Algorithmes efficaces pour le contrôle et l'apprentissage par renforcement

CJC-MA 2021, Palaiseau

Eloïse BERTHIER

Jeudi 28 octobre 2021

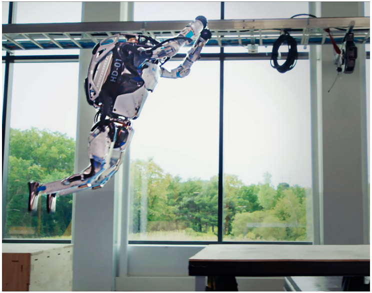
Plan

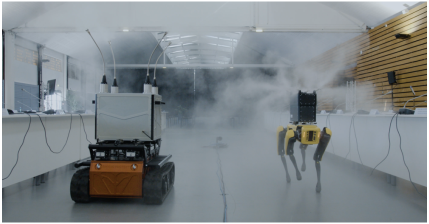
Contrôle optimal & Apprentissage par renforcement

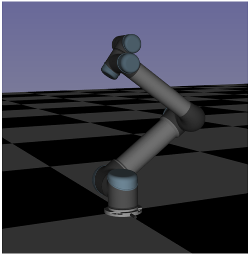
$$\min_{u(.)} \int_0^T L(x(t), u(t))\mathrm{d}t$$
$$\dot{x}(t) = f(x(t), u(t)).$$

$$\max_{a_0,\dots,a_{n-1}} \mathbb{E}\left[\sum_{t=1}^{n} r_t\right]$$
$$s_{t+1} \sim P(s_t, a_t)$$
$$r_{t+1} \sim R(s_t, a_t).$$

Motivation : Application à la robotique







France 2030 : Emmanuel Macron annonce 800 millions d'euros pour développer le secteur de la robotique

Publié le 25/10/2021 15:13

Motivation : Application à la robotique

Contrôler un robot présente plusieurs difficultés :

- Les systèmes sont en **dimension** (relativement) grande
  $\Longrightarrow$ on ne peut espérer résoudre exactement les problèmes de contrôle optimal (*malédiction de la dimension*).
- Les systèmes dynamiques sont **non-linéaires**
  $\Longrightarrow$ pas d'utilisation directe des outils du contrôle linéaire.
- Le **modèle** du système est imparfait
  $\Longrightarrow$ une solution exacte aurait peu d'intérêt.
- Certains calculs doivent être faits en **temps réel**, ou sur des **systèmes embarqués**
  $\Longrightarrow$ puissance et temps de calcul disponibles sont limités.

Plan

Idée générale

On considère un **Processus de Décision Markovien** (MDP) à espace d'état continu (temps et contrôle discrets). On cherche à le **discrétiser** en un MDP fini (état discret), par exemple pour approximer la fonction valeur avec l'algorithme de value iteration.

Problème : Une discrétisation naïve n'a pas de notion de proximité spatiale. Pour capturer la dynamique, pour une discrétisation de pas ε, la taille en mémoire $O(\varepsilon^{-d})$ explose avec la dimension d.

On suit l'approche de [McE03, AGL08], pour calculer une approximation max-plus linéaire de la fonction valeur [1].

1. E. Berthier, F. Bach. Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes. *IEEE Control Systems Letters*, 4(3) :767-772, 2020.

Processus de décision Markovien

On considère un MDP déterministe, à horizon infini, defini par :

- un **espace d'états** borné $\mathcal{S} \subset \mathbb{R}^d$,
- un **espace d'actions** $\mathcal{A}$ fini,
- une **fonction de récompense** $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$,
- une **dynamique** $\varphi_.(.) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$,
- un **facteur d'actualisation** $0 \leq \gamma < 1$.

La **fonction valeur** $V^* : \mathcal{S} \to \mathbb{R}$ est la récompense cumulée actualisée maximale :

$$V^*(s) = \max_{\pi} \sum_{t=0}^{\infty} \gamma^t r_t, \quad s_0 = s.$$

Elle correspond à une **politique optimale** $\pi^* : \mathcal{S} \to \mathcal{A}$:

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

Value Iteration

L'algorithme de **value iteration** consiste à calculer V^* comme l'unique **point fixe** de l'opérateur de Bellman $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$:

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

L'algorithme calcule itérativement $V_{k+1} = TV_k$, et converge linéairement vers V^*. Si $\mathcal{S}$ est un ensemble fini, l'algorithme fait $O(|\mathcal{A}| \cdot |\mathcal{S}|)$ opérations, et stocke $O(|\mathcal{S}|)$ valeurs à chaque itération.

Propriétés max-plus de l'opérateur de Bellman

Le semi-anneau max-plus est défini par $(\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$, où $\oplus$ représente l'opérateur maximum, et $\otimes$ la somme usuelle.

La structure de l'opérateur de Bellman :

$$T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$$
$$TV(s) = \max_{a \in \mathcal{A}} r(s,a) + \gamma V(\varphi_a(s))$$

est naturellement **compatible** avec l'algèbre max-plus. T est max-plus additif et homogène :

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

L'additivité n'est plus vérifiée pour des MDP stochastiques.

Approximation max-plus linéaire

Soit $\mathcal{W}$ un dictionnaire de fonctions $w : \mathcal{S} \to \mathbb{R}$. On approxime V comme une combinaison **max-plus linéaire** de fonctions de $\mathcal{W}$:

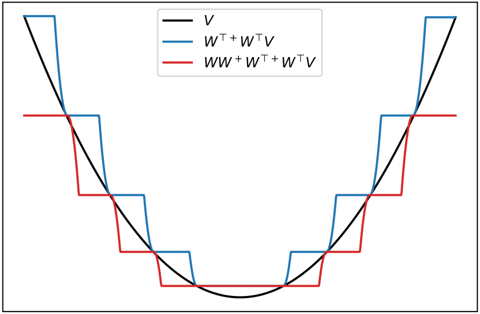
$$V(s) = \bigoplus_{w \in \mathcal{W}} \alpha(w) \otimes w(s) = \max_{w \in \mathcal{W}} \alpha(w) + w(s).$$

Dictionnaires de fonctions :

- Lisses : $w(s) = -c\|s - s_0\|^2$
- Indicatrices : $w(s) = \begin{cases} 0 & \text{si } s \in A_w \\ -\infty & \text{sinon.} \end{cases}$
- ...

Les indicatrices permettent d'approximer V^* avec une fonction constante par morceaux (discrétisation).

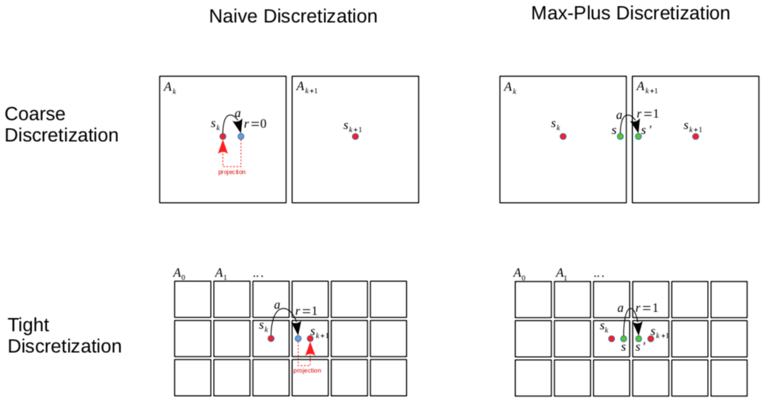
Approximation max-plus linéaire



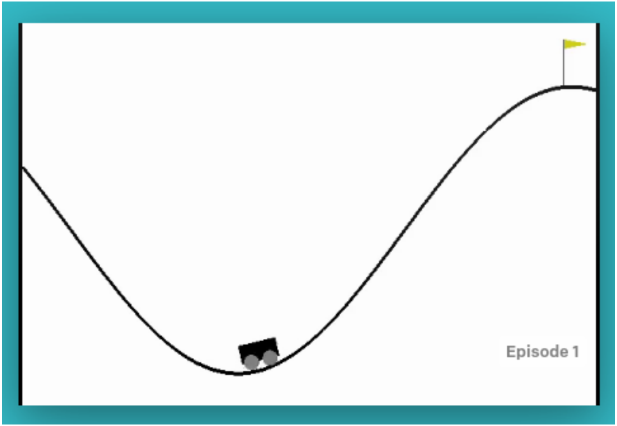
Exemple d'approximation max-plus linéaire avec indicatrices.

Max-plus value iteration

Pour calculer cette approximation : Max-plus value iteration.

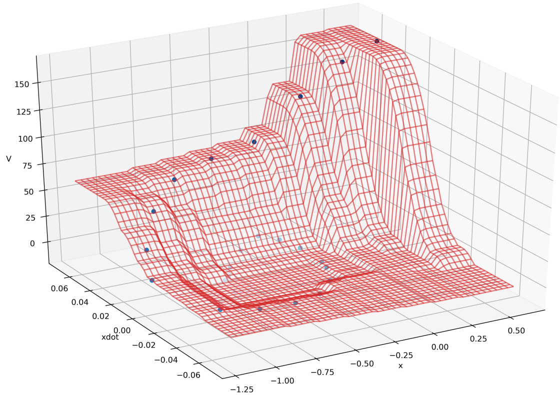


Exemple numérique



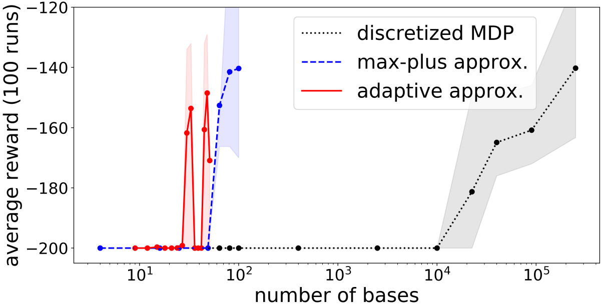
Mountain MDP ($d = 2$).

Exemple numérique



Fonction valeur obtenue avec la max-plus value iteration.

Exemple numérique



On obtient une discrétisation plus compacte.

Plan

Idée générale

Constat : Les problèmes de contrôle optimal sont difficiles à résoudre numériquement, même pour des systèmes de dimension modeste.

On étend la méthode de [LHPT08] pour en calculer une approximation [2] :

- dans le cas des **problèmes lisses**,
- de façon "**black-box**" (à partir d'observations, sans gradients),
- qui peut passer à l'échelle en **grande dimension**.

2. E. Berthier, J. Carpentier, A. Rudi, F. Bach. Infinite-Dimensional Sums-of-Squares for Optimal Control. *Technical Report*, ⟨hal-03377120⟩, 2021.

Problème de contrôle optimal

On considère le problème à horizon fini, avec des espaces d'état et de contrôle compacts $\mathcal{X} \subset \mathbb{R}^d$ et $\mathcal{U} \subset \mathbb{R}^p$:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t))\mathrm{d}t + M(x(T))$$
$$\forall t \in [t_0, T],\ \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0. \tag{OCP}$$

Formulation faible du problème de contrôle

Sous certaines conditions le problème est équivalent à :

$$\begin{aligned} &\sup_{V \in C^1([0,T]\times\mathcal{X})} \quad V(0,x_0) \\ \forall(t,x,u),\ &\frac{\partial V}{\partial t}(t,x) + L(t,x,u) + \nabla V(t,x)^\top f(t,x,u) \geq 0 \\ &\forall x,\ V(T,x) \leq M(x). \end{aligned} \tag{P}$$

On cherche V dans un espace de dimension fini $\mathcal{F}$ paramétrisé par $\theta \in \mathbb{R}^m$. Si $V^* \in \mathcal{F}$, (P) est un programme linéaire, avec **un ensemble dense de contraintes** de la forme :

$$\forall(t,x,u), \quad H(t,x,u) \geq 0.$$

Première idée : discrétisation naïve

Sous-échantiller les contraintes donne une **relaxation** simple :

$$\sup_{\theta \in \mathbb{R}^m} \quad V_\theta(0, x_0) - \lambda_\theta \|\theta\|_2^2$$
$$\forall i \in I, \quad H(t^{(i)}, x^{(i)}, u^{(i)}) \geq 0. \tag{LP}$$

Cette relaxation est connue dans le cas de l'**optimisation globale** :

$$\sup c \;\; \text{s.t. } \forall y \in \mathbb{R}^p, \; g(y) - c \geq 0.$$

Le sous-échantillonnage équivaut à $\min g \simeq \min\{g(x_i)\}$. Il faut $O(\varepsilon^{-p})$ échantillons pour approximer $\min g$ avec précision ε.
Si $g \in C^s(\mathbb{R}^p)$ est lisse, ce taux peut être amélioré à $O(\varepsilon^{-p/s})$ en utilisant une représentation "sommes de carrés".

Représentation de fonctions positives par sommes de carrés

On a un ensemble dense de contraintes :

$$\forall (t, x, u), \quad H(t, x, u) \geq 0.$$

Dans le cas polynomial : la méthode "moments - sommes de carrés" ou hiérarchie de Lasserre [Las10] consiste à représenter les polynômes positifs par une somme de carrés de polynômes :

$$p(x) = \sum_{k=1}^{\ell} p_k(x)^2 \geq 0.$$

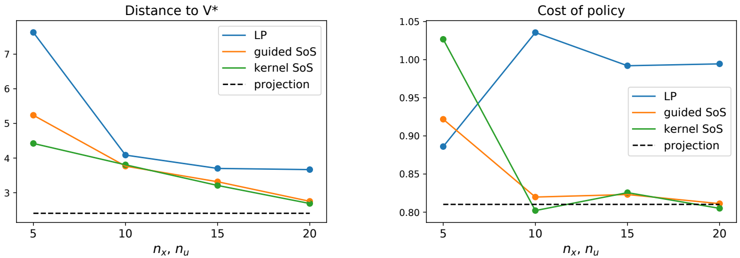
Dans le cas lisse : une représentation similaire a été introduite récemment [MFBR20, RMFB20] pour représenter des fonctions positives dans un espace de Sobolev, en utilisant une **méthode à noyaux**.

Problème en dimension finie

Après sous-échantillonnage, la méthode conduit à un **programme SDP** de la forme :

$$\sup_{B \succcurlyeq 0, \theta \in \mathbb{R}^m} c^\top \theta - \lambda_\theta \|\theta\|_2^2 - \lambda \mathrm{Tr}(B) + C$$
$$\text{such that } \forall i \in \{1, \ldots, n\},\ b_i + a_i^\top \theta = (\Phi_i)^\top B \Phi_i. \qquad \text{(KSOS)}$$

Exemple numérique



Cas d'un problème LQR en dimension 2.

Conclusion

- On regarde des problèmes où on est confronté à la "**maléduction de la dimension**".
- On cherche à faire le lien entre **contrôle optimal** et **apprentissage par renforcement** (discret et continu, model-based et model-free...).
- On cherche à **exploiter la structure** des problèmes de contrôle, comme c'était le cas des problèmes d'optimisation (lisse, fortement convexe...) et d'apprentissage (modèles linéaires...)

$\rightarrow$ *Comment reconnaître les problèmes "faciles" en contrôle ?*

Références

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.

Jean-Bernard Lasserre, *Moments, positive polynomials and their applications*, vol. 1, World Scientific, 2010.

Jean-Bernard Lasserre, Didier Henrion, Christophe Prieur, and Emmanuel Trélat, *Nonlinear optimal control via occupation measures and LMI-relaxations*, SIAM J. on Control and Optim. **47** (2008), no. 4, 1643–1666.

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.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi, *Non-parametric models for non-negative functions*, Advances in Neural Information Processing Systems, 2020.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach, *Finding global minima via kernel approximations*, Tech. Report 2012.11978, arXiv, 2020.