W}^+ V_*\|_\infty + \|\mathcal{W}\mathcal{W}^+ V_* - V_*\|_\infty$. Using the non-expansivity of $\mathcal{W}\mathcal{W}^+$, we get

$$\|\hat{T}V_* - V_*\|_\infty \leqslant \|\mathcal{Z}^{\top +}\mathcal{Z}^\top V_* - V_*\|_\infty + \|\mathcal{W}\mathcal{W}^+ V_* - V_*\|_\infty \leqslant 2\eta.$$

This imples implies that $\|V_\infty - V_*\|_\infty \leqslant \frac{2\eta}{1-\gamma}$ using the same reasoning as in the proof of Lemma 1.

The result extends to assumptions on $\min_{\alpha\in\mathbb{R}^w} \|\mathcal{W}\alpha - V_*\|_\infty$ instead of $\|\mathcal{W}\mathcal{W}^+V_* - V_*\|_\infty$ (and similarly for $\mathcal{Z}$). Indeed, we have, with $\tilde{\alpha} = \mathcal{W}^+ V_*$,

$$\begin{aligned}
\min_{\alpha\in\mathbb{R}^{\mathcal{W}}} \|\mathcal{W}\alpha - V_*\|_\infty &\leqslant \|\mathcal{W}\tilde{\alpha} - V_*\|_\infty = \|\mathcal{W}\mathcal{W}^+V_* - V_*\|_\infty, \text{ and for any } \alpha, \\
\|\mathcal{W}\mathcal{W}^+V_* - V_*\|_\infty &\leqslant \|\mathcal{W}\mathcal{W}^+V_* - \mathcal{W}\alpha\|_\infty + \|\mathcal{W}\alpha - V_*\|_\infty \leqslant \|\mathcal{W}^+V_* - \alpha\|_\infty + \|\mathcal{W}\alpha - V_*\|_\infty
\end{aligned}$$

by non-expansivity of $\mathcal{W}$. By taking the infimum over α, we get that

$$\min_{\alpha\in\mathbb{R}^{\mathcal{W}}} \|\mathcal{W}\alpha - V_*\|_\infty \leqslant |\mathcal{W}\mathcal{W}^+V_* - V_*\|_\infty \leqslant 2 \min_{\alpha\in\mathbb{R}^{\mathcal{W}}} \|\mathcal{W}\alpha - V_*\|_\infty.$$

In order to prove the result when replacing T by T^ρ, we simply have to notice that the contraction factor of τ^ρ is γ^ρ and that for $\tau = (1-\gamma)^{-1} \geqslant 1$ and $\rho \geqslant 1$, we have

$$\frac{1}{1-(1-1/\tau)^\rho} \leqslant (1 + \tau/\rho).$$

B Approximation properties of max-plus-linear combinations

Here we provide proofs for approximation properties of max-plus linear combinations.

B.1 Proof of Prop. 2 (piecewise-constant functions)

We have, for any $s \in \mathcal{S}$:

$$V(s) - \mathcal{W}\mathcal{W}^+V(s) = V(s) - \max_{w\in\mathcal{W}} w(s) + \min_{s'\in\mathcal{S}} V(s') - w(s'),$$

which is thus always non-negative (and valid for any family $\mathcal{W}$ of functions).

Thus, for partition-based set of functions, if w is chosen so that $s \in A(w)$, then, w is the maximizer above and s' above is restricted to $A(w)$, because the value of w is $-\infty$ outside of $A(w)$. Thus

$$\begin{aligned} V(s) - \mathcal{W}\mathcal{W}^+ V(s) &= V(s) - \min_{s' \in A(w)} V(s') \\ &= \max_{s' \in A(w)} V(s') - V(s) \leqslant \max_{s' \in A(w)} |V(s') - V(s)| \\ &\leqslant \mathrm{Lip}_p(V) \max_{s', s'' \in A(w)} d(s', s'')^p. \end{aligned}$$

The result follows from the fact that with the choice of partition, $\max_{s', s'' \in A(w)} d(s', s'')$ is less than $2\eta(n, \mathcal{S})$.

B.2 Proof of Prop. 3 (distance functions)

The proof is similar to [1] (but slightly tighter). First, we have, for $c \geqslant \mathrm{Lip}_1(V)$, and any $w \in \mathcal{S}$:

$$V(w) = \min_{s' \in \mathcal{S}} V(s') + c \cdot d(s', w).$$

Indeed, (a) $V(w)$ is equal to $V(s') + c \cdot d(s', w)$ for $s' = w$, which implies $V(w) \geqslant \min_{s' \in \mathcal{S}} V(s') + c \cdot d(s', w)$; moreover, (b) $V(s') + c \cdot d(s', w) \geqslant V(w) - \mathrm{Lip}_1(V) \cdot d(s'w) + c \cdot d(s', w) \geqslant V(w)$, for all $s' \in \mathcal{S}$, which implies that $V(w) \leqslant \min_{s' \in \mathcal{S}} V(s') + c \cdot d(s', w)$.

Therefore, we get:

$$\begin{aligned} \|V - \mathcal{W}\mathcal{W}^+ V\|_\infty &= \max_{s \in \mathcal{S}} V(s) - \max_{w \in \mathcal{W}} -c \cdot d(w, s) + \min_{s' \in \mathcal{S}} V(s') + c \cdot d(s', w) \\ &= \max_{s \in \mathcal{S}} \min_{w \in \mathcal{W}} V(s) + c \cdot d(w, s) - V(w) \\ &\leqslant \max_{s \in \mathcal{S}} \min_{w \in \mathcal{W}} c \cdot d(w, s) + \mathrm{Lip}_1(V) \cdot d(w, s) \\ &\leqslant 2c \cdot \max_{s \in \mathcal{S}} \min_{w \in \mathcal{W}} d(w, s). \end{aligned}$$

B.3 Bregman basis functions smooth functions

In this section, we consider approximations of functions on a subspace of $\mathbb{R}^d$. These results can be used for discretizations of smooth problems.

Assuming that $\mathcal{S}$ is a convex compact subset of $\mathbb{R}^d$, and h is any convex function from $\mathbb{R}^d$ to $\mathbb{R}$, then, we consider the functions of the form $-\mathcal{D}_h(s, v) = -h(s) + h(v) - h'(v)^\top (s - v)$, which is the negative Bregman divergence associated to h, which is an extension of [1] from quadratic to all convex functions h. Since the functions need only be defined up to constants, we can reparameterize them with $w = -h'(v)$, and thus consider $w(s) = -h(s) + w^\top s$, where $\mathcal{W}$ is identified to a convex subset of $\mathbb{R}^d$. Then $\mathcal{W}\mathcal{W}^+ V$ is related to $(V + h)^{**} - h$, where g^* is the Fenchel-conjugate [6] of a function g. More precisely, we have (see proof in App. B.4).

Proposition 4 *(a) If $\mathcal{W}$ contains the domain of $(V+h)^*$, then $\mathcal{W}\mathcal{W}^+V = (V+h)^{**} - h$.*
(b) More generally, for any norm Ω on $\mathbb{R}^d$,

$$\|(V+h)^{**} - h - \mathcal{W}\mathcal{W}^+V\|_\infty \leqslant \mathrm{diam}_\Omega(\mathcal{S}) \max_{w \in \mathrm{domain}((V+h)^*)} \min_{w' \in \mathcal{W}} \Omega^*(w - w').$$

(c) If $(V+h)^{**} - g$ is convex, for a convex function g, then:

$$\|(V+h)^{**} - h - \mathcal{W}\mathcal{W}^+V\|_\infty \leqslant \max_{w \in \mathrm{domain}((V+h)^*)} \min_{w' \in \mathcal{W}} \mathcal{D}_{g^*}(w', w).$$

Thus, when $\mathcal{W}$ is not discretized, projection onto the image space of $\mathcal{W}$ corresponds to projection on the set of functions such that $V+h$ is convex: indeed, (a) implies that if $V+h$ is convex, $(V+h)^{**} = V+h$ and $\mathcal{W}\mathcal{W}^+V = V$. When $\mathcal{W}$ is discretized, then we get an approximation of the result above, with an approximation error that vanishes when $\mathcal{W}$ covers the domain of $(V+h)^*$. Similarly, for $\mathcal{Z}$, we obtain a similar behavior but for concave functions, because $\mathcal{Z}^{\top +}\mathcal{Z}^\top V = -\mathcal{Z}\mathcal{Z}^+(-V)$.

Smooth functions. If we make the assumption that the function V is smooth with respect to the Bregman divergence defined from the convex function $\frac{1}{2}h$ [3], that is, for all s, s', $-\frac{1}{2}\mathcal{D}_h(s', s) \leqslant V(s') - V(s) - \nabla V(s)^\top (s' - s) \leqslant \frac{1}{2}\mathcal{D}_h(s', s)$, that is, $V + \frac{1}{2}h$ and $\frac{1}{2}h - V$ are convex; this implies that with $g = \frac{1}{2}h$, $(V+h)^{**} - g = V + h - \frac{1}{2}h = V + \frac{1}{2}h$ is convex, and thus statement (c) from Prop. 4 leads to approximation guarantee for $\|V - \mathcal{W}\mathcal{W}^+V\|_\infty$. Similarly, the fact that $\frac{1}{2}h - V$ is convex leads to a similar guarantee for $\|\mathcal{Z}^{\top +}\mathcal{Z}^\top V - V\|_\infty$.

Morever, with $g = \frac{1}{2}h$, if h is strongly-convex, then g^* is smooth and we get a squared norm $\Omega(w - w')^2$ in the guarantees in Prop. 4.

In order to obtain guarantees from a finite number of basis functions, we just need a cover of the state-space for the Bregman divergence.

B.4 Proof of Prop. 4

The proof follows the same structure as [1], but extended to all convex functions h (and not only quadratic). We have, by definition of the Fenchel-conjugate $(V+h)^*$ of $V+h$ (see [6]):

$$\begin{aligned} \mathcal{W}\mathcal{W}^+V(s) &= \max_{w \in \mathcal{W}} w^\top s - h(s) + \min_{s' \in \mathcal{S}} V(s') + h(s') - w^\top s' \\ &= \max_{w \in \mathcal{W}} w^\top s - h(s) - (V+h)^*(w). \end{aligned} \tag{9}$$

Thus, if $\mathcal{W}$ contains the domain of $(V+h)^*$, we have $\mathcal{W}\mathcal{W}^+V(s) = (V+h)^{**}(s) - h(s)$, which shows (a). Moreover, this implies that in all situations, we have $\mathcal{W}\mathcal{W}^+V(s) \leqslant (V+h)^{**}(s) - h(s)$.

We now denote $w^*(s)$ the unconstrained minimizer in the optimization problem in Eq. (9) but do not assume that $\mathcal{W}$ contains the domain of $(V+h)^*$. Moreover, since $\mathcal{S}$ is compact, the function $(V+h)^*$ has subgradients in $\mathcal{S}$, thus the function $w \mapsto w^\top s - (V+h)^*(w)$ has gradients bounded in the norm Ω by $\mathrm{diam}(\mathcal{S})$. We thus get:

$$\begin{aligned} (V+h)^{**}(s) - h(s) - \mathcal{W}\mathcal{W}^+V(s) &= \min_{w \in \mathcal{W}} -w^\top s + (V+h)^*(w) - w^*(s)^\top s - (V+h)^*(w^*(s)) \\ &\leqslant \mathrm{diam}(\mathcal{S}) \min_{w \in \mathcal{W}} \Omega^*(w - w^*(s)). \end{aligned}$$

This leads to (b).

Since $w^*(s)$ is the unconstrained maximizer of $w^\top s - h(s) - (V+h)^*(w)$, it is such that $s \in \partial(V+h)^*(w^*(s))$. Since $(V+h)^{**} - g$ is convex, $g^* - (V+h)^*$ is convex, and then, for any w,

$$g^*(w) - (V+h)^*(w) \;\geqslant\; g^*(w^*(s)) - (V+h)^*(w^*(s)) + (w - w^*(s))^\top \big[\nabla g^*(w^*(s)) - s\big],$$

leading to, by rearranging terms:

$$-(V+h)^*(w) \geqslant -\mathcal{D}_{g^*}(w, w^*(s)) - (V+h)^*(w^*(s)) - (w - w^*(s))^\top s.$$

This leads to, for any $w \in \mathcal{W}$,

$$(V+h)^*(w) - w^\top s - (V+h)^*(w^*(s)) + w^*(s)^\top s \leqslant \mathcal{D}_{g^*}(w, w^*(s)).$$

Taking the infimum with respect to $w \in \mathcal{W}$, we get:

$$-\mathcal{W}\mathcal{W}^+ V(s) - h(s) + (V+h)^{**}(s) \leqslant \min_{w \in \mathcal{W}} \mathcal{D}_{g^*}(w, w^*(s)),$$

which in turn leads to (c), since the two quantities above are non-negative.

C Detailed proofs of complexities of iterations

Compile time for sparse 1D graph. Given an order k chain graph, the square of the adjacency matrix is of order $2k$ and getting it can be done in $O(k^2|\mathcal{S}|)$ (with $2k|\mathcal{S}|$ elements to obtain, each with $O(k)$ complexity). Thus, the overall complexity is $\sum_{i=0}^{\log_2 \rho} (2^i)^2 |\mathcal{S}| = O(|\mathcal{S}|\rho^2)$.

Compile time for sparse graph (general dimension). Given an order k d-dimensional grid (where each node is connected to $2d$ neighbors, two per dimension), up to multiplicative constants depending on d, the square of the adjacency matrix is of order[2] k^d and getting it can be done in $O(k^{2d}|\mathcal{S}|)$ (with $k^d|\mathcal{S}|$ elements to obtain, each with $O(k^d)$ complexity). Thus, the overall complexity is $\sum_{i=0}^{\log_2 \rho} (2^i)^{2d} |\mathcal{S}| = O(|\mathcal{S}|\rho^{2d})$.

D Convex optimization for estimating z_{new} and w_{new}

In order to go beyond a finite set of functions, one can optimze z_{new} as follows (the optimization problem for w_{new} would follow similarly).

[2]This degree $\deg(\rho, d)$ is equal to the number of elements of the elements of the ℓ_1-ball of radius ρ with integer coordinates. Following [29], we have $\deg(\rho, d) = \sum_{i=0}^{\min\{d,\rho\}} 2^i \binom{d}{i}\binom{\rho}{i}$. In particular $\deg(\rho, d) = \deg(d, \rho)$, and we have the bound $\deg(\rho, d) \leqslant \rho^d \frac{2^d}{d!}$, which is the volume of the ℓ_1-ball of radius ρ in dimension d, which grows as ρ^d.

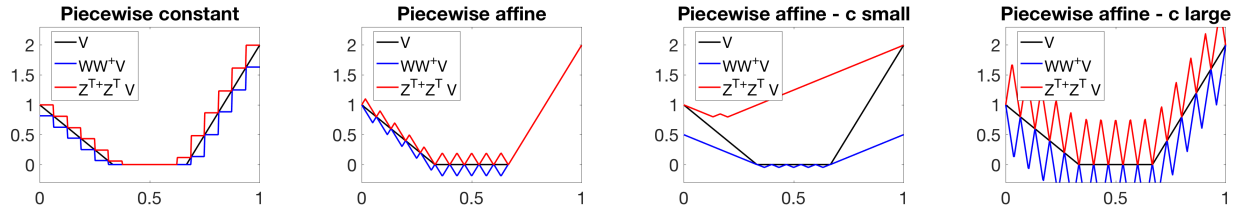

Figure 4: Approximation of a function V with different finite basis with 16 elements. One-dimensional case (with a convex optimal value function). From left to right: piece-wise constant basis function, piecewise affine basis functions with well chosen value of c, then too small and too large.

We can write the criterion that needs to be optimized from Section 5 as follows:

$$\begin{aligned}
&\max_{s\in\mathcal{S}} \min\left\{U(s) - TV(s), -TV(s) - z_{\text{new}}(s) + \langle TV | z_{\text{new}}\rangle, \right\} \\
=&\max_{s\in\mathcal{S}} \min_{\eta(s,z_{\text{new}})\in[0,1]} \eta(s,z_{\text{new}})\big[U(s) - TV(s)\big] + (1-\eta(s,z_{\text{new}}))\big[-TV(s) - z_{\text{new}}(s) + \langle TV|z_{\text{new}}\rangle\big] \\
\leqslant&\max_{s\in\mathcal{S}} \eta^*(s,\tilde{z}_{\text{new}})\big[U(s) - TV(s)\big] + (1-\eta^*(s,\tilde{z}_{\text{new}}))\big[-TV(s) - z_{\text{new}}(s) + \langle TV|z_{\text{new}}\rangle\big],
\end{aligned}$$

where $\eta^*(s,\tilde{z}_{\text{new}})$ is the minimizer for a fixed $\tilde{z}_{\text{new}}$. The previous function is *convex* in z_{new}. This leads to a natural majorization-minimization algorithm, which needs to be initialized in a problem dependent way, and can thus be used to learn linear parametrizations of z_{new}.

E Extra experiments

In this section, we provide extra experiments and complements on one-dimensional and two-dimensional problems.

E.1 One-dimensional

We consider pairs (b, V) of reward and optimal value functions that satisfy the Hamilton-Jacobi-Bellman (HJB) equation [11, 25]: $V(x)\log\eta + |V'(x)| + b(x) = 0$. We consider two functions $V(x)$, from which we can recover the function $b(x)$:

- $V(x) = (1-3x)_+ + (6x-4)_+ + (1-36(x-1/2)^2)_+$, as plotted in Figure 1.
- $V(x) = (1-3x)_+ + (6x-4)_+$, as plotted in Figure 4.

We consider $\eta = 1/2$, a number of nodes equal to $S = 362 \approx 2^{17/2}$, such that $\gamma \approx 0.9981$ and $\tau \approx 521$.

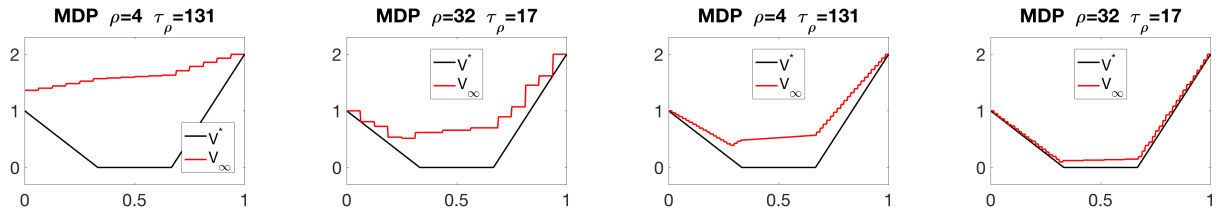

Figure 5: Approximation of a function V with different finite basis with 16 or 64 elements, *within the MDP*, and with values of ρ that are 4 and 32. One-dimensional case (with a convex optimal value function). From left to right: $(n = 16, \rho = 4)$, $(n = 16, \rho = 32)$, $(n = 64, \rho = 4)$, $(n = 16, \rho = 32)$.

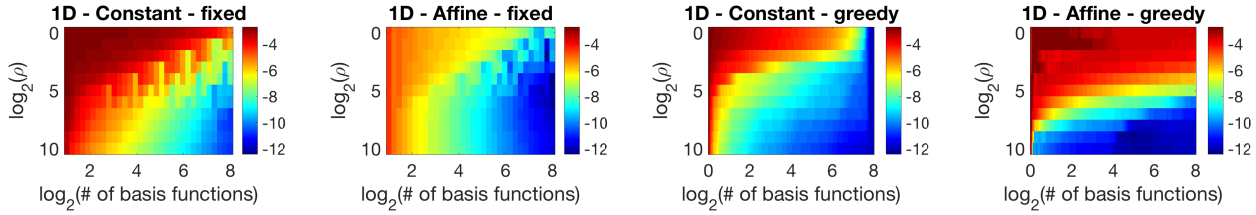

Figure 6: Approximation error of a function V as a function of ρ and the number of basis functions, for piecewise constant functions and piecewise affine functions. One-dimensional case (with a convex optimal value function). Left: fixed, right: greedy.

E.2 Two-dimensional

We consider control problems on $[0, 1]^d$, with $d = 2$, with $2d$ potential discrete actions (one in each coordinate direction). The corresponding Hamilton-Jacobi equation is

$$V(x) \log \eta + \max_{i \in \{1,\dots,d\}} \left| \frac{\partial V}{\partial x_i}(x) \right| + b(x) = 0,$$

with natural boundary conditions. We consider the following functions:

- $V(x_1, x_2) = (1 - 3x_1)_+ + (6x_1 - 4)_+$. In the main paper, we consider this value function (with potential for variable selection) plotted in Figure 7, with performance plots in the bottom of Figure 3.
- $V(x_1, x_2) = (1 - 3x_1)_+ + (6x_1 - 4)_+ + (1 - 3x_2)_+ + (6x_2 - 4)_+$. This function, without potential for variable selection, is plotted in Figure 8, with performance plots in Figure 9.

We consider $\eta \approx 0.919$, a number of nodes per dimension equal to $\tilde{S} = 45 \approx 2^{11/2}$ and thus a total number of nodes equal to $S = \tilde{S}^2 = 2025$, such that $\gamma \approx 0.9981$ and $\tau \approx 521$ (like in the one-dimensional case).

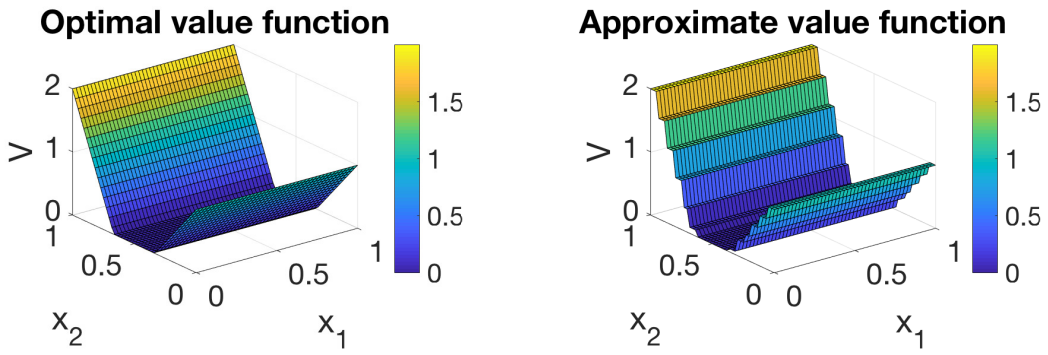

Figure 7: Value function with potential for variable selection. Left: optimal, right: approximation with piecewise constant functions.

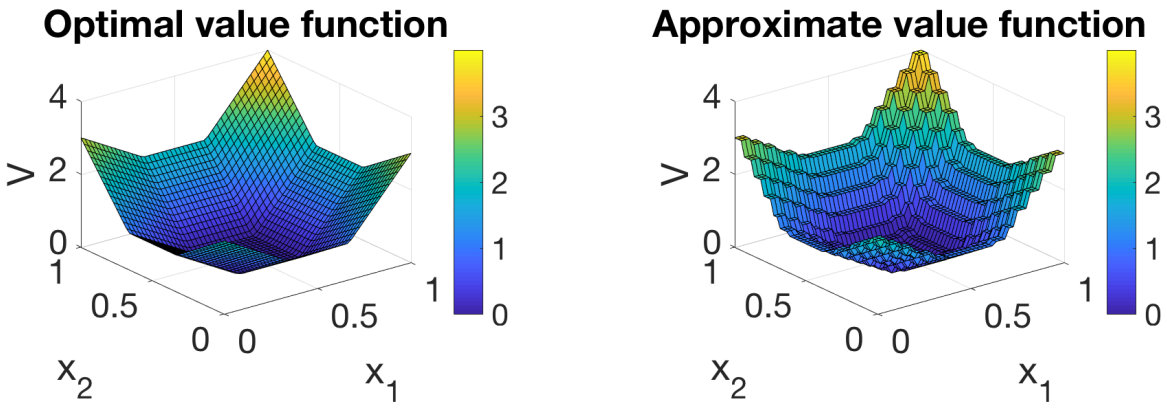

Figure 8: Value function without potential for variable selection. Left: optimal, right: approximation with piecewise constant functions.

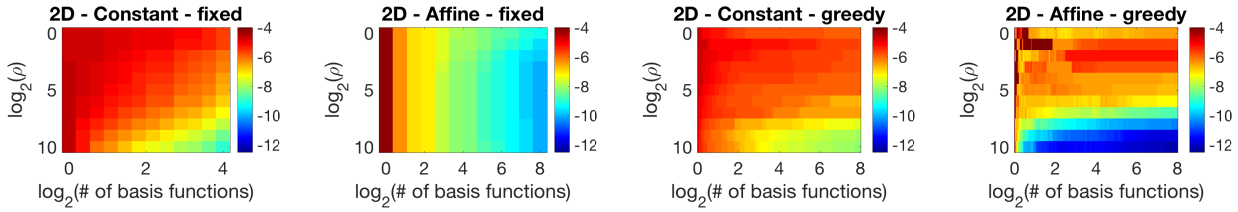

Figure 9: Approximation error of a function V as a function of ρ and the number of basis functions, for piecewise constant functions and piecewise affine functions. Two-dimensional case corresponding to the function in Figure 8. Left: fixed, right: greedy. Compared to the bottom of Figure 3 in the main paper, the gains in performance of the greedy method are not as large because the optimal reward function depends on all variables.