Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes

Eloïse Berthier, Francis Bach

MAX-PLUS LINEAR APPROXIMATIONS FOR DETERMINISTIC CONTINUOUS-STATE MARKOV DECISION PROCESSES

ELOÏSE BERTHIER AND FRANCIS BACH

Inria - Ecole Normale Supérieure
PSL Research University, Paris, France

Abstract. We consider deterministic continuous-state Markov decision processes (MDPs). We apply a max-plus linear method to approximate the value function with a specific dictionary of functions that leads to an adequate state-discretization of the MDP. This is more efficient than a direct discretization of the state space, typically intractable in high dimension. We propose a simple strategy to adapt the discretization to a problem instance, thus mitigating the curse of dimensionality. We provide numerical examples showing that the method works well on simple MDPs.

1. Introduction

Reinforcement learning problems [SB18] are generally formulated as Markov decision processes (MDPs). Dynamic programming provides simple algorithms, such as value iteration, to compute the optimal value function and an optimal policy for a discrete MDP, when the model is known.

Yet many problems formalized as MDPs are time- and space-discretizations of control problems, with a continuous underlying state space. To faithfully reproduce the dynamics of the control problem, one needs to compute a sharp space-discretization, subject to the curse of dimensionality: for high-dimensional problems, the space-discretized MDP will not even fit in memory.

Following the method of [McE03] and [AGL08], we compute approximations of the optimal value function for deterministic MDPs, namely max-plus linear approximations within a dictionary of functions. These methods have been developed for optimal control and deal with continuous state spaces. For certain choices of function dictionaries, they can be viewed as an efficient way to discretize the state-continuous MDP while preserving its dynamics. Adaptively choosing the basis functions used to approximate the value function is a way to circumvent the curse of dimensionality when the true value function has a sparse representation.

Our contributions are the following:

- we propose a new approximation method to solve subproblems appearing in the max-plus value iteration algorithm, namely to optimize some objectives over the state-space with gradient ascent (section III, D);
- we present a specific dictionary of functions simplifying the method, and show how it can be used to build an adaptive discretization of the state space (section VI);
- we provide numerical simulations on MDPs where this adaptive max-plus approximation method computes nearly optimal policies with significantly less parameters than discretized value iteration (section VII).

Setting: We consider a deterministic, time-homogeneous, infinite-horizon, discounted MDP [HLL12] defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, a bounded reward function $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$ for some $R \geq 0$, a dynamics $\varphi_{\cdot}(.) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and a discount factor $0 \leq \gamma < 1$, with the following assumptions:

(1) the state space $\mathcal{S}$ is a bounded subset of $\mathbb{R}^d$ ($d \geq 1$);
(2) the action space $\mathcal{A}$ is finite.

E-mail address: eloise.berthier@inria.fr, francis.bach@inria.fr.

We want to approximate the optimal value function $V^* : \mathcal{S} \to \mathbb{R}$ corresponding to an optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$ maximizing the cumulative discounted reward. The greedy policy π corresponding to a value function V is obtained by:

$$\pi(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} r(s,a) + \gamma V(\varphi_a(s)).$$

The value iteration algorithm consists in computing V^* as the unique fixed point of the Bellman operator $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ ($\mathbb{R}^{\mathcal{S}}$ denotes the set of functions from $\mathcal{S}$ to $\mathbb{R}$) defined as:

$$TV(s) := \max_{a \in \mathcal{A}} r(s,a) + \gamma V(\varphi_a(s)).$$

The value iteration algorithm iteratively computes the recursion $V_{k+1} = TV_k$ that converges to V^*, with linear rate since T is strictly contractive with factor $\gamma < 1$. But if $\mathcal{S}$ is a finite set, it requires $O(|\mathcal{A}| \cdot |\mathcal{S}|)$ computations, and the storage of $O(|\mathcal{S}|)$ values of V_k at each step.

From now on we consider that $\mathcal{S}$ is a compact but potentially not discrete set. In this case one can directly look for a discretization of the MDP and perform value iteration, but this will become intractable in high dimension since the size of the discretized state space grows exponentially with the dimension. Alternatively one can consider the space-continuous MDP and compute an approximation of the optimal value function, without having to discretize the MDP.

2. Max-Plus Linear Approximations

Let $\mathcal{W}$ be a finite dictionary of functions $w : \mathcal{S} \to \mathbb{R}$. The value function can be approximated by a "linear" combination of functions in $\mathcal{W}$, with an adapted definition of linearity. The max-plus semiring [GP97] is defined as $(\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$, where $\oplus$ represents the maximum operator, and $\otimes$ represents the usual sum. Like linear combinations in the usual ring, for $\alpha \in \mathbb{R}^{\mathcal{W}}$, we define the max-plus linear combination:

$$V(s) = \bigoplus_{w \in \mathcal{W}} \alpha(w) \otimes w(s) = \max_{w \in \mathcal{W}} \alpha(w) + w(s).$$

The Bellman operator's structure is naturally compatible with max-plus operations, as it is max-plus additive and homogeneous: for $c, V, V' \in \mathbb{R}^{\mathcal{S}}$, we have

$$\begin{aligned} T(V \oplus V') &= T(\max\{V, V'\}) = \max\{TV, TV'\} = TV \oplus TV' \\ T(c \otimes V) &= T(c + V) = \gamma c + TV = c^{\otimes \gamma} TV. \end{aligned}$$

The basis functions used in [AGL08] and [McE03] are smooth ($w_i(s) := -c\|s - s_i\|^2$ for some $s_i \in \mathcal{S}$) or Lipschitz-continuous ($w_i(s) := -c\|s - s_i\|$). However, the scale $c > 0$ of such functions must be chosen according to the regularity of the true value function. Since it is unknown in practice it needs to be tuned as a hyperparameter. Other somewhat simpler choices of basis functions can be considered as well. Let $(A(w_1), ..., A(w_n))$ be a partition of the state space, where each w_i is defined as the max-plus indicator of a set $A(w_i)$:

$$w_i(s) := \begin{cases} 0 & \text{if } s \in A(w_i) \\ -\infty & \text{otherwise.} \end{cases}$$

Then the max-plus linear combinations of $(w_1, ..., w_n)$ span the set of value functions that are piecewise constant with respect to the partition [Bac19]. This is thus a way to discretize the value function.

Following the notations of [Bac19], for a given dictionary of functions $\mathcal{W}$, we define the following four operators:

$$\begin{aligned} W : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{S}}, &\quad W\alpha(s) := \max_{w \in \mathcal{W}} \alpha(w) + w(s) \\ W^+ : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{W}}, &\quad W^+V(w) := \inf_{s \in \mathcal{S}} V(s) - w(s) \\ W^\top : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{W}}, &\quad W^\top V(w) := \sup_{s \in \mathcal{S}} V(s) + w(s) \\ W^{\top +} : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{S}}, &\quad W^{\top +}\alpha(s) := \min_{w \in \mathcal{W}} \alpha(w) - w(s). \end{aligned}$$

W maps a vector α to a function $W\alpha$ that is the max-plus linear combination of the dictionary $\mathcal{W}$ with coefficient α. W^+ is known as the residuation [CGQ04] of W and acts as a pseudo-inverse: $W\alpha \leq V \Leftrightarrow \alpha \leq W^+V$. The transposed notation for $W^\top$ comes from the definition of a max-plus *dot product* (it is only a max-plus bilinear form) between functions on $\mathcal{S}$, which will be used in the rest of the paper:

$$\forall z, w \in \mathbb{R}^{\mathcal{S}}, \quad \langle z, w \rangle := \sup_{s \in \mathcal{S}} z(s) + w(s).$$

The lower-projection of a function $V \in \mathbb{R}^{\mathcal{S}}$ onto the span of the dictionary is computed as WW^+V, and $W^{\top+}W^\top V$ is its upper-projection. Both projection operators WW^+ and $W^{\top+}W^\top$ are idempotent and non-expansive for the ℓ_∞ norm.

3. Approximate Value Iteration

These max-plus tools can be used to compute a tractable approximation of the optimal value function of an MDP.

3.1. Projection Method. A simple way to approximate the value function has been proposed in [AGL08], as an extension of the method of [McE03], both for control problems. Following [Bac19], we apply it to MDPs. The idea is to represent the value function as a max-plus linear combination in a dictionary of functions, and to apply alternatingly the Bellman operator and a projection onto the span of the dictionary: $V_{k+1} = WW^+TV_k$. Hence if V_k is represented as $W\alpha_k$, then α_{k+1} is given by $\alpha_{k+1} = W^+TW\alpha_k$, where the operator $W^+TW : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{W}}$ is computed by:

$$\alpha_{k+1}(w) = \inf_{s \in \mathcal{S}} \max_{w' \in \mathcal{W}} \gamma\alpha_k(w') + Tw'(s) - w(s).$$

This computation is a min/max problem, which is not easy to solve in general. If $\mathcal{S}$ is finite, this requires to compute $|\mathcal{S}| \cdot |\mathcal{W}|$ values at each iteration.

3.2. Variational Method. A slightly more involved approximation method has been also proposed by [AGL08]. Let us define two dictionaries of functions $\mathcal{W}$ and $\mathcal{Z}$. $\mathcal{W}$ plays the same role as before, while $\mathcal{Z}$ is a set of test functions which can be taken equal to $\mathcal{W}$. The value iteration recursion $V_{k+1} = TV_k$ is replaced by a variational formulation:

$$\langle z, V_{k+1} \rangle = \langle z, TV_k \rangle \quad \forall z \in \mathcal{Z},$$

of which we consider the maximal solution in span(W), given by ([AGL08], Proposition 4): $V_{k+1} = WW^+Z^{\top+}Z^\top TV_k$. It can be interpreted as a first projection on the min-plus span generated by $\mathcal{Z}$, before a second projection on the max-plus span of $\mathcal{W}$. Again if V_k is represented by $W\alpha_k$, we have the following recursion:

$$\alpha_{k+1} = W^+Z^{\top+}Z^\top TW\alpha_k.$$

The operator $W^+Z^{\top+}Z^\top TW : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{W}}$ decomposes as $M \circ K$, with $K = Z^\top TW : \mathbb{R}^{\mathcal{W}} \to \mathbb{R}^{\mathcal{Z}}$ and $M = W^+Z^{\top+} : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{W}}$. The recursion may be recast as:

$$\begin{aligned}
\beta_{k+1}(z) = K\alpha_k(z) &= \sup_{s \in \mathcal{S}} z(s) + \max_{w \in \mathcal{W}} \gamma\alpha_k(w) + Tw(s) \\
&= \max_{w \in \mathcal{W}} \gamma\alpha_k(w) + \langle z, Tw \rangle \\
\alpha_{k+1}(w) = M\beta_{k+1}(w) &= \inf_{s \in \mathcal{S}} -w(s) + \min_{z \in \mathcal{Z}} \beta_{k+1}(z) - z(s) \\
&= \min_{z \in \mathcal{Z}} \beta_{k+1}(z) - \langle z, w \rangle.
\end{aligned}$$

The operator $W^+Z^{\top+}Z^\top TW$ is a γ-contraction, hence the recursion will converge with linear rate to the unique fixed point. An interesting property is that the $|\mathcal{Z}| \cdot |\mathcal{W}|$ values $\langle z, Tw \rangle$ for $(z, w) \in \mathcal{Z} \times \mathcal{W}$ can be precomputed at a cost that is independent of the horizon $1/(1-\gamma)$ of the MDP. The main difficulty here is their prior computation. Unlike in [Bac19] where $\mathcal{S}$ is finite, for a continuous state space these computations might only be performed approximately.

3.3. Basis Functions and Clustered MDP. Discrete versions of the MDP can be built from the preceding approximation methods with $\mathcal{W}$ a set of max-plus indicators corresponding to a partition of the state space, as mentioned earlier. Indeed, when $\mathcal{W} = \mathcal{Z}$ and the $(w_i)_{1\leq i\leq n}$ are max-plus indicators, the above operator M is the identity and $W^+W^{\top+}W^\top TW = W^\top TW$ (a max/max problem), to be compared with W^+TW (a min/max problem) for the projection method. Note that with the approximate indicators introduced below, M will not be equal to the identity, even though $\mathcal{W} = \mathcal{Z}$.

With max-plus indicators and the variational method, the approximate value iteration becomes:

$$\alpha_{k+1}(w) = \max_{w'\in\mathcal{W}} \langle w, Tw'\rangle + \gamma\alpha_k(w'),$$

which we interpret as classical value iteration on the MDP formed with the clusters $(A(w))_{w\in\mathcal{W}}$ as states, and as rewards the maximal achievable reward going from one cluster to the other, that is:

$$\begin{aligned} R(w,w') = \langle w, Tw'\rangle &= \sup_{s\in\mathcal{S}} w(s) + Tw'(s) \\ &= \sup_{s\in A(w)} \max_{\substack{a\in\mathcal{A} \text{ s.t.} \\ \varphi_a(s)\in A(w')}} r(s,a), \end{aligned}$$

with $R(w,w') = -\infty$ if $\{(s,\varphi_a(s)) \mid s\in\mathcal{S}, a\in\mathcal{A}\} \cap A(w)\times A(w') = \varnothing$. This reduced problem is appealing but hard to solve in a continuous state space. Even finding if $R(w,w')$ is finite is both a controllability and reachability problem [Lib11], whose solution is not straightforward. A differentiable version of the max-plus indicators is the following:

$$w(s) = -c\text{dist}(s, A(w))^2,$$

where $\text{dist}(s, A(w))$ is the euclidean distance between s and the set $A(w)$, and $c > 0$ is a hyper-parameter, typically chosen large compared to the scale of the true value function. We refer to such basis functions as soft indicators. When $c \to +\infty$, we recover the preceding clustered MDP and elements in the span of W are *almost* (asymptotically) piecewise constant with respect to the partition.

3.4. Oracle Subproblem. We now take a closer look at the subproblems that must be solved before running the approximate value iteration recursion, namely $\langle z, w\rangle$ and $\langle z, Tw\rangle$ for the variational method. First, $\langle z, w\rangle$ is independent of the MDP and can be computed in closed form for general choices of dictionaries, and:

$$\begin{aligned} \langle z, Tw\rangle &= \sup_{s\in\mathcal{S}} z(s) + Tw(s) \\ &= \sup_{s\in\mathcal{S},\ a\in\mathcal{A}} z(s) + r(s,a) + \gamma w(\varphi_a(s)). \end{aligned}$$

This is a discrete-time control problem, easier than the original one (finding the optimal value function) since its horizon is one time step. As mentioned by [AGL08], this is a perturbed version of the computation of $\langle z, w\rangle$ as soon as T is close to the identity, that is, in the context of optimal control when the time-discretization of the MDP is small.

In [AGL08], $\langle z, Tw\rangle$ is approximated using the Hamiltonian of a control problem. For general MDPs, we may look at this problem from a different perspective. It is an optimization problem, and, as noted by [AGL08], even though computing $\langle z, Tw\rangle$ is not a concave maximization problem, choosing strongly concave basis functions z and w has a regularizing effect.

Hence an approximation of $\langle z, Tw\rangle$ can be computed by gradient ascent on $f_a(s) := z(s) + r(s,a) + \gamma w(\varphi_a(s))$, for each $a\in\mathcal{A}$, and then taking the maximum on a. For differentiable z, w, φ_a and $r(.,a)$, f_a is differentiable with:

$$\nabla f_a(s) = \nabla z(s) + \nabla r(s,a) + \gamma J_{\varphi_a}(s)^\top \nabla w(\varphi_a(s)),$$

where J_{φ_a} denotes the Jacobian of φ_a and ∇r is the gradient of r with respect to s. Seeing this problem like [AGL08] as a perturbation of $\langle z, w\rangle$, an efficient initialization for gradient ascent on this problem is given by $s_0 \in \text{argmax}_s\ z(s) + w(s)$. Furthermore, for continuous basis functions, reward function and dynamics, since $\mathcal{S}$ is compact by assumption, the supremum in $\langle z, Tw\rangle$ is a maximum attained at some $(s,a)\in\mathcal{S}\times\mathcal{A}$. The full procedure to obtain the approximate value function is described in Algorithm 1.

As noted in [Bac19], the Bellman operator T can be replaced by T^ρ for some integer $\rho \geq 1$, replacing accordingly γ by γ^ρ. This makes the computation of $\langle z, T^\rho w\rangle$ more complicated, as it requires to run $|\mathcal{A}|^\rho$ gradient ascents. A simplification is to consider only sequences of constant actions for ρ steps.

Comparison with existing methods: Approximate value iteration is usually performed by fitted value iteration [SB18], with a linear parameterization of the value function. With max-plus parameterizations, the projections are computed efficiently, which spares the repeated use of stochastic optimization, and leads to an explicit error analysis. Nonlinear approximations can be handled with Q-learning (see [MM09] for continuous MDPs), with weaker convergence guarantees [SB18].

Algorithm 1 Max-plus Approximate Value Iteration

Input: MDP, $\mathcal{W}$ and $\mathcal{Z}$, gradient steps k, step size ξ
Output: approximate value function V

Precomputations:
for $z \in \mathcal{Z}, w \in \mathcal{W}$ **do**
  $s, \langle z, w\rangle \leftarrow \mathrm{argmax}, \max_{s\in\mathcal{S}}\ z(s) + w(s)$
  for $a \in \mathcal{A}$ **do**
    $\langle z, Tw\rangle \leftarrow z(s) + r(s,a) + w(\varphi_a(s))$
    for $i = 1$ **to** k **do**
      $g \leftarrow \nabla z(s) + \nabla r(s,a) + J_{\varphi_a}(s)^\top \nabla w(\varphi_a(s))$
      $s \leftarrow s + \xi g$
      $f \leftarrow z(s) + r(s,a) + w(\varphi_a(s))$
      $\langle z, Tw\rangle \leftarrow \max\{f, \langle z, Tw\rangle\}$
Reduced value iteration:
$\alpha \leftarrow 0$
repeat
  for $z \in \mathcal{Z}$ **do**
    $\beta(z) \leftarrow \max_{w\in\mathcal{W}}\ \gamma\alpha(w) + \langle z, Tw\rangle$
  for $w \in \mathcal{W}$ **do**
    $\alpha(w) \leftarrow \min_{z\in\mathcal{Z}}\ \beta(z) - \langle z, w\rangle$
until convergence
return $V = W\alpha$

4. Error Analysis

4.1. Error Decomposition. The operator $\bar{T} := WW^+Z^{\top+}Z^\top T$ is γ-contractive, since T is γ-contractive and both WW^+ and $Z^{\top+}Z^\top$ are non-expansive. If the $\langle z, w\rangle$ and $\langle z, Tw\rangle$ are computed exactly, the error of the *exact* max-plus approximation is controlled only by projection errors. In practice, the values $K_{z,w} := \langle z, Tw\rangle$ are approximated by some $\hat{K}_{z,w}$ obtained by gradient ascent with some error due to the finite number of iterations and to the non-concavity of the objective function.

Proposition 1. *Let V^* be the optimal value function of the MDP, $\hat{V} = W\hat{\alpha}$, where $\hat{\alpha}$ is the fixed point of $M \circ \hat{K}$, and*

$$\|\hat{K} - K\|_\infty := \sup_{z\in\mathcal{Z}, w\in\mathcal{W}} |\hat{K}_{z,w} - K_{z,w}|.$$

$$\text{Then:}\quad \|\hat{V} - V^*\|_\infty \leq \frac{1}{1-\gamma}\Big(\|WW^+V^* - V^*\|_\infty + \|Z^{\top+}Z^\top V^* - V^*\|_\infty + \|\hat{K} - K\|_\infty\Big).$$

Proof. Let V_∞ be the unique fixed point of $\bar{T}$:

$$\begin{aligned}\|V_\infty - V^*\| &\leq \|\bar{T}V_\infty - \bar{T}V^*\| + \|\bar{T}V^* - V^*\| \\ &\leq \gamma\|V_\infty - V^*\| + \|WW^+Z^{\top+}Z^\top V^* - V^*\|\end{aligned}$$

$$\begin{aligned}(1-\gamma)\|V_\infty - V^*\| &\leq \|WW^+Z^{\top+}Z^\top V^* - WW^+V^*\| \\ &\quad + \|WW^+V^* - V^*\| \\ &\leq \|Z^{\top+}Z^\top V^* - V^*\| + \|WW^+V^* - V^*\|,\end{aligned}$$

because WW^+ is non-expansive. Since $\hat{\alpha} = M \circ \hat{K}\hat{\alpha}$:

$$\begin{aligned}\|\hat{V} - V_\infty\| &\leq \|\hat{V} - \bar{T}\hat{V}\| + \gamma\|\hat{V} - V_\infty\| \\ (1-\gamma)\|\hat{V} - V_\infty\| &\leq \|WW^+Z^{\top+}\hat{K}\hat{\alpha} - WW^+Z^{\top+}K\hat{\alpha}\| \\ &\leq \|Z^{\top+}\hat{K}\hat{\alpha} - Z^{\top+}K\hat{\alpha}\| \leq \|\hat{K}\hat{\alpha} - K\hat{\alpha}\| \\ &\leq \|\hat{K} - K\|_\infty.\end{aligned}$$

The last two inequalities result from the reverse triangle inequality for the infinity norm. Combining the last line with the upper bound on $\|V_\infty - V^*\|$, the result follows. □

In numerical implementations, the fact that reduced value iteration is stopped after a finite number of iterations causes a last source of error. Since the convergence is fast, it will often be negligible compared to the other approximations.

Proposition 2. *For $\alpha_0 \in \mathbb{R}^{\mathcal{W}}$, let $\alpha_{k+1} = M\hat{K}\alpha_k$ for $k \geq 1$. Then denoting as $\hat{\alpha}$ the unique fixed point of $M \circ \hat{K}$:*

$$\|W\alpha_k - W\hat{\alpha}\|_\infty \leq \|\alpha_k - \hat{\alpha}\|_\infty \leq \gamma^k\|\alpha_0 - \hat{\alpha}\|_\infty.$$

4.2. Projection Error. For any $V \in \mathbb{R}^{\mathcal{S}}$, $W^{\top+}W^\top V = -WW^+(-V)$, so we only consider the projection error for WW^+ (lower projection).

Proposition 3. *Let $c > 0$ and $(A_1, ..., A_n)$ a partition of $\mathcal{S}$ where each A_i is convex, compact and non-empty, and let $D = \max_{1\leq i\leq n} \mathrm{diam}(A_i)$ where $\mathrm{diam}(A_i) = \max_{s,s'\in A_i} \|s - s'\|$. Let $\mathcal{W}_1 = \{w_1^1, ..., w_n^1\}$ and $\mathcal{W}_2 = \{w_1^2, ..., w_n^2\}$ defined by:*

$$\forall i \in \{1, ..., n\}, \forall s \in \mathcal{S}, \quad \begin{cases} w_i^1(s) & = -c_1 \mathrm{dist}(s, A_i) \\ w_i^2(s) & = -c_2 \mathrm{dist}(s, A_i)^2. \end{cases}$$

If V has Lipschitz constant L and $c_1 \geq L$, $c_2 \geq \frac{L}{4D}$, then

$$\|V - W_1W_1^+V\|_\infty \leq LD$$

$$\|V - W_2W_2^+V\|_\infty \leq LD + \frac{L^2}{4c_2} \leq 2LD.$$

Proof. For $s \in \mathcal{S}$,

$$WW^+V(s) \leq \max_w w(s) + V(s) - w(s) \leq V(s).$$

On the other hand, $\exists i \in \{1, ..., n\}$ s.t. $s \in A_i$. Then:

$$\begin{aligned}WW^+V(s) &= \max_w w(s) + \inf_{s'} V(s') - w(s') \\ &\geq \underbrace{w_i(s)}_{=0} + \inf_{s'} V(s') - w_i(s').\end{aligned}$$

The Lipschitz continuity of V implies that for $p \in \{1, 2\}$:

$$WW^+V(s) \geq V(s) + \inf_{s'\in\mathcal{S}} \left\{c_p \mathrm{dist}(s', A_i)^p - L\|s - s'\|\right\},$$

and the results follow using $\|s - s'\| \leq \mathrm{diam}(A_i) + \mathrm{dist}(s', A_i)$.

□

Unlike for smooth or Lipschitz-continuous basis functions [AGL08, Bac19], there is no dependency in c in the bound, for c large enough. This avoids oscillations of the approximation when c is chosen too large and simplifies parameter selection.

5. Comparison with the Method of Akian, Gaubert & Lakhoua for Control Problems

Deterministic MDPs and optimal control problems are closely related. Applying our method to an MDP that is a time-discretization of a control problem is similar to directly applying the original method by [AGL08] to the control problem. Let $0 \leq \eta < 1$ be a discount factor, $\bar{r} : \mathcal{S} \times \mathcal{A} \to [-R, R]$, $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and define the optimal control problem [FS06]:

$$\sup_{a(.)} \int_0^{+\infty} \eta^t \bar{r}(s(t), a(t)) \mathrm{d}t,$$

with $s(0) = s_0, \forall t \geq 0, \dot{s}(t) = f(s(t), a(t))$, $(s(t), a(t)) \in \mathcal{S} \times \mathcal{A}$.

5.1. Time-Discretization of a Control Problem. A control problem can be approximated by a state-continuous MDP by time-discretization. The corresponding time-discretized MDP with step $\tau > 0$ and Euler scheme is:

$$r(s, a) = \tau \bar{r}(s, a), \quad \varphi_a(s) = s + \tau f(s, a), \quad \gamma = \eta^\tau.$$

For $\tau > 0$, the continuous- and discrete-time Bellman operators $S_\tau, T_\tau : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ are defined by, for each $u \in \mathbb{R}^{\mathcal{S}}$:

$$\begin{aligned}(S_\tau u)(s) &:= \sup_{a(.)} \int_0^\tau \eta^t \bar{r}(s(t), a(t)) \mathrm{d}t + \eta^\tau u(s(\tau)) \\ (T_\tau u)(s) &:= \max_a \tau \bar{r}(s, a) + \eta^\tau u(s + \tau f(s, a)).\end{aligned}$$

Under regularity assumptions, the value function of the MDP converges to the value function of the control problem [TBO19]. This is obtained in a similar way as the Hamilton-Jacobi-Bellman (HJB) equation [FS06].

5.2. Hamiltonian Approximation for the Oracle Subproblem. For a continuous-time control problem, in the approximation method, T is the continuous Bellman operator S_τ. The HJB equation provides a first order approximation [AGL08] of $S_\tau w$ with respect to the horizon τ:

$$\begin{aligned}S_\tau w(s) &= \sup_{a(.)} \int_0^\tau \eta^t \bar{r}(s(t), a(t)) \mathrm{d}t + \eta^\tau w(s(t + \tau)) \\ &= \sup_{a \in \mathcal{A}} \tau \bar{r}(s, a) + o(\tau) + (1 + \tau \log \eta + o(\tau)) \\ &\qquad \times (w(s) + \tau \nabla w(s) \cdot f(s, a) + o(\tau)) \\ &= (1 + \tau \log \eta) w(s) + \tau H(s, \nabla w) + o(\tau),\end{aligned}$$

where $H(s, p) := \sup_{a \in \mathcal{A}} \bar{r}(s, a) + p \cdot f(s, a)$ is the Hamiltonian of the control problem. Instead of optimizing over $\mathcal{S}$, a second approximation made by [AGL08] is to consider only $s_0 \in S(z, w) := \operatorname{argmax}_s z(s) + w(s)$, since $\langle z, Tw \rangle$ is a perturbation of $\langle z, w \rangle$ for τ small. The final approximation is:

$$\langle z, Tw \rangle \simeq \sup_{s \in S(z,w)} z(s) + (1 + \tau \log \eta) w(s) + \tau H(s, \nabla w).$$

This is a valid approximation up to $O(\tau\sqrt{\tau})$ terms, if the Hamiltonian is Lipschitz-continuous and $z + w$ is strongly concave. This prevents the use of Lipschitz bases for $\mathcal{Z}$ and $\mathcal{W}$ at the same time in [AGL08]. Without these assumptions, the approximation is weaker, in $O(\tau)$, breaking the convergence of the method. In this case, one cannot avoid optimizing on s. [McE03] and [AGL08] use the first order approximation, but τ is a parameter of their method that can be made arbitrarily small. In the context of MDPs, τ is fixed and in principle it cannot be modified while solving the MDP. Besides, some MDPs are not natural time-discretizations of control problems.

For control problems, time-discretization and Hamiltonian approximation result in the same approximation of $\langle z, Tw \rangle$, up to $o(\tau)$ terms (or $O(\tau^2)$ assuming more regularity on w):

$$\begin{aligned}\hat{K}_{z,w} &= \sup_{s,a} z(s) + \tau \bar{r}(s, a) + \eta^\tau w(s + \tau f(s, a)) \\ &= \sup_s z(s) + (1 + \tau \log \eta) w(s) + \tau H(s, \nabla w) + o(\tau).\end{aligned}$$

If τ is not negligible, the Hamiltonian approximation is no longer valid, nor the approximate computation of $\langle z, S_\tau w\rangle$. On the other hand, the computation of $\langle z, T_\tau w\rangle$ is still valid, but the MDP no longer approximates the control problem.

After convergence of the reduced value iteration, our method provides an approximation of the value function of the control problem with an error of order

$$O\left(D/\tau + \tau + \|\hat{K} - K\|_\infty/\tau\right),$$

where D is the maximal diameter of the partition, which is similar to [AGL08]. Reaching a fixed precision requires a number of basis functions exponential in the dimension d. Exploiting the structure of the problem like [GMQ11] may reduce this effect.

Remark on the use of T^ρ: As previously mentioned, T_τ^ρ can be used for $\rho \geq 1$, instead of T_τ. In the error bounds, τ is replaced by $\tau\rho$, which can be advantageous for D fixed. Considering only constant actions during ρ steps, T_τ^ρ is very close to $T_{\tau\rho}$, up to the Euler scheme used to compute the dynamics (ρ steps of size τ *vs.* one step of size $\tau\rho$).

6. Adaptive Selection of Basis Functions

From a partition $(A_1, ..., A_n)$ of the state space, we define a dictionary $\mathcal{W} = \mathcal{Z}$ of soft-indicators $w_i(.) = -c\text{dist}(., A_i)^2$. Running Algorithm 1 with this dictionary returns a value function that is *almost* piecewise constant with respect to the partition (when c is large). This is a way to discretize the MDP, but the performance of the final policy will depend on the partition [MM02, BS08]. Typically, a uniform partition of $\mathcal{S}$ might not be the best choice for all MDPs. For instance some areas of $\mathcal{S}$ with very low optimal value function are usually not encountered in optimal trajectories, hence spending computational power there would be useless. On the contrary, a sharper approximation of the value function in other areas is critical to the performance of the policy.

We propose an algorithm to build the partition adaptively, with a simple greedy heuristic. Starting from a coarse partition, we compute the approximate value function, and then we select one of the $(A_i)_{1\leq i\leq n}$ that we want to refine, according to some criterion to be described later. Then we split this cluster into new sub-clusters partitioning it and replace it by them in the partition. If the clusters are rectangular parallelepipeds in dimension d, a simple splitting strategy is to subdivide it into 2^d smaller parallelepipeds, by a middle cut along each dimension. In a two-dimensional state space, this corresponds to building a quadtree [FB74]. Formally, a cluster A with a soft-indicator w is split into $C_1, ..., C_{2^d}$:

$$A = \bigcup_{j=1}^{2^d} C_j \quad \text{with } \forall i \neq j \in \{1, ..., 2^d\}, C_i \cap C_j = \varnothing.$$

Criterion for cluster selection: The efficiency of the partition hinges on the strategy used to select the cluster to split at each step. We maintain a dictionary $\mathcal{W}$ of soft-indicators associated to a partition $(A_1, ..., A_n)$ and another dictionary $\mathcal{Z}$ with partition $(B_1, ..., B_n)$. Following the idea of matching pursuit [MZ93], a simple heuristic is to split the cluster with highest Bellman error $|TV(s) - V(s)|$. Since two dictionaries are maintained, the origin of this error will be shared between $\mathcal{W}$ and $\mathcal{Z}$, which will lead to a possibly different cluster selected in each dictionary.

We define a grid $G = (s_1, ..., s_p)$ covering $\mathcal{S}$ where we evaluate the Bellman error. Assuming the $\langle z, Tw\rangle$ are computed exactly, after convergence of reduced value iteration, we obtain fixed points α and β such that:

$$\begin{cases} \beta = K\alpha = Z^\top TW\alpha \\ \alpha = M\beta = W^+ Z^{\top +}\beta. \end{cases}$$

Let $V = W\alpha$ and $U = Z^{\top +}\beta$, we get $V = WW^+U$ and $Z^{\top +}Z^\top TV = U$, and then the decomposition:

$$\begin{aligned} V - TV &= (V - Z^{\top +}Z^\top TV) + (Z^{\top +}Z^\top TV - TV) \\ &= (V - U) + (U - TV). \end{aligned}$$

For $s \in G$, $|V(s) - TV(s)| \leq |V(s) - U(s)| + |U(s) - TV(s)|$. The first term is the difference between U and its lower projection on the span of $\mathcal{W}$, the second one between TV and its upper projection on the span of $\mathcal{Z}$. This suggests to:

- select cluster $A \ni s \in \operatorname{argmax}_{s \in G} U(s) - V(s)$ in $\mathcal{W}$,
- select cluster $B \ni s \in \operatorname{argmax}_{s \in G} U(s) - TV(s)$ in $\mathcal{Z}$.

This strategy greedily targets areas of $\mathcal{S}$ where the projection errors should be reduced and refines the dictionaries locally. One could imagine other selection criteria, such as favoring areas with high value function or near the initialization of the trajectories if it is fixed. Alternative strategies [Bac19] include using basis functions depending on a subset of the state variables to capture local lower-dimensional dependencies. Furthermore, online methods could be applied to incorporate exploration, especially techniques based on upper-confidence bounds, as done in [BS08] on a similar problem.

7. Experiments

Setting: We test our method on two standard deterministic MDPs from the `gym` library of reinforcement learning environments [BCP+16]. Both are time-discretizations of control problems, with state dimension 2 in `Mountain`, 4 in `Cartpole`. We test uniform max-plus partitions and the adaptive basis procedure, with respectively $\rho = 5$ and $\rho = 1$ for problems 1 and 2. For comparison, we also run standard value iteration on discretizations of the MDPs. To ensure differentiability, the reward function is slightly smoothed as a sigmoid function for all three methods; γ is set identical across methods ($\gamma = 0.999$ in problem 1, 0.99 in problem 2).

The optimal value function V^* being unknown, the methods cannot be evaluated by $\|V - V^*\|_\infty$. Instead we evaluate the performance of the greedy policy π with respect to V on the task. The standard performance criterion proposed by `gym` is the cumulative reward averaged on 100 consecutive runs. The randomness only comes from the initialization of the trajectories drawn from a Gaussian around equilibrium positions. The results are plotted in Figures 1 and 2. We give the mean cumulative reward in solid line, as well as the first and third quartiles in shaded colors. The x-axis represents the number of parameters of the value function, that is, either the number of basis functions in the dictionaries $\mathcal{W}$ or $\mathcal{Z}$ or the number of states in the direct discretization of the MDP.

Results: Value iteration on the discretized MDP requires a very sharp discretization to get an efficient policy. While it is still achievable for such small MDPs, it is not reliable in higher dimension. The max-plus approximation computes *almost* piecewise constant value functions that lead to efficient policies for a much smaller number of parameters. On `Mountain`, the number of parameters is reduced from 10^5 to 10^2 from a direct discretization to the max-plus discretization, for similar performances of the policies. Finally, the adaptive basis selection method further improves the ratio between performance and number of parameters. It provides compact representations of V, faster to compute and leading to faster online evaluations of π during inference.

8. Conclusion

The max-plus linear approximation method for deterministic continuous-state MDPs with a suitable choice a basis functions provides an intuitive state-discretization. While it is still subject to the curse of dimensionality, the discretization can be adapted to a specific MDP and turns out to be effective in numerical examples. The same approach can be adapted to the Q-function for deterministic MDPs, although the potential benefits are unclear in a model-based setting. The extension to non deterministic MDPs is not straightforward and provides an interesting avenue for future research.

Acknowledgements

This work was supported by the Direction Générale de l'Armement, and by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

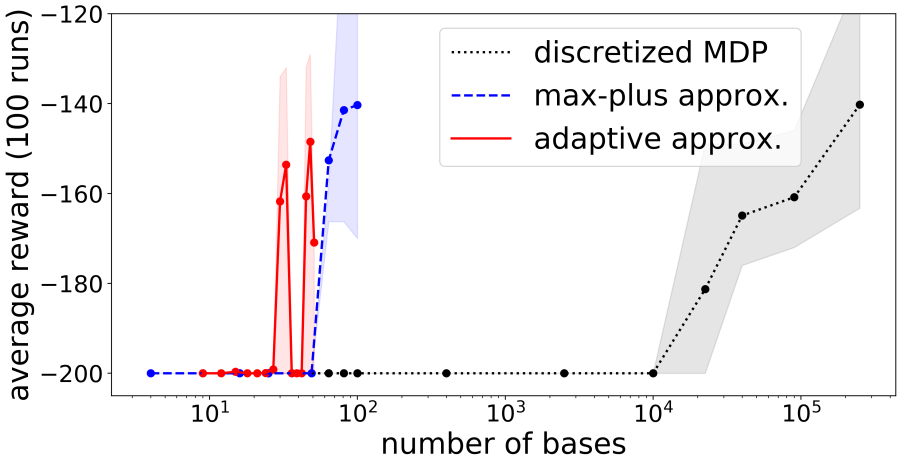

FIGURE 1. Average performance of the three approximation methods on `Mountain` as a function of the number of parameters.

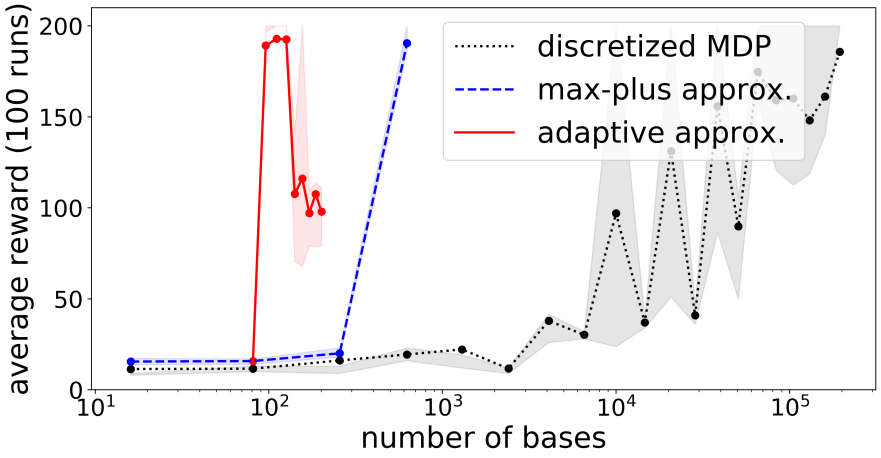

FIGURE 2. Average performance of the three approximation methods on `Cartpole` as a function of the number of parameters.

REFERENCES

[AGL08] 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** (2008), no. 2, 817–848.

[Bac19] Francis Bach, *Max-plus matching pursuit for deterministic Markov decision processes*, working paper or preprint, June 2019.

[BCP+16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba, *OpenAI gym*, 2016.

[BS08] Andrey Bernstein and Nahum Shimkin, *Adaptive aggregation for reinforcement learning with efficient exploration: Deterministic domains.*, COLT, 2008, pp. 323–334.

[CGQ04] Guy Cohen, Stéphane Gaubert, and Jean-Pierre Quadrat, *Duality and separation theorems in idempotent semimodules*, Linear Algebra and its Applications **379** (2004), 395–422.

[FB74] Raphael Finkel and Jon Bentley, *Quad trees: A data structure for retrieval on composite keys.*, Acta Inf. **4** (1974), 1–9.

[FS06] Wendell H Fleming and Halil Mete Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer Science & Business Media, 2006.

[GMQ11] Stephane Gaubert, William McEneaney, and Zheng Qu, *Curse of dimensionality reduction in max-plus based approximation methods: Theoretical estimates and improved pruning algorithms*, 2011 50th IEEE Conference on Decision and Control, IEEE, 2011, pp. 1054–1061.

[GP97] Stephane Gaubert and Max Plus, *Methods and applications of (max,+) linear algebra*, Annual Symposium on Theoretical Aspects of Computer Science, Springer, 1997, pp. 261–282.

[HLL12] Onésimo Hernández-Lerma and Jean B Lasserre, *Discrete-time Markov Control Processes: Basic Optimality Criteria*, vol. 30, Springer Science & Business Media, 2012.

[Lib11] Daniel Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*, Princeton University Press, 2011.

[McE03] William M McEneaney, *Max-plus eigenvector representations for solution of nonlinear H infinity problems: basic concepts*, IEEE Transactions on Automatic Control **48** (2003), no. 7, 1150–1163.

[MM02] Rémi Munos and Andrew Moore, *Variable resolution discretization in optimal control*, Machine Learning **49** (2002), no. 2-3, 291–323.

[MM09] Prashant Mehta and Sean Meyn, *Q-learning and Pontryagin's minimum principle*, Proceedings of the 48h IEEE Conference on Decision and Control, IEEE, 2009, pp. 3598–3605.

[MZ93] Stéphane G Mallat and Zhifeng Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Transactions on Signal Processing **41** (1993), no. 12, 3397–3415.

[SB18] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, second ed., The MIT Press, 2018.

[TBO19] Corentin Tallec, Léonard Blier, and Yann Ollivier, *Making deep Q-learning methods robust to time discretization*, International Conference on Machine Learning, 2019, pp. 6096–6104.