Tropical Algebra for Value Function Approximation

Theory and Implementation

Emile Esmaili

Reinforcement Learning - 6892 - Prof. Javad Ghaderi

Table of Contents

Table of Contents

Opening Remarks

This project is

- Part theory and details on existing literature with proofs
- Part implementation of papers' results

Motivation and Scope

We look into the issue of control problems with large deterministic state-spaces (ie robotics)
Consider a continuous-state MDP (discrete-time, discrete-control). We want to discretize it into a finite MDP (discrete-state), e.g. to approximate the value function with value iteration.

Problem: A naive discretization has no notion of spatial proximity, hence we would need a very large state-discretization, not even fitting in memory for problems of moderate dimensions.

Motivation and Scope

We consider a deterministic, time-homogeneous, infinite-horizon, discounted MDP defined by:

- a state space S,
- an action space A,
- a bounded reward function $r : S \times A \to [-R, R]$,
- dynamics $\phi(\cdot) : S \times A \to S$,
- and a discount factor $0 \leq \gamma < 1$.

We make the following assumptions:

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

Value Iteration

The optimal value function $V^* : S \to \mathbb{R}$ corresponds to an optimal policy $\pi^* : S \to A$ maximizing the cumulative discounted reward. The greedy policy π corresponding to a value function V is then:

$$\pi(s) \in \arg\max_{a \in A} \left[r(s,a) + \gamma V(\phi_a(s)) \right].$$

The value iteration algorithm consists in computing V^* as the unique fixed point of the Bellman operator $T : \mathbb{R}^S \to \mathbb{R}^S$:

$$TV(s) := \max_{a \in A} \left[r(s,a) + \gamma V(\phi_a(s)) \right].$$

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

Value Function Approximation

We have seen a regular linear parameterization of the value function, as

$$V(s) = \sum_{w \in W} \alpha_w \cdot w(s)$$

where W is a set of basis functions $w : S \to \mathbb{R}$.

Idea: What if we use a 'tropical' or **max-plus** linear approximation instead?

Table of Contents

Primer on Tropical Algebra

In an exotic country, children are taught that:

$$\text{“}a + b\text{”} = \max(a, b) \quad ; \text{“}a \times b\text{”} = a + b$$

So

- $\text{“}2 + 3\text{”} = 3$
- $\text{“}2 \times 3\text{”} = 5$
- $\text{“}5/2\text{”} = 3$
- $\text{“}2^3\text{”} = \text{“}2 \times 2 \times 2\text{”} = 6$
- $\text{“}\sqrt{-1}\text{”} = -0.5$

Primer on Tropical Algebra

The max-plus semiring $(\mathbb{R}_{\max}, \oplus, \otimes)$ is the set $\mathbb{R} \cup \{-\infty\}$, equipped with the two operations:

$$x \oplus y = \max\{x, y\}$$
$$x \otimes y = x + y$$

The relations $\oplus$ and $\otimes$ are associative and commutative. The 0 element for $\oplus$ is $-\infty$, which is such that:

$$x \oplus (-\infty) = \max\{x, -\infty\} = x$$

The **1** element for $\otimes$ is 0, such that $x \otimes 0 = x + 0 = x$. All non-zero elements (i.e., different from $-\infty$) have an inverse for $\otimes$, equal to $-x$ (hence making the structure a semifield):
An interesting property is that the semiring is idempotent:

$$x \oplus x = \max\{x, x\} = x$$

Max-Plus Linear Algebra

Consider the following linear system, with unknown $z = (x, y) \in \mathbb{R}^2$:

$$\begin{pmatrix} \mathbf{1} & 2 \\ -4 & \mathbf{1} \end{pmatrix} \otimes \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Unrolling the max-plus notations, this is equivalent to the following system of equations:

$$\max\{x, y + 2\} = 1$$
$$\max\{x - 4, y\} = 2$$

The first line is equivalent to:

$$(x = 1 \text{ and } y + 2 \leq 1) \text{ or } (x \leq 1 \text{ and } y + 2 = 1)$$

with a similar condition for the second line:

$$(x - 4 = 2 \text{ and } y \leq 2) \text{ or } (x - 4 \leq 2 \text{ and } y = 2).$$

Table of Contents

Max-Plus linearity of Bellman backup

The structure of the Bellman operator $T : \mathbb{R}^S \to \mathbb{R}^S$ is naturally compatible with max-plus algebra. It is max-plus additive and homogeneous:

Bellman backup TV(s) is MaxPlus linear

Proof:

$$T(V \oplus V_0) = T(\max\{V, V_0\}) = \max\{TV, TV_0\} = TV \oplus TV_0$$

$$T(c \otimes V) = T(c + V) = \gamma c + TV = c^{\otimes\gamma} TV.$$

Max Plus linear combinations

Let W be a finite dictionary of functions $w : S \to \mathbb{R}$.
For $\alpha \in \mathbb{R}^W$, we define the max-plus linear combinations:

$$V(s) = \bigoplus_{w \in W} \alpha(w) \otimes w(s) = \max_{w \in W} \left[\alpha(w) + w(s)\right].$$

and we write it more compactly:

$$V = W\alpha.$$

We can also define a dot product:

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

Max-Plus basis functions

Idea from Bach [1] : the value function can be approximated by a max-plus linear combination of functions in W.
The functions $w(s)$ form a **basis** in the max-plus linear approximation of V
Most common dictionaries of functions:

- Smooth: $w_i(s) = -c\|s - s_i\|^2$
- Lipschitz: $w_i(s) = -c\|s - s_i\|$
- Indicator: $w_i(s) = \begin{cases} 0 & \text{if } s \in A(w_i) \\ -\infty & \text{otherwise} \end{cases}$
- Soft indicator: $w_i(s) = -c dist(s, A(w_i))^2$

Smooth or Lipschitz basis functions are used to approximate value functions of the same regularity, controlled by c. (Akian et al. [2])

Piecewise constant value functions are good candidates for a discretization. They are used in Bach [1] to cluster similar states in discrete MDPs.

Max-Plus Linear Projections

Define the following four operators:

- $W : \mathbb{R}^W \to \mathbb{R}^S$, $W\alpha(s) := \max_{w \in W}[\alpha(w) + w(s)]$
- $W^+ : \mathbb{R}^S \to \mathbb{R}^W$, $W^+ V(w) := \inf_{s \in S}[V(s) - w(s)]$
- $W^\top : \mathbb{R}^S \to \mathbb{R}^W$, $W^\top V(w) := \sup_{s \in S}[V(s) + w(s)]$
- $W^{\top +} : \mathbb{R}^W \to \mathbb{R}^S$, $W^{\top +}\alpha(s) := \min_{w \in W}[\alpha(w) - w(s)]$

Max-Plus Linear Projections

W^+ acts like a pseudo inverse

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

$$\begin{aligned} &\forall s \in S,\ W\alpha(s) \leq V(s) \\ &\iff \forall (s, w) \in S \times W,\ \alpha(w) + w(s) \leq V(s) \\ &\iff \forall (s, w) \in S \times W,\ \alpha(w) \leq V(s) - w(s) \\ &\iff \forall w \in W,\ \alpha(w) \leq W^+V(w). \end{aligned}$$

As shown in Akian et al. [2], $WW^+ = W$ and $W^+W^+ = W^+$
Therefore W^+ plays a role of pseudo-inverse, and WW^+ the role of projection on the image of W.

Max Plus Linear Projections

Idea: Projection on the range of W

Replace $V_{t+1} = TV_t$ by $V_{t+1} = WW^+V_t$

If we consider V_t of the form $V_t = W\alpha_t$, then $V_{t+1} = W\alpha_{t+1}$ with
$\alpha_{t+1}(w) = W^+TW\alpha_t(w) = \min_{s\in S}\{\max_{w'\in W}\gamma\alpha_t(w') + Tw'(s)\} - w(s)$

Which comes from Max-plus homogeneity of $T(W\alpha)$

$$T(W\alpha) = T(\bigoplus w \otimes \alpha) = \bigoplus \alpha\gamma + Tw = \max_{w} \gamma\alpha + Tw$$

This requires to solve at each iteration an infimum problem over S, which is computationally expensive as $O(|S| \cdot |W|)$, which is typically worse than classical value iteration. Not good!

Variational Trick

Better Idea from [1]: Use a variational formulation with another basis of functions Z. Define, similarly to what we did with W:

- $Z^\top V(z) = \max_{s \in S} V(s) + z(s)$.
- $Z^{\top +} \beta(s) = \min_{z \in Z} \beta(z) - z(s)$.

The operator $Z^{\top +} Z^\top$ on functions from S to $\mathbb{R}$ is the projection on the image of $Z^{\top +}$.

The value iteration recursion $V_{k+1} = TV_k$ is replaced by a variational formulation:

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

of which we consider the maximal solution in span(W) [2]:

$$V_{k+1} = WW^+ Z^{\top +} Z^+ TV_k.$$

If $V_k = W\alpha_k$, we have the following recursion:

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

Reduced Value Iteration

The operator $W^+ Z^{\top +} Z^\top T W : \mathbb{R}^{\mathbb{W}} \to \mathbb{R}^{\mathbb{W}}$ decomposes as $M \circ K$ with
$K = Z^\top W : \mathbb{R}^{\mathbb{W}} \to \mathbb{R}^{\mathbb{Z}}$
$M = W^+ Z^{\top +} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{W}}$

The recursion can be reformulated :

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

We can then recover the optimal Value function as

$$V^* = W\alpha$$

Reduced Value Iteration: Convergence guarantee

Proposition 1 from Bach [1]

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

Table of Contents

Implementations

- We **reproduce and implement** the results from Bach [1]
- and add other illustrations

Implementing Reduced VI from [1]

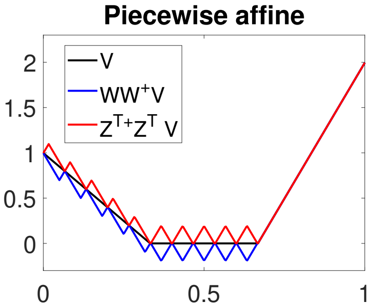
First in a 1D state-space.
We reproduce the results of [1] using the following setup:

- $|S| = 2^8, |A| = 2$, Discretized MDP from continuous control problem
- discount factor for continous control problem $\eta = 0.5$, for MDP $\gamma = \eta/|S|$
- convex and non-convex reward functions
- convex reward is given by

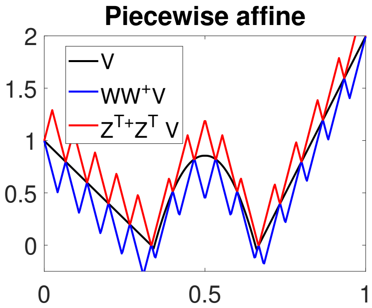
$$R(x) = |(1 - 3x) \cdot \mathbf{1}_{x<1/3} + (6x - 4) \cdot \mathbf{1}_{x>2/3}| \\ -log(\eta)(-3) \cdot \mathbf{1}_{x<1/3} + (6) \cdot \mathbf{1}_{x>2/3}$$

- This is a theoretical setup where we **know** V^*

What do the projections look like in a 1D space?



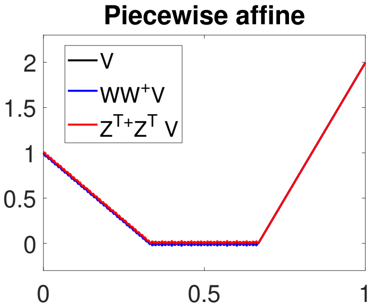
(a) 16 affine bases, convex reward



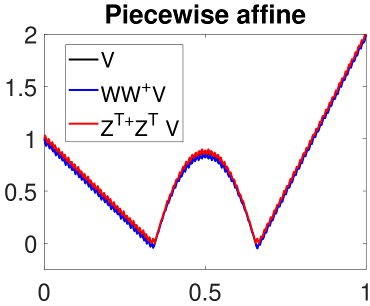
(b) 16 affine bases, non-convex reward

Figure: Upper and lower projections error for 16 basis functions

What do the projections look like in a 1D space?



(a) 100 affine bases, convex reward

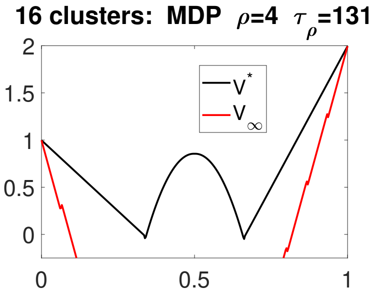


(b) 100 affine bases, non-convex reward

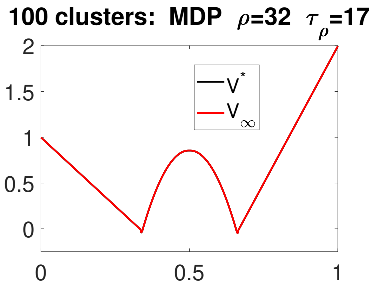
Figure: Near-perfect approximation with upper and lower projections

Solving the MDP with reduced VI (1D)

$\tau = (1-\gamma)^{-1}$ (larger = large horizon)
ρ is such that discount factor is γ^ρ



(a) 16 affine bases, nonconvex reward



(b) 100 affine bases, non-convex reward

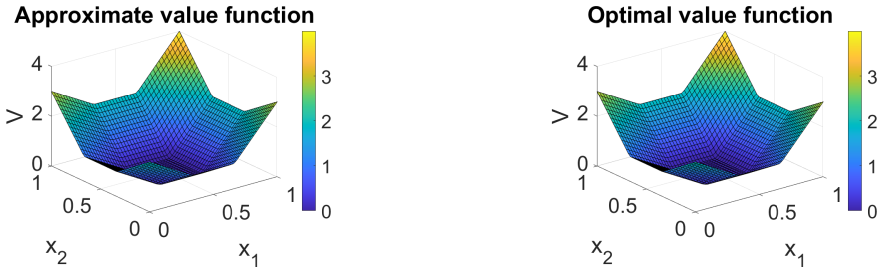
Figure: Solving a control problem with reduced VI

Reduced VI: 2D state space

The setup is now a 2D state space with $|S| = 2^5 \times 2^5$
We adapt the rewards to be multivariable functions $R(x, y)$
This is already more realistic for control problems

Reduced VI: 2D state space

64 basis functions



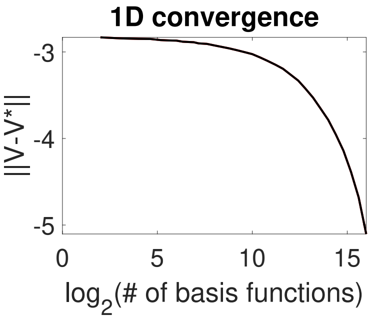
(a) Approximate Value function

(b) Optimal Value function

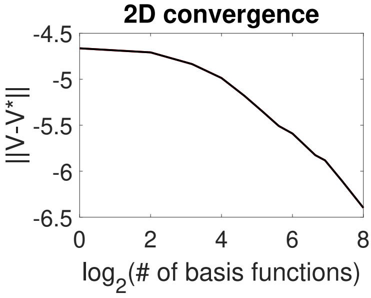
Figure: Max-plus approximation of V with a 2D state space

Performance plots

Now let's look at the convergence $||V^* - V_{approx}||$ as a function of the number of basis functions



(a) 1D state-space



(b) 2D state-space

Figure: Convergence plots

Table of Contents

Extensions

This is very theoretical but some recent papers looked at extensions:

- When the MDP does not come from an underlying continuous-time problem, the quantity $\langle z, Tw \rangle$ can be hard to compute. Berthier and Bach [3] use a gradient ascent technique to use Reduced Value Iteration on MDPs.
- Gonçalves [4] discusses extension to online learning. Possible extensions:
  - Q-values! What if we approximate $Q(s, a)$ with tropical linear projections?
  - what about stochastic MDPs?

References

[1] Francis R. Bach. Max-plus matching pursuit for deterministic markov decision processes. *CoRR*, abs/1906.08524, 2019. URL http://arxiv.org/abs/1906.08524.

[2] 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. doi: 10.1137/060655286. URL https://doi.org/10.1137/060655286.

[3] Eloïse Berthier and Francis Bach. Max-plus linear approximations for deterministic continuous-state markov decision processes. *IEEE Control Systems Letters*, 4(3):767–772, 2020.

[4] Vinicius Mariano Gonçalves. Max-plus approximation for reinforcement learning. *Automatica*, 129:109623, 2021. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2021.109623. URL https://www.sciencedirect.com/science/article/pii/S0005109821001436.