Efficient Algorithms for Control and Reinforcement Learning

Eloïse Berthier

Supervised by Francis Bach

October 27, 2022

Prelude: A diversity of control problems

Prelude: A diversity of control problems

Prelude: A diversity of control problems

Prelude: A diversity of control problems

Contents

1 Introduction
- Optimal Control
- Reinforcement Learning
- Research Questions & Contributions

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

The optimal control problem

An optimization problem [Liberzon, 2011]:

$$\inf_{u(\cdot)} \int_0^T L(x(t), u(t))\,\mathrm{d}t + M(x(T))$$
$$\text{s.t.} \quad \forall t \in [0, T], \quad \dot{x}(t) = f(x(t), u(t))$$
$$x(0) = x_0.$$

Ingredients:

The optimal control problem

An optimization problem [Liberzon, 2011]:

$$\begin{aligned} \inf_{u(\cdot)} \int_0^T & L(x(t), u(t))\,\mathrm{d}t + M(x(T)) \\ \text{s.t.} \quad \forall t \in [0, T], & \quad \boxed{\dot{x}(t) = f(x(t), u(t))} \\ x(0) &= x_0. \end{aligned}$$

Ingredients:

- A controlled dynamics

The optimal control problem

An optimization problem [Liberzon, 2011]:

$$\begin{aligned} \inf_{u(\cdot)} \quad & \int_0^T \boxed{L(x(t), u(t))}\, \mathrm{d}t + \boxed{M(x(T))} \\ \text{s.t.} \quad & \forall t \in [0, T], \quad \dot{x}(t) = f(x(t), u(t)) \\ & x(0) = x_0. \end{aligned}$$

Ingredients:

- A controlled dynamics
- A running cost and a terminal cost

The optimal control problem

An optimization problem [Liberzon, 2011]:

$$\boxed{\inf_{u(\cdot)}} \int_0^T L(x(t), u(t))\,\mathrm{d}t + M(x(T))$$
$$\text{s.t.} \quad \forall t \in [0, T], \quad \dot{x}(t) = f(x(t), u(t))$$
$$x(0) = x_0.$$

Ingredients:

- A controlled dynamics
- A running cost and a terminal cost
- An infinite-dimensional minimization problem

Optimality conditions

Parallel approaches to solve optimal control problems [Trélat, 2005]:

Optimality conditions

Parallel approaches to solve optimal control problems [Trélat, 2005]:

- **Pontryagin's Maximum Principle** [Pontryagin et al., 1974]: generalization of the Karush–Kuhn–Tucker necessary conditions. $\rightarrow$ indirect shooting methods.

Optimality conditions

Parallel approaches to solve optimal control problems [Trélat, 2005]:

- **Pontryagin's Maximum Principle** [Pontryagin et al., 1974]:
  generalization of the Karush–Kuhn–Tucker necessary conditions.
  $\rightarrow$ indirect shooting methods.

- **Bellman's Optimality Principle** [Bellman, 1954]:
  "Whatever the first decisions, the remaining ones must be optimal with regard to the state resulting from the first decisions."
  $\rightarrow$ dynamic programming.

Optimality conditions: the value function

Key object: the **value function**

$$
\begin{aligned}
V^*(t_0, x_0) = \inf_{u(\cdot)} & \int_{t_0}^{T} L(x(t), u(t))\,\mathrm{d}t + M(x(T)) \\
\text{s.t.} \quad & \forall t \in [t_0, T], \quad \dot{x}(t) = f(x(t), u(t)) \\
& x(t_0) = x_0.
\end{aligned}
$$

Optimality conditions: the value function

Key object: the **value function**

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

The Hamilton-Jacobi-Bellman PDE [Crandall, Evan and Lions, 1984]:

$$
\begin{aligned}
\forall (t, x), \quad & \frac{\partial V}{\partial t}(t, x) + \inf_{u \in \mathcal{U}} \left\{ L(x, u) + \nabla V(t, x)^\top f(x, u) \right\} = 0 \\
\forall x, \quad & V(T, x) = M(x).
\end{aligned}
$$

Contents

1 Introduction
- Optimal Control
- Reinforcement Learning
- Research Questions & Contributions

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

The reinforcement learning problem

A stochastic optimization problem [Sutton and Barto, 2018]:

$$\begin{aligned} &\max_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}_p\left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t))\right] \\ \text{s.t.} \quad & \forall t \in \mathbb{N}, \quad s_{t+1} \sim p(s' \mid s = s_t, a = \pi(s_t)) \\ & s_0 = s. \end{aligned}$$

Ingredients:

The reinforcement learning problem

A stochastic optimization problem [Sutton and Barto, 2018]:

$$\max_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \right]$$

$$\text{s.t.} \quad \forall t \in \mathbb{N}, \quad \boxed{s_{t+1} \sim p(s' \mid s = s_t, a = \pi(s_t))}$$

$$s_0 = s.$$

Ingredients:

- An **unknown** controlled stochastic dynamics

The reinforcement learning problem

A stochastic optimization problem [Sutton and Barto, 2018]:

$$\max_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \boxed{\gamma^t r(s_t, \pi(s_t))} \right]$$

$$\text{s.t.} \quad \forall t \in \mathbb{N}, \quad s_{t+1} \sim p(s' \mid s = s_t, a = \pi(s_t))$$

$$s_0 = s.$$

Ingredients:

- An **unknown** controlled stochastic dynamics
- An **unknown** discounted reward

The reinforcement learning problem

A stochastic optimization problem [Sutton and Barto, 2018]:

$$\begin{aligned} &\boxed{\max_{\pi:\mathcal{S}\to\mathcal{A}}} \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \right] \\ \text{s.t.} \quad &\forall t \in \mathbb{N}, \quad s_{t+1} \sim p(s' \mid s = s_t, a = \pi(s_t)) \\ &s_0 = s. \end{aligned}$$

Ingredients:

- An **unknown** controlled stochastic dynamics
- An **unknown** discounted reward
- A maximization problem

Dynamic programming

Key object: the **value function**

$$V^*(s) = \max_{\pi} \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \;\middle|\; s_0 = s \right].$$

Dynamic programming

Key object: the **value function**

$$V^*(s) = \max_{\pi} \ \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \,\middle|\, s_0 = s \right].$$

V^* is the fixed point of the Bellman operator T defined by:

$$TV(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(\cdot|s,a)} V(s') \right\}$$

Dynamic programming

Key object: the **value function**

$$V^*(s) = \max_{\pi} \ \mathbb{E}_p \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, \pi(s_t)) \ \Big| \ s_0 = s \right].$$

V^* is the fixed point of the Bellman operator T defined by:

$$TV(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{p(\cdot|s,a)} V(s') \right\}$$

Algorithms:

- *Value Iteration*: $V_k = T^k V_0$ converges to V^* if $\gamma \in [0, 1)$.
- *Temporal-Difference Learning*: estimate the Bellman operator from observed transitions, for policy evaluation.

Contents

1 Introduction
- Optimal Control
- Reinforcement Learning
- Research Questions & Contributions

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

Requirements for modern applications

- The dynamical systems are nonlinear
  $\Rightarrow$ linear control methods cannot be used directly.

Requirements for modern applications

- The dynamical systems are nonlinear
  $\Rightarrow$ linear control methods cannot be used directly.
- The dimensions of the systems are (relatively) large
  $\Rightarrow$ approximation is needed.

Requirements for modern applications

- The dynamical systems are nonlinear
  $\Rightarrow$ linear control methods cannot be used directly.
- The dimensions of the systems are (relatively) large
  $\Rightarrow$ approximation is needed.
- There are modeling uncertainties
  $\Rightarrow$ estimation is needed.

Requirements for modern applications

- The dynamical systems are nonlinear
  $\Rightarrow$ linear control methods cannot be used directly.
- The dimensions of the systems are (relatively) large
  $\Rightarrow$ approximation is needed.
- There are modeling uncertainties
  $\Rightarrow$ estimation is needed.
- Some computations are done in real-time, embedded systems
  $\Rightarrow$ memory/time efficient algorithms are needed.

Research questions

Questions explored throughout this thesis:

1. How to exploit partial knowledge of the model? [estimation]
2. How to represent the value function? [approximation]

Q1: How to exploit partial knowledge of the model?

"*The controller*"

"*The reinforcement learner*"

Q1: How to exploit partial knowledge of the model?

"The controller"

"The reinforcement learner"

known
model

Q1: How to exploit partial knowledge of the model?

"The controller"

"The reinforcement learner"

known model

approximate model

Q1: How to exploit partial knowledge of the model?

"The controller"

"The reinforcement learner"

known model | approximate model | offline observations

Q1: How to exploit partial knowledge of the model?

"The controller"

"The reinforcement learner"

$\longrightarrow$

known model	approximate model	offline observations	online observations

Q1: How to exploit partial knowledge of the model?

"*The controller*"

"*The reinforcement learner*"

known model → approximate model → offline observations → online observations → partial observability

Q2: How to represent the value function?

- If $\mathcal{S}$ is a finite set: tabular storage of $V(s)$, $s \in \{1, ..., |\mathcal{S}|\}$
  $\rightarrow$ does not fit in memory if $|\mathcal{S}|$ is too large ⚠
- If $\mathcal{S}$ is a continuous set: parameterization V_θ, $\theta \in \mathbb{R}^p$
  $\rightarrow$ curse of dimensionality if $\dim(\mathcal{S})$ is large ⚠

Q2: How to represent the value function?

- If $\mathcal{S}$ is a finite set: tabular storage of $V(s)$, $s \in \{1, ..., |\mathcal{S}|\}$
  $\rightarrow$ does not fit in memory if $|\mathcal{S}|$ is too large ⚠
- If $\mathcal{S}$ is a continuous set: parameterization V_θ, $\theta \in \mathbb{R}^p$
  $\rightarrow$ curse of dimensionality if $\dim(\mathcal{S})$ is large ⚠

Solution: exploit some regularity or structure on V.

Tools used in our work:

- Max-plus linear parameterization
- Non-parametric representations in an RKHS

Contributions

- E. B. and F. Bach, "Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes," in *IEEE Control Systems Letters*, July 2020.

- E. B., J. Carpentier and F. Bach, "Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems," *European Control Conference (ECC)*, July 2021.

- E. B., J. Carpentier, A. Rudi and F. Bach, "Infinite-dimensional Sums-of-Squares for Optimal Control," *Conference on Decision and Control (CDC)*, Dec. 2022.

- E. B., Z. Kobeissi and F. Bach, "A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022.

Contributions

- E. B. and F. Bach, "Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes," in *IEEE Control Systems Letters*, July 2020. [model known]
- E. B., J. Carpentier and F. Bach, "Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems," *European Control Conference (ECC)*, July 2021. [model in a robust class]
- E. B., J. Carpentier, A. Rudi and F. Bach, "Infinite-dimensional Sums-of-Squares for Optimal Control," *Conference on Decision and Control (CDC)*, Dec. 2022. [batch of observations]
- E. B., Z. Kobeissi and F. Bach, "A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022. [online observations]

Contributions

- E. B. and F. Bach, "Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes," in *IEEE Control Systems Letters*, July 2020. [model known] [V max-plus linear]

- E. B., J. Carpentier and F. Bach, "Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems," *European Control Conference (ECC)*, July 2021. [model in a robust class]

- E. B., J. Carpentier, A. Rudi and F. Bach, "Infinite-dimensional Sums-of-Squares for Optimal Control," *Conference on Decision and Control (CDC)*, Dec. 2022. [batch of observations] [$H \geq 0$ with an RKHS]

- E. B., Z. Kobeissi and F. Bach, "A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022. [online observations] [V in an RKHS]

Contributions

- E. B. and F. Bach, "Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes," in *IEEE Control Systems Letters*, July 2020. [model known] [V max-plus linear]

- E. B., J. Carpentier and F. Bach, "Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems," *European Control Conference (ECC)*, July 2021. [model in a robust class]

- E. B., J. Carpentier, A. Rudi and F. Bach, "Infinite-dimensional Sums-of-Squares for Optimal Control," *Conference on Decision and Control (CDC)*, Dec. 2022. [batch of observations] [$H \geq 0$ with an RKHS]

- E. B., Z. Kobeissi and F. Bach, "A Non-asymptotic Analysis of Non-parametric Temporal-Difference Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2022. [online observations] [V in an RKHS]

Contents

1 Introduction

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

State-discretization of an MDP

Consider a deterministic MDP defined by:

- a continuous state space $\mathcal{S} \subset \mathbb{R}^d$,
- a discrete action space $\mathcal{A}$,
- a bounded reward function $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$,
- a dynamics $\varphi_{\cdot}(.) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$.

We want to discretize it into a finite MDP, to run value iteration.

State-discretization of an MDP

Consider a deterministic MDP defined by:

- a continuous state space $\mathcal{S} \subset \mathbb{R}^d$,
- a discrete action space $\mathcal{A}$,
- a bounded reward function $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$,
- a dynamics $\varphi_.(.) : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$.

We want to discretize it into a finite MDP, to run value iteration.

Problem: A naive discretization requires a very tight state-discretization to capture the dynamics, whose size blows up with the dimension.

$\rightarrow$ *Can we build a better discretization?*

Max-plus linear approximation

The max-plus semiring is defined as $(\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$, where $\oplus$ represents the maximum operator, and $\otimes$ represents the usual sum.

Let $W = (w_1, ..., w_k)$ be a dictionary of functions $w_i : \mathcal{S} \to \mathbb{R}$.

For $\alpha \in \mathbb{R}^k$, we define the max-plus linear combination [Fleming and McEneaney, 2000]:

$$V(s) = \bigoplus_{i=1}^{k} \alpha_i \otimes w_i(s) = \max_{1 \le i \le k} \alpha_i + w_i(s) \, .$$

Dictionaries for discretization

Piecewise constant value functions are natural candidates for a discretization, suggesting the following dictionaries:

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

Dictionaries for discretization

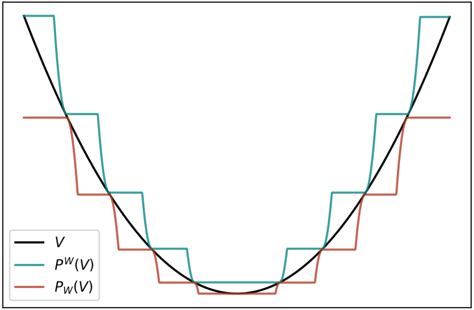
Piecewise constant value functions are natural candidates for a discretization, suggesting the following dictionaries:

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

- Soft indicator: $w(s) = -c \,\mathrm{dist}(s, A)^2$, with c large.

Max-plus projection

A function $V \in \mathbb{R}^{\mathcal{S}}$ can be lower- (or upper-) projected onto W.



Max-plus projection

A function $V \in \mathbb{R}^{\mathcal{S}}$ can be lower- (or upper-) projected onto W.

Proposition ([Berthier and Bach, 2020])

Let $(A_1, \ldots, A_k)$ *a partition of* $\mathcal{S}$ *where each* A_i *is convex, compact and non-empty, and let* $D = \max_{1 \leq i \leq k} \mathrm{diam}(A_i)$.
Let $W = (w_1, \ldots, w_k)$ *defined by:*

$$w_i(s) = -c \; \mathrm{dist}(s, A_i)^2$$

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

$$\|V - P_W(V)\|_\infty \leq 2LD$$

Max-plus projection

A function $V \in \mathbb{R}^{\mathcal{S}}$ can be lower- (or upper-) projected onto W.

Proposition ([Berthier and Bach, 2020])

Let $(A_1, \ldots, A_k)$ a partition of $\mathcal{S}$ where each A_i is convex, compact and non-empty, and let $D = \max_{1 \le i \le k} \operatorname{diam}(A_i)$.
Let $W = (w_1, \ldots, w_k)$ defined by:

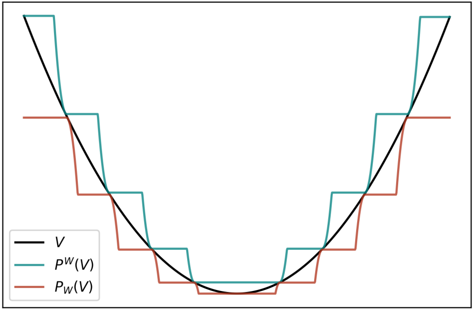
$$w_i(s) = -c \operatorname{dist}(s, A_i)^2$$

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

$$\|V - P_W(V)\|_\infty \leq \boxed{2LD} \leftarrow \text{independent of } c$$

Max-plus projection

A function $V \in \mathbb{R}^{\mathcal{S}}$ can be lower- (or upper-) projected onto W.



Can we compute $P_W(V^*)$ without knowing V^*?

Approximate value iteration

We follow the method of [Akian et al., 2008]. Using the max-plus linearity of the Bellman operator, it decouples into two steps:

1. k^2 precomputations of the form:

$$K_{ij} = \sup_{s \in \mathcal{S},\ a \in \mathcal{A}} w_i(s) + r(s, a) + \gamma w_j(\varphi_a(s))\,.$$

2. A reduced value iteration algorithm on a finite MDP with k states and k actions, which uses the K_{ij}.

Approximate value iteration

We follow the method of [Akian et al., 2008]. Using the max-plus linearity of the Bellman operator, it decouples into two steps:

1. k^2 precomputations of the form:

$$\boxed{K_{ij} = \sup_{s \in \mathcal{S},\ a \in \mathcal{A}} w_i(s) + r(s,a) + \gamma w_j(\varphi_a(s))\,.}$$

2. A reduced value iteration algorithm on a finite MDP with k states and k actions, which uses the K_{ij}.

Approximate precomputations

$$K_{ij} = \max_{a \in \mathcal{A}} \sup_{s \in \mathcal{S}} w_i(s) + r(s, a) + \gamma w_j(\varphi_a(s))\,.$$

Approximate precomputations

$$K_{ij} = \max_{a \in \mathcal{A}} \underbrace{\sup_{s \in \mathcal{S}} \; w_i(s) + r(s, a) + \gamma w_j(\varphi_a(s))}_{\text{gradient ascent on } s \ (\simeq \text{ concave}) \ \to \ \hat{K}_{ij}} \; .$$

Approximate precomputations

$$K_{ij} = \max_{a \in \mathcal{A}} \underbrace{\sup_{s \in \mathcal{S}} \; w_i(s) + r(s,a) + \gamma w_j(\varphi_a(s))}_{\text{gradient ascent on } s \ (\simeq \text{ concave}) \ \to \ \hat{K}_{ij}} \; .$$

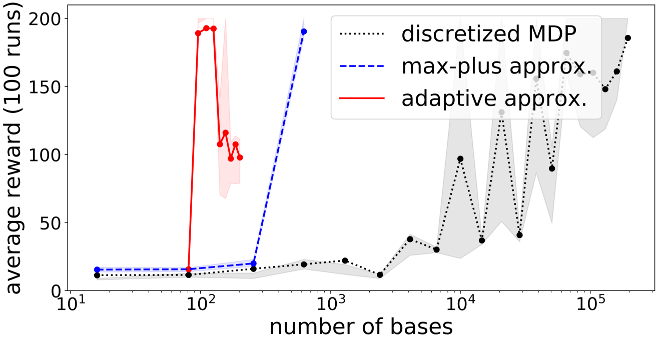
Decomposition of errors:

Theorem ([Berthier and Bach, 2020])

Let V be the result of the reduced value iteration step. Then:

$$\|V - V^*\|_\infty \leq \frac{1}{1-\gamma}\Big(\|P_W(V^*) - V^*\|_\infty + \|P^W(V^*) - V^*\|_\infty + \|\hat{K} - K\|_\infty \Big).$$

Experiment (Cartpole, $d = 4$)



Contents

1 Introduction

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

Sample-based optimal control

We want to solve the optimal control problem:

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

without knowing f and L.

Sample-based optimal control

We want to solve the optimal control problem:

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

without knowing f and L.

We only observe samples:

$$f(t^{(i)}, x^{(i)}, u^{(i)}), \quad L(t^{(i)}, x^{(i)}, u^{(i)}),$$
$$\text{for } i \in \{1, ..., n\} = I.$$

Weak-formulation of optimal control

The optimal control problem:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t))\mathrm{d}t + M(x(T))$$

$$\forall t \in [t_0, T],\ \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0.$$

Weak-formulation of optimal control

The optimal control problem:

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

is equivalent (under convexity assumptions) to finding a maximal subsolution of the HJB equation [Lasserre et al., 2010]:

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

Weak-formulation of optimal control

The optimal control problem:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t)) \mathrm{d}t + M(x(T))$$

$$\forall t \in [t_0, T], \; \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0.$$

is equivalent (under convexity assumptions) to finding a maximal subsolution of the HJB equation [Lasserre et al., 2010]:

$$\sup_{V \in C^1([0,T] \times \mathcal{X})} V(0, x_0)$$

$$\forall (t, x, u), \boxed{\frac{\partial V}{\partial t}(t, x) + L(t, x, u) + \nabla V(t, x)^\top f(t, x, u) \geq 0}$$

$$\forall x, \; V(T, x) \leq M(x).$$

$$H(t, x, u) \geq 0$$

A simple baseline: linear programming

Using a linear parameterization of V, and simply subsampling inequalities leads to an LP:

$$\sup_{\theta \in \mathbb{R}^m} \quad V_\theta(0, x_0)$$
$$\forall i \in I, \quad H_\theta(t^{(i)}, x^{(i)}, u^{(i)}) \geq 0.$$

This readily gives a first numerical method.

A simple baseline: linear programming

Using a linear parameterization of V, and simply subsampling inequalities leads to an LP:

$$\sup_{\theta \in \mathbb{R}^m} \quad V_\theta(0, x_0)$$
$$\forall i \in I, \quad H_\theta(t^{(i)}, x^{(i)}, u^{(i)}) \geq 0.$$

This readily gives a first numerical method.

Can we do any better?

SoS representation of non-negative functions

$$\sup_{\theta \in \mathbb{R}^m} V_\theta(0, x_0)$$

$$\forall (t, x, u), \boxed{H_\theta(t, x, u) \geq 0.}$$

SoS representation of non-negative functions

$$\sup_{\theta \in \mathbb{R}^m} V_\theta(0, x_0)$$
$$\forall (t, x, u), \boxed{H_\theta(t, x, u) \geq 0.}$$

If we represent some g_k of the form:

$$g_k(y) = \langle \alpha_k, \varphi(y) \rangle.$$

Then we can generate a non-negative function as a sum-of-squares:

$$g(y) = \sum_{k=1}^{m} g_k(y)^2$$

SoS representation of non-negative functions

$$\sup_{\theta\in\mathbb{R}^m} V_\theta(0,x_0)$$

$$\forall(t,x,u),\ \boxed{H_\theta(t,x,u)\geq 0.}$$

If we represent some g_k of the form:

$$g_k(y)=\langle\alpha_k,\varphi(y)\rangle.$$

Then we can generate a non-negative function as a sum-of-squares:

$$g(y)=\sum_{k=1}^m g_k(y)^2=\boxed{\langle\varphi(y),A\varphi(y)\rangle.}$$

where $A=\sum_{k=1}^m \alpha_k\otimes\alpha_k\succeq 0.$

SoS representation of the Hamiltonian

Theorem ([Berthier, Carpentier, Rudi and Bach, 2022])

Assume that:

- f *is control-affine:* $f(t,x,u) = g(t,x) + B(t,x)u$;
- L *is strongly convex in* u;
- L, B *and* V^* *are sufficiently smooth;*

Then H^* *is a SoS of* p *smooth functions* $(w_j)_{1\leq j\leq p} \in C^s(\Omega)$*:*

$$\forall (t,x,u) \in \Omega, \quad H^*(t,x,u) = \sum_{j=1}^{p} w_j(t,x,u)^2.$$

SoS representation of the Hamiltonian

Theorem ([Berthier, Carpentier, Rudi and Bach, 2022])

Assume that:

- *f is control-affine:* $f(t,x,u) = g(t,x) + B(t,x)u$;
- *L is strongly convex in u;*
- *L, B and V^* are sufficiently smooth;*

Then H^* is a SoS of p smooth functions $(w_j)_{1 \leq j \leq p} \in C^s(\Omega)$:

$$\forall (t,x,u) \in \Omega, \quad H^*(t,x,u) = \sum_{j=1}^{p} w_j(t,x,u)^2.$$

⚠ In general V^* is not even C^1.

An algorithm for smooth optimal control

$$\sup_{V \in C^1([0,T]\times\mathcal{X})} \quad V(0, x_0)$$
$$\forall (t,x,u),\ H(t,x,u) \geq 0$$
$$\forall x,\ V(T,x) \leq M(x)$$

Steps:

An algorithm for smooth optimal control

$$\sup_{\theta \in \mathbb{R}^m} V_\theta(0, x_0)$$
$$\forall (t, x, u),\ H_\theta(t, x, u) \geq 0$$

Steps:

- linear parameterization of V

An algorithm for smooth optimal control

$$\sup_{\theta \in \mathbb{R}^m,\ \mathcal{A} \in \mathbb{S}_+(\mathcal{H})} V_\theta(0, x_0)$$

$$\forall (t, x, u),\ \ H_\theta(t, x, u) = \langle \varphi(t, x, u), \mathcal{A}\varphi(t, x, u) \rangle$$

Steps:

- linear parameterization of V
- SoS representation of the Hamiltonian

An algorithm for smooth optimal control

$$\sup_{\theta \in \mathbb{R}^m,\ \mathcal{A} \in \mathbb{S}_+(\mathcal{H})} V_\theta(0, x_0) - \lambda \mathrm{Tr}(\mathcal{A})$$

$$\forall i,\ \ H_\theta(t^{(i)}, x^{(i)}, u^{(i)}) = \langle \varphi(t^{(i)}, x^{(i)}, u^{(i)}), \mathcal{A}\varphi(t^{(i)}, x^{(i)}, u^{(i)}) \rangle$$

Steps:

- linear parameterization of V
- SoS representation of the Hamiltonian
- subsampling equalities

An algorithm for smooth optimal control

$$\sup_{\theta \in \mathbb{R}^m,\ B \succeq 0} \quad V_\theta(0, x_0) - \lambda \mathrm{Tr}(B)$$

$$\forall i,\ H_\theta(t^{(i)}, x^{(i)}, u^{(i)}) = \Phi_i^\top B \Phi_i$$

Steps:

- linear parameterization of V
- SoS representation of the Hamiltonian
- subsampling equalities
- kernel trick

An algorithm for smooth optimal control

$$\sup_{\theta \in \mathbb{R}^m,\ B \succeq 0} V_\theta(0, x_0) - \lambda \mathrm{Tr}(B)$$
$$\forall i,\ H_\theta(t^{(i)}, x^{(i)}, u^{(i)}) = \Phi_i^\top B \Phi_i$$

Steps:

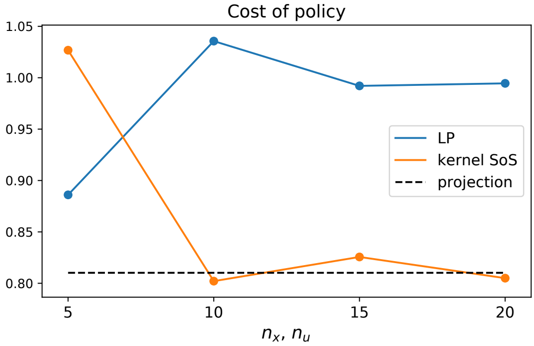
- linear parameterization of V
- SoS representation of the Hamiltonian
- subsampling equalities
- kernel trick

$\rightarrow$ This is an SDP of size $n \times n$.

Sample-based version of the method of [Lasserre et al., 2010].

Numerical example

On a simple linear quadratic regulator:



Contents

1 Introduction

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

Policy evaluation

Given a fixed policy π, we want to evaluate:

$$V^*(x) = \mathbb{E}_\pi \left[\sum_{n=0}^{+\infty} \gamma^n r(x_n) \middle| x_0 = x \right],$$

without knowing $r \in L^2$ nor the transition probabilities.

Policy evaluation

Given a fixed policy π, we want to evaluate:

$$V^*(x) = \mathbb{E}_\pi \left[\sum_{n=0}^{+\infty} \gamma^n r(x_n) \middle| x_0 = x \right],$$

without knowing $r \in L^2$ nor the transition probabilities.

We only observe samples of transitions from the Markov chain:

$$(x_k, r(x_k), x'_k)_{1 \leq k \leq n}$$

TD(0) with linear function approximation

Linear approximation of the value function:

$$V^*(x) \simeq \xi^\top \varphi(x), \text{ for some } \xi \in \mathbb{R}^p.$$

TD(0): sample a transition $(x_n, r(x_n), x_n')$ and update:

$$\xi_n = \xi_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x_n') - V_{n-1}(x_n) \right] \varphi(x_n),$$

TD(0) with linear function approximation

Linear approximation of the value function:

$$V^*(x) \simeq \xi^\top \varphi(x), \text{ for some } \xi \in \mathbb{R}^p.$$

TD(0): sample a transition $(x_n, r(x_n), x'_n)$ and update:

$$\xi_n = \xi_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x'_n) - V_{n-1}(x_n) \right] \varphi(x_n),$$

Converges under classical assumptions for stochastic approximation,
⚠ to something different from V^* if $V^* \notin \text{span}(\varphi_1, ..., \varphi_p)$.
[Tsitsiklis and Van Roy, 1997], [Bhandari et al., 2018]

TD(0) with linear function approximation

Linear approximation of the value function:

$$V^*(x) \simeq \xi^\top \varphi(x), \text{ for some } \xi \in \mathbb{R}^p.$$

TD(0): sample a transition $(x_n, r(x_n), x_n')$ and update:

$$\xi_n = \xi_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x_n') - V_{n-1}(x_n) \right] \varphi(x_n),$$

Converges under classical assumptions for stochastic approximation,
⚠ to something different from V^* if $V^* \notin \mathrm{span}(\varphi_1, \dots, \varphi_p)$.
[Tsitsiklis and Van Roy, 1997], [Bhandari et al., 2018]

Can we fix this with a universal approximator?

TD(0) with linear function approximation

Linear approximation of the value function:

$$V^*(x) \simeq \xi^\top \varphi(x), \text{ for some } \xi \in \mathbb{R}^p.$$

TD(0): sample a transition $(x_n, r(x_n), x_n')$ and update:

$$\xi_n = \xi_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x_n') - V_{n-1}(x_n) \right] \varphi(x_n),$$

Converges under classical assumptions for stochastic approximation,
⚠ to something different from V^* if $V^* \notin \text{span}(\varphi_1, \ldots, \varphi_p, \ldots)$.
[Tsitsiklis and Van Roy, 1997], [Bhandari et al., 2018]

Can we fix this with a universal approximator?

Non-parametric TD(0)

Sample a transition $(x_n, r(x_n), x_n')$ and update:

$$V_n = V_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x_n') - V_{n-1}(x_n) \right] K(x_n, \cdot),$$

where K is the reproducing kernel of an RKHS $\mathcal{H} \subset L^2$.

Non-parametric TD(0)

Sample a transition $(x_n, r(x_n), x'_n)$ and update:

$$V_n = V_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x'_n) - V_{n-1}(x_n) \right] K(x_n, \cdot),$$

where K is the reproducing kernel of an RKHS $\mathcal{H} \subset L^2$.

- the iterates are in $\mathcal{H}$ (functional space)

Non-parametric TD(0)

Sample a transition $(x_n, r(x_n), x'_n)$ and update:

$$V_n = V_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x'_n) - V_{n-1}(x_n) \right] K(x_n, \cdot),$$

where K is the reproducing kernel of an RKHS $\mathcal{H} \subset L^2$.

- the iterates are in $\mathcal{H}$ (functional space)
- recovers linear approximation with $K(x, y) = \varphi(x)^\top \varphi(y)$

Non-parametric TD(0)

Sample a transition $(x_n, r(x_n), x'_n)$ and update:

$$V_n = V_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x'_n) - V_{n-1}(x_n) \right] K(x_n, \cdot),$$

where K is the reproducing kernel of an RKHS $\mathcal{H} \subset L^2$.

- the iterates are in $\mathcal{H}$ (functional space)
- recovers linear approximation with $K(x, y) = \varphi(x)^\top \varphi(y)$
- universal kernel such that $\overline{\mathcal{H}} = L^2$ (*e.g.*, Sobolev kernel)

$\rightarrow$ convergence to V^* in L^2-norm, even if $V^* \notin \mathcal{H}$.

Non-parametric TD(0)

Sample a transition $(x_n, r(x_n), x_n')$ and update:

$$V_n = V_{n-1} + \rho_n \left[r(x_n) + \gamma V_{n-1}(x_n') - V_{n-1}(x_n) \right] K(x_n, \cdot),$$

where K is the reproducing kernel of an RKHS $\mathcal{H} \subset L^2$.

- the iterates are in $\mathcal{H}$ (functional space)
- recovers linear approximation with $K(x, y) = \varphi(x)^\top \varphi(y)$
- universal kernel such that $\overline{\mathcal{H}} = L^2$ (*e.g.*, Sobolev kernel)

$\rightarrow$ convergence to V^* in L^2-norm, even if $V^* \notin \mathcal{H}$.

Let us define the covariance operator [De Vito et al., 2005]:

$$\Sigma = \mathbb{E}[K(x, \cdot) \otimes K(x, \cdot)].$$

Main convergence result

Theorem ([Berthier, Kobeissi and Bach, 2022])

Assume that for some $\theta \in (-1, 1]$*:*

$$\|\Sigma^{-\theta/2} V^*\|_{\mathcal{H}} < +\infty \,. \qquad \text{(source condition)}$$

Then with suitable regularization, step size and averaging scheme:

$$\mathbb{E}\left[\|\overline{V}_n - V^*\|_{L^2}^2\right] = O\left((\log n)^2 n^{-\frac{1+\theta}{2+\theta}}\right) .$$

Main convergence result

Theorem ([Berthier, Kobeissi and Bach, 2022])

Assume that for some $\theta \in (-1, 1]$*:*

$$\|\Sigma^{-\theta/2} V^*\|_{\mathcal{H}} < +\infty \,. \qquad \text{(source condition)}$$

Then with suitable regularization, step size and averaging scheme:

$$\mathbb{E}\left[\|\overline{V}_n - V^*\|_{L^2}^2\right] = O\left((\log n)^2 n^{-\frac{1+\theta}{2+\theta}}\right).$$

- $\theta = 0$: $V^* \in \mathcal{H}$ recovers known $1/\sqrt{n}$ parametric rate.
- $\theta \in (0, 1]$: stronger assumption, faster rate.
- $\theta = -1$: $V^* \in L^2$, only asymptotic convergence.
- $\theta \in (-1, 0)$: $V^* \notin \mathcal{H}$, weaker assumption, slower rate.

Main convergence result

Theorem ([Berthier, Kobeissi and Bach, 2022])

Assume that for some $\theta \in (-1, 1]$:

$$\|\Sigma^{-\theta/2} V^*\|_{\mathcal{H}} < +\infty \,. \qquad \text{(source condition)}$$

Then with suitable regularization, step size and averaging scheme:

$$\mathbb{E}\left[\|\overline{V}_n - V^*\|_{L^2}^2\right] = O\left((\log n)^2 n^{-\frac{1+\theta}{2+\theta}}\right) .$$

- Theorem proved in the *i.i.d.* sampling setting.
- Extends to sampling from a Markov chain with exponential mixing, with an additional boundedness assumption.
- Results are similar to SGD ($\gamma = 0$) [Dieuleveut and Bach, 2016].

Sketch of the proof

1. The ODE method: study the average update in continuous-time

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = \mathbb{E}\Big[(r(x) + \gamma V_t(x') - V_t(x))K(x,\cdot)\Big]$$

Sketch of the proof

1. The ODE method: study the average update in continuous-time
2. Prove the stability of the ODE with a Lyapunov function

$$\frac{\mathrm{d}}{\mathrm{d}t}\|V_t - V^*\|_{\mathcal{H}}^2 < 0.$$

Sketch of the proof

1. The ODE method: study the average update in continuous-time
2. Prove the stability of the ODE with a Lyapunov function

$$\frac{\mathrm{d}}{\mathrm{d}t}\|V_t - V^*\|_{\mathcal{H}}^2 < 0.$$

With Polyak-Ruppert averaging:

$$\|\overline{V_t} - V^*\|_{L^2}^2 \leq \frac{1}{2(1-\gamma)} \frac{\|V^*\|_{\mathcal{H}}^2}{t}.$$

Sketch of the proof

1. The ODE method: study the average update in continuous-time
2. Prove the stability of the ODE with a Lyapunov function

$$\frac{\mathrm{d}}{\mathrm{d}t}\|V_t - V^*\|_{\mathcal{H}}^2 < 0.$$

With Polyak-Ruppert averaging:

$$\|\overline{V_t} - V^*\|_{L^2}^2 \leq \frac{1}{2(1-\gamma)} \frac{\boxed{\|V^*\|_{\mathcal{H}}^2}}{t}.$$

Sketch of the proof

1. The ODE method: study the average update in continuous-time
2. Prove the stability of the ODE with a Lyapunov function
3. If $V^* \notin \mathcal{H}$, add an extra regularization

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = \mathbb{E}\Big[(r(x) + \gamma V_t(x') - V_t(x))K(x,\cdot)\Big] - \lambda V_t$$

Sketch of the proof

1. The ODE method: study the average update in continuous-time
2. Prove the stability of the ODE with a Lyapunov function
3. If $V^* \notin \mathcal{H}$, add an extra regularization

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = \mathbb{E}\Big[\big(r(x) + \gamma V_t(x') - V_t(x)\big)K(x,\cdot)\Big] - \lambda V_t$$

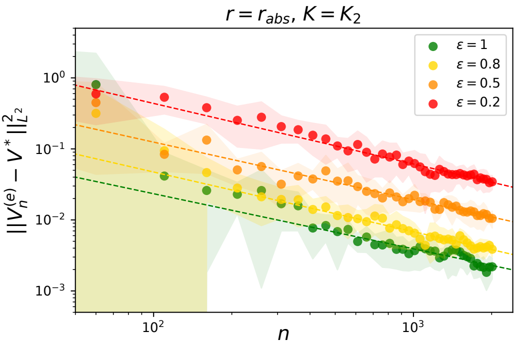
$$\begin{tikzcd} V_t \arrow[r, "t\to+\infty"'] & V^*_\lambda \arrow[d, "\lambda\to 0"'] \\ & V^* \end{tikzcd}$$

$\rightarrow$ tradeoff in the choice of λ, depending on θ.

Numerical experiment

Sobolev kernel of regularity s on the 1d torus.
Source condition θ: decrease of Fourier coefficients of V^*:

$$|\hat{V}_0^*|^2 + \sum_{\omega \neq 0} |\omega|^{2s(1+\theta)} |\hat{V}_\omega^*|^2 < \infty.$$



Contents

1 Introduction

2 Max-Plus Discretization of Deterministic MDPs

3 Infinite-Dimensional Sums-of-Squares for Optimal Control

4 Convergence of Non-parametric Temporal-Difference Learning

5 Conclusion & Perspectives

Summary of the contributions

1. A max-plus approximation scheme applied to the discretization of deterministic MDPs.
2. A method for estimating stability regions on robust classes of dynamical systems.
3. A sample-based algorithm for optimal control problems, based on a SoS representation of non-negative functions.
4. Convergence rates for non-parametric TD learning.

Perspectives

Control problems from a machine learning viewpoint:

- **approximation** – model of the value function? the Hamiltonian?
- **estimation** – sample complexities? stochastic approximation?
- **optimization** – primal-dual formulation? link with SGD?

Thank you for your attention!

References (1)

M. Akian, S. Gaubert, and A. 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.

J. Bhandari, D. Russo, and R. Singal.
A finite time analysis of temporal difference learning with linear function approximation.
In *Conference on Learning Theory*, pages 1691–1692, 2018.

E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, and P. Bartlett.
Learning from examples as an inverse problem.
Journal of Machine Learning Research, 6(5), 2005.

A. Dieuleveut and F. Bach.
Nonparametric stochastic approximation with large step-sizes.
The Annals of Statistics, 44(4):1363–1399, 2016.

References (2)

W. H. Fleming and W. M. McEneaney.
A max-plus-based algorithm for a Hamilton–Jacobi–Bellman equation of nonlinear filtering.
SIAM Journal on Control and Optimization, 38(3):683–710, 2000.

J.-B. Lasserre, D. Henrion, C. Prieur, and E. Trélat.
Nonlinear optimal control via occupation measures and LMI-relaxations.
SIAM Journal on Control and Optimization, 47(4):1643–1666, 2008.

D. Liberzon.
Calculus of Variations and Optimal Control Theory: A Concise Introduction.
Princeton University Press, 2011.

E. Novak.
Deterministic and Stochastic Error Bounds in Numerical Analysis.
Springer, 2006.

References (3)

A. Rudi, U. Marteau-Ferey, and F. Bach.
Finding global minima via kernel approximations.
Technical Report 2012.11978, arXiv, 2020.

R. S. Sutton and A. G. Barto.
Reinforcement Learning: An Introduction.
MIT Press, 2018.

E. Trélat.
Contrôle Optimal: Théorie & Applications, volume 36.
Vuibert Paris, 2005.

J. N. Tsitsiklis and B. Van Roy.
An analysis of temporal-difference learning with function approximation.
IEEE Transactions on Automatic Control, 42(5):674–690, 1997.