Quantifying Uncertainty: Stochastic, Adversarial, and Beyond
Sep 12—16, 2022

SIMONS INSTITUTE
for the Theory of Computing

Function approximation and large-scale MDP planning

Csaba Szepesvári
DeepMind & University of Alberta

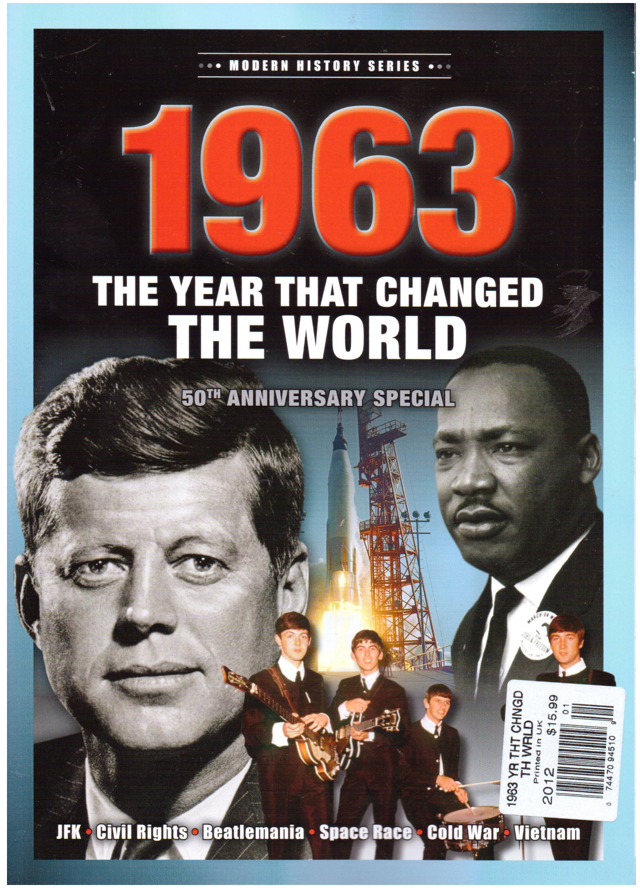
Thanks to..

- **Gellért Weisz**

- Philip Amortilla
- Barnabás Janzer
- Nan Jiang
- Yasin Abbasi-Yadkori
- András György

- Foundations team and more @ DeepMind
- AMII, RLAI @ UofA

Gellért discovering that this wall is unbounded from above..



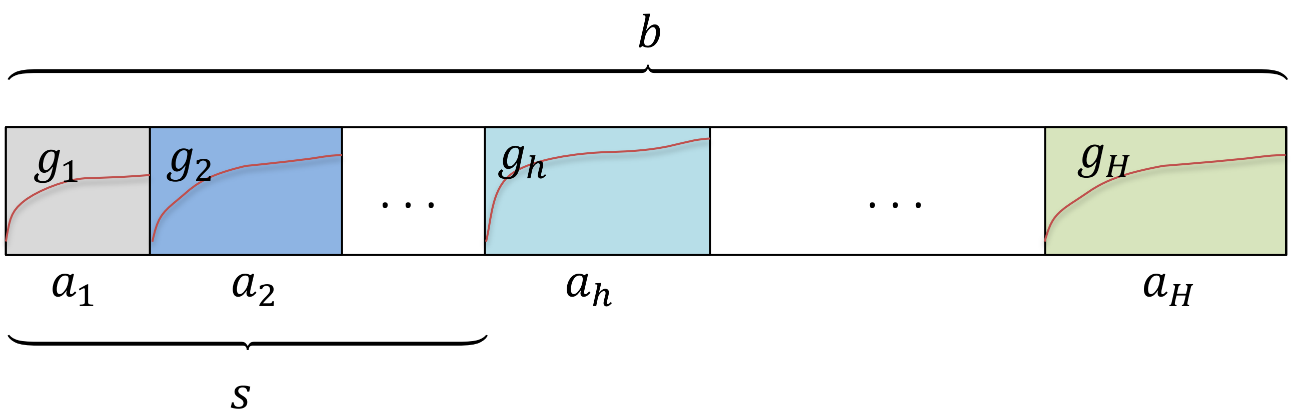
Polynomial Approximation—A New Computational Technique in Dynamic Programming: Allocation Processes

Math.Comput.
17:155—161
1963

By Richard Bellman, Robert Kalaba, and Bella Kotkin

Resource allocation with nonlinear utilities to H projects

Given $g_1, \dots, g_H \colon [0,b]^D \to \mathbb{R}$, find $v^* = \max_{a_1+\cdots+a_H \le b\mathbf{1}, a_i \ge 0} g_1(a_1) + \cdots + g_H(a_H)$



$v_h^*(s)$: optimal value achievable over $[h, H]$ if resource used before this stage is $s \in [0,b]^D$

$$v_h^*(s) = \max_{0 \le a \le (b-s)\mathbf{1}} g_h(a) + v_{h+1}^*(s+a) \qquad 1 \le h \le H-1$$

$$v_H^*(s) = g_H(b-s) \text{ [say, } g_h \text{ is increasing]}$$

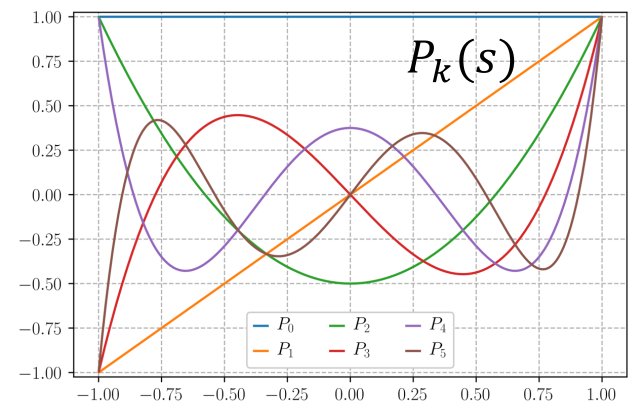
How to compute $v_1^*(0)$? .. and the optimal "policy" ($a_h^*(s) = ?$)

"Represent" v_h^* somehow.. Discretization? Bad $\Omega(2^D)$ scaling when $a \in [0,b]^D$

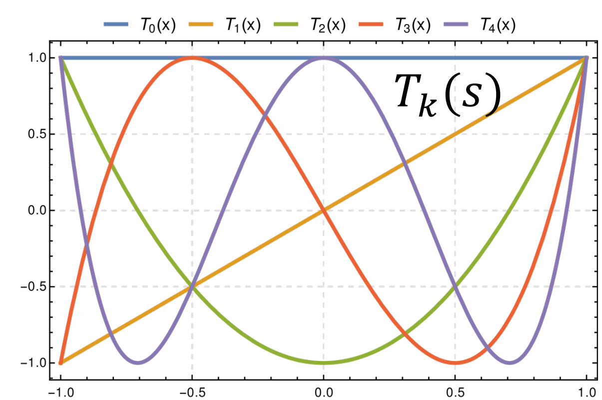
New idea (in 1963): Generalized polynomial approximation

$$f(s) = \Sigma_{k=1}^{d} \theta_k \phi_k(s), \qquad s \in [-1,1]$$

$\phi_k(s) = s^{k-1}$, or $\cos((k-1)s)$, or
$P_k(s)$, or $T_k(s)$



$\{P_k\}$ (or $\{T_k\}$) $\Rightarrow$ orthonormal set w.r.t. uniform measure on [-1,1]



$$\theta_k = \int_{-1}^{1} f \phi_k$$

https://en.wikipedia.org/wiki/Legendre_polynomials

https://en.wikipedia.org/wiki/Chebyshev_polynomials

$$(*)\ v_h^*(s) = \underbrace{\max_{0\le a\le (b-s)\mathbf{1}} g_h(a) + v_{h+1}^*(s+a)}_{(Tv_{h+1}^*)(s)},\quad 1 \le h \le H-1$$

Successive approximations

Idea: $v_h^* \approx v_{\theta_h} := \Phi\theta_h$ for some $\theta_h \in \mathbb{R}^d$ for all h.

Getting θ_h from θ_{h+1}:

$$v_{\theta_h} = \Pi_{span(\Phi)}(T\, v_{\theta_{h+1}})$$

Fitted value iteration

BKK63 used an ONB and Gaussian quadratures for approximating the projection

Results

Benchmarks! 2 dimensional problems! Good results!

"Finally, if we combine these techniques – polynomial approximations and Lagrange multipliers – with that of successive approximations, there should be very few allocation processes which still resist our efforts."

(Lagrange multipliers: Because actions may be constrained)

Why the optimism?
No discretization of the state space, just need to guess Φ
➔ no "curse of dimensionality" if guess is correct. Yes?

Questions

1. **Approximation**: How large should be the degree of polynomials used to approximate v^*? How to choose the basis functions?

   Smoothness, approximation theory, systems theory.. Someone else's problem☺

2. **Computation**: 🡸 FOCUS
   Given that we can approximate well v^*, say,

   $v^*(x) = \Sigma_{i=1}^{d} \theta_i^* \phi_i(x)$,

   how much computation is needed to get $\theta^* = (\theta_1^*, \dots, \theta_d^*)$? How many <u>queries</u>?

   Can we do it in $\mathrm{poly}(A, H, d, 1/\varepsilon)$ regardless of dimension (state space size)?

Contents

1. ~~The origins (1963)~~
2. ~~The questions~~
3. Optimistic Constraint Propagation (2013)
4. Misspecification (Du-Kakade-Wang-Yang 2021) under strong FA
5. Weak FA results
6. Future

MDPs and Bellman equations

$$v_h^*(s) = \max_{a \in \mathcal{A}(s)} \underbrace{r(s,a) + \mathbb{E}_\xi[v_{h+1}^*(f(s,a,\xi))]}_{q_h^*(s,a)}$$

$s \in \mathcal{S}$ -- States

$r(s,a)$ -- Rewards and $r(s,a) = r_a(s)$

$f(s,a,\xi)$ -- Stochastic transitions to a next state,

$s' \sim P_a(s)$. $\mathbb{E}_\xi\left[v\left(f(s,a,\xi)\right)\right] = \langle P_a(s), v \rangle$

$\mathcal{A}(s)$ -- Admissible actions

For simplicity, $\mathcal{A}(s) = \mathcal{A}$

Optimistic Constraint Propagation

Deterministic MDPs

Zheng Wen
Ben van Roy
2013

$$q_h^*(s,a) = \varphi_h(s,a)^\top \theta^* =: q_h(s,a;\theta^*)$$

$$\mathrm{TD}_h(s,a,s',\theta) := r_h(s,a) + \max_{a'} q_{h+1}(s',a';\theta) - q_h(s,a;\theta)$$

(*) $\quad \mathrm{TD}_h(s,a,s',\theta^*) = 0 \qquad \forall h, s, a, s' = f_h(s,a)$

Start with $\Theta_0 = \{\, \theta : \|\theta\|_1 \le B \,\}$

Iteration $i = 0,1,\ldots$:

Pick any $\theta \in \Theta_i$ s.t. $\max_a q_1(s_0,a;\theta)$ is maximized over Θ_i (ASK ME)

Roll out with $\pi_h(s) = \mathrm{argmax}_a q_h(s,a;\theta) \to ((s_h,a_h)_h)$

$\Theta_{i+1} = \{\theta \in \Theta_i \;:\; \mathrm{TD}_h(s_h,a_h,s_{h+1},\theta) = 0 \;\; \forall h\,\}$

Return $\pi_h(s_0)$ if $\Theta_{i+1} = \Theta_i$

Sample Complexity

Theorem [WR13]:

For any deterministic system, the previous algorithm stops after

$$\mathrm{poly}(B, d, H, A)$$

interactions with the system and returns an optimal action at s_0.

Further, the total computation effort is also poly in the same quantities.

8 years later..

- Du-Kakade-Wang-Yang 2021, Lattimore-Szepesvari-Weisz 2021

- Setup: $Q(\Pi) \subset \mathcal{F} \bigoplus [-\varepsilon, \varepsilon]^d$
- Result: the query complexity to get a δ −optimal action at s_0 is exponential in $\min(H, d)$ unless $\delta \geq \sqrt{d}\varepsilon$

- For $\delta \geq \sqrt{d}\,\varepsilon$, fitted policy iteration under global access returns with δ −optimal action in poly time

- Insight: Extrapolation based on finite data unavoidably inflates best approximation error
  Note! $\sqrt{d}$ is the maximum blowup. Blowup may not happen

Strong ⇒ Weak function approximation

- Strong function approximation:
$$T_*\mathcal{F} \subset \mathcal{F}$$
$$\text{or}$$
$$Q(\Pi) \subset \mathcal{F}$$

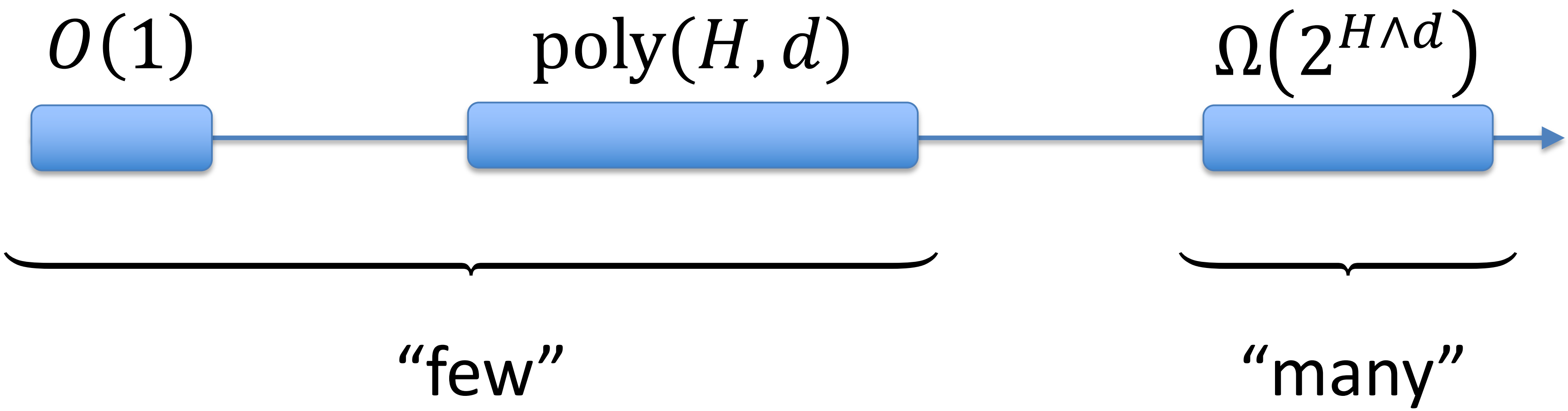
- Weak function approximation:
$$v^* \in \mathcal{F}$$
(or $q^* \in \mathcal{F}$).

- Why weak? The approximation space is not large enough to hold all kind of functions, just the optimal value function
- More ambitious
- But no misspecification

Contents

1. ~~The origins (1963)~~
2. ~~The questions~~
3. ~~Optimistic Constraint Propagation (2013)~~
4. ~~Misspecification (Du-Kakade-Wang-Yang 2021) under strong FA~~
5. Weak FA results
6. Future

The size of the action space



Many actions

Source	Action count	MDP class	Poly(.) compl?

** even under global access!
*   only under global access!

B: $\|\theta^*\| \leq B$

d: number of features (parameters)

H: horizon

A: number of actions

Norm of features ≤ 1

Simulator access models

- Global access:
  - Gets the description of the full state space
  - Gets all features at all states (state-action pairs) upfront
  - Can ask for a transition at any state-action pair
- Local access:
  - Does not get the description of the full state space
  - Only gets features associated with states visited
  - Simulation starts at some initial state
  - Simulator can be reset to a a previously visited state
- Online access:
  - Like local access, except that resetting to previously visited states is not possible

WSzGy21

Few actions

Action count	MDP class	Poly(.) compl?
$O(1)$ *, **	$\mathcal{M}^{v^*}_{B,d,H,A}$	✓
$O(1)$ **	$\mathcal{M}^{q^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	✓
$O(1)$ **	$\mathcal{M}^{v^*/q^*\mathrm{reach}}_{B,d,H,A}$	✓

* result by WAJAYJ*Sz*21
** even under local access
*** even under global access

Why not hard?

$O(1)$ actions: Why not hard?

Stochastic transitions
v^* realizability
local access

TensorPlan

WAJAYJ*Sz*21

$q_h(s,a;\theta) \coloneqq r_h(s,a) + \langle P_a(s), v_{h+1}(\cdot;\theta)\rangle$

TensorPlan

$$v_h^*(s) = \varphi_h(s)^\top \theta^* =: v_h(s;\theta^*)$$

$$\mathrm{TD}_h(s,a,\theta) \coloneqq r_h(s,a) + \langle P_a(s), v_{h+1}(\cdot;\theta)\rangle - v_h(s;\theta)$$

(*) $\quad \Pi_a \, \mathrm{TD}_h(s,a,\theta^*) = 0 \qquad \forall s,h \qquad$ Algebraic Bellman!

Start with $\Theta_0 = \{\, \theta : \|\theta\| \le B \,\}$

Iteration $i = 0,1,\ldots$:

- Pick $\theta = \mathrm{argmax}_{\theta' \in \Theta_i} v_0(s_0;\theta')$ # optimism
- Roll out/test with $\pi_h(s) = \mathrm{argmax}_a q_h(s,a;\theta) \to \left(\left(s_{j,h}, a_{j,h}\right)_{j,h}\right)$
- $\Theta_{i+1} = \{\theta \in \Theta_i \;:\; \Pi_a \widehat{\mathrm{TD}}_h\left(s_{j,h}, a_{j,h}, \theta\right) \approx 0 \;\; \forall j,h\,\}$
- Return $\pi_h(s_0)$ if $\Theta_{i+1} = \Theta_i$

Why will TensorPlan stop changing Θ?

$$\Pi_a \mathrm{TD}_h(s,a,\theta) = 0$$

$$\Leftrightarrow$$

$$\left\langle \otimes_a \overline{r_a(s)\left(P_a(s)^\top \phi_{h+1} - \phi_h(s)\right)}, \otimes_a \overline{1\ \theta} \right\rangle = 0$$

$\otimes_a \overline{1\ \theta} \in \mathbb{R}^{(d+1)^A}$

$\Rightarrow$ must stop after $(d+1)^A$ constraint violations

What's the role of optimism?

Consider the TensorPlan that in state s at stage h chooses the first action a s.t. $\mathrm{TD}_h(s, a, \theta) = 0$

..not necessarily a maximizing action

Let π be the corresponding policy

If $v_h^\pi(s) = \phi_h(s)^\top \theta \; \forall h, s$, TensorPlan could return $\pi(s_0)$!

..Problem? Not if $v_0^\pi(s_0) \geq v_0^*(s_0)$!

Since $\theta^* \in \Theta_i$, $v_0^\pi(s_0) = \max_{\theta \in \Theta_i} v_0(s_0; \theta) \geq v_0(s_0; \theta^*) = v_0^*(s_0)$

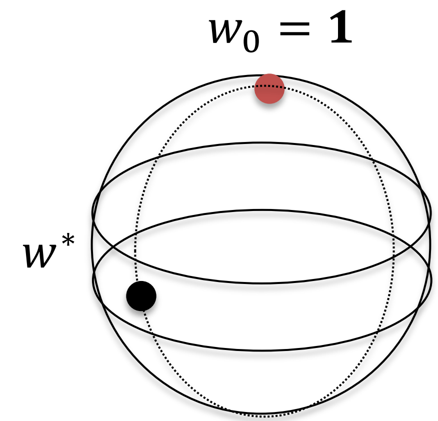
Theorem:

The number of simulator calls C performed by TensorPlan satisfies

$$C = O\left(\mathrm{poly}\left(\left(\frac{dH}{\delta}\right)^{A}, B\right)\right)$$

while TensorPlan induces a δ-optimal policy.

Hardness with poly actions



Challenges:

1. Algorithms can measure local consistency (w.r.t. TD error)
2. Large reward at stage H gives away θ^* (bandits!)
3. Need large total reward to keep action-gap large at s_0

Two-step approach:

1. Structured combinatorial semi-bandit where reward is the product of low-order polynomials with values in (0.1,0.9) action is chosen in K stages, need to "hit" nbh of $w^* \in \{-1,1\}^p$
2. Realize the semi-bandit with MDP with linear v^*

Summary

Source	Action count	MDP class	Poly(.) compl?
WAJAYJ*Sz*21	$O(1)$	$\mathcal{M}^{v^*}_{B,d,H,A}$	✓
W*Sz*Gy21	$O(1)$	$\mathcal{M}^{q^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	✓
W*Sz*Gy21	$O(1)$	$\mathcal{M}^{v^*/q^*\mathrm{reach}}_{B,d,H,A}$	✓
W*Sz*Gy21	$\Omega(d^{1/4} \wedge H^{1/2})$	$\mathcal{M}^{q^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	×
W*Sz*Gy21	$\Omega(d^{1/4} \wedge H^{1/2})$	$\mathcal{M}^{v^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	×
W*Sz*Gy21	$\Omega(d^{1/4} \wedge H^{1/2})$	$\mathcal{M}^{v^*/q^*\mathrm{reach}}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	×
WA*Sz*21	$2^{\Omega(d\wedge H)}$	$\mathcal{M}^{q^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{Pdet}}$	×
WR13	any	$\mathcal{M}^{q^*}_{B,d,H,A} \cap \mathcal{M}^{\mathrm{det}}$	✓
DKLLMSW21	any	$\mathcal{M}^{v^*/q^*}_{B,d,H,A}$	✓

poly compute for green lines? ☹☹☹ (KLLM’22)

> “Finally, if we combine these techniques – polynomial approximations and Lagrange multipliers – with that of successive approximations, there should be very few allocation processes which still resist our efforts.” [BKK63]

- Successive approximations?

  ..only for strong FA,

  ..for weak FA: constraint propagation/version space pruning

  and in stochastic systems, optimism
- Even with strong FA, we need to live with approximation error blowup due to extrapolation!
- Unlike in bandits, large action spaces cause hardness!

Some open problems

- Query complexity when

  $\mathcal{M}^{q^*}_{B,d,H,A}$, $A = O(1)$ AND transitions are stochastic

- Computational complexity when

  $A = O(1), \mathcal{M}^{v^*/q^*}_{B,d,H,A}$?

- Online access under $Q(\Pi) \subset \mathcal{F}$?

- Nonlinear fapp?

- Models that work for continuous action spaces?

Specializing the MDP class

- Deterministic dynamics is helpful
- Factored linear dynamics? Yes, eg,
$$s_{h+1} = f(s_h, a_h) + \eta, \eta \sim N(0, \Sigma)$$
  - Or just $T_* \mathcal{F} \subset \mathcal{F}$ or some variant of this
- Other special structure?
  - "Allocation processes"?
  - Linear dynamics, linear cost/reward, feasible action set is a polytope
  - …
- General characterization of query complexity (Foster, Kakade, Qian, Rakhlin)

Main references

[WR13] Z. Wen and B. Van Roy. 2017. "Efficient Reinforcement Learning in Deterministic Systems with Value Function Generalization", *Mathematics of Operations Research*, 42(3):762–782. [arXiv 2013]

[DKWY20] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. 2020. "Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?" ICLR and arXiv:1910.03016.

[LSzW20] Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. 2020. "Learning with Good Feature Representations in Bandits and in RL with a Generative Model." ICML and arXiv:1911.07676.

[WASz21] Gellért Weisz, Philip Amortila, Csaba Szepesvári. 2021. Exponential Lower Bounds for Planning in MDPs With Linearly-Realizable Optimal Action-Value Functions, ALT and arXiv:2010.01374

[WAJAJSz21] Gellért Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, Csaba Szepesvári. 2021. On Query-efficient Planning in MDPs under Linear Realizability of the Optimal State-value Function, COLT arXiv:2102.02049

[WGySz22] Gellért Weisz, András György, Csaba Szepesvári, 2022: TensorPlan and the Few Actions Lower Bound for Planning in MDPs under Linear Realizability of Optimal Value Functions. ALT

[DKLLMSW21] Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, Ruosong Wang, 2021 : Bilinear Classes: A Structural Framework for Provable Generalization in RL, ICML arXiv:2103.10897

References

- https://rltheory.github.io (RL Theory Lecture notes, CMPUT 652 @UofA, 2021-2022)
- Agarwal, Jiang, Kakade, Sun. Reinforcement Learning: Theory and Algorithms https://rltheorybook.github.io/
- RL Theory virtual seminar series: https://sites.google.com/view/rltheoryseminars

Basics

- Martin L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1994
- Chen, Y., & Wang, M. (2017). Lower bound on the computational complexity of discounted markov decision problems. arXiv preprint arXiv:1705.07312. [link]
- Singh, S. P., & Yee, R. C. (1994). An upper bound on the loss from approximate optimal-value functions. Machine Learning, 16(3), 227-233.
- Feinberg, E. A., Huang, J., & Scherrer, B. (2014). Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming. Operations Research Letters, 42(6-7), 429-431. [link]
- Scherrer, B. (2016). Improved and generalized upper bounds on the complexity of policy iteration. Mathematics of Operations Research, 41(3), 758-774. [link]
- Ye, Y. (2011). The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. Mathematics of Operations Research, 36(4), 593-603. [link]
- Kearns, M., Mansour, Y., & Ng, A. Y. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. Machine learning, 49(2), 193-208. [link]
- Remi Munos (2014). From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning. Foundations and Trends in Machine Learning: Vol. 7: No. 1, pp 1-129.

Online Learning

- Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. The Journal of Machine Learning Research, 11:1563–1600, 2010.
- Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. arXiv preprint arXiv:1906.05110, 2019.
- Bourel, Hippolyte, Odalric Maillard, and Mohammad Sadegh Talebi. 2020. “Tightening Exploration in Upper Confidence Reinforcement Learning.” Edited by Hal Daumé Iii and Aarti Singh, Proceedings of Machine Learning Research, 119: 1056–66.
- Fruit, Ronan, Matteo Pirotta, and Alessandro Lazaric. n.d. “Improved Analysis of UCRL2 with Empirical Bernstein Inequality.” https://rlgammazero.github.io/docs/ucrl2b_improved.pdf.
- Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
- L.J. Savage, The theory of statistical decision, J. Amer. Statist. Assoc. 46 (1951) 55-67.
- Wald, Statistical Decision Functions, Wiley, New York, 1950.

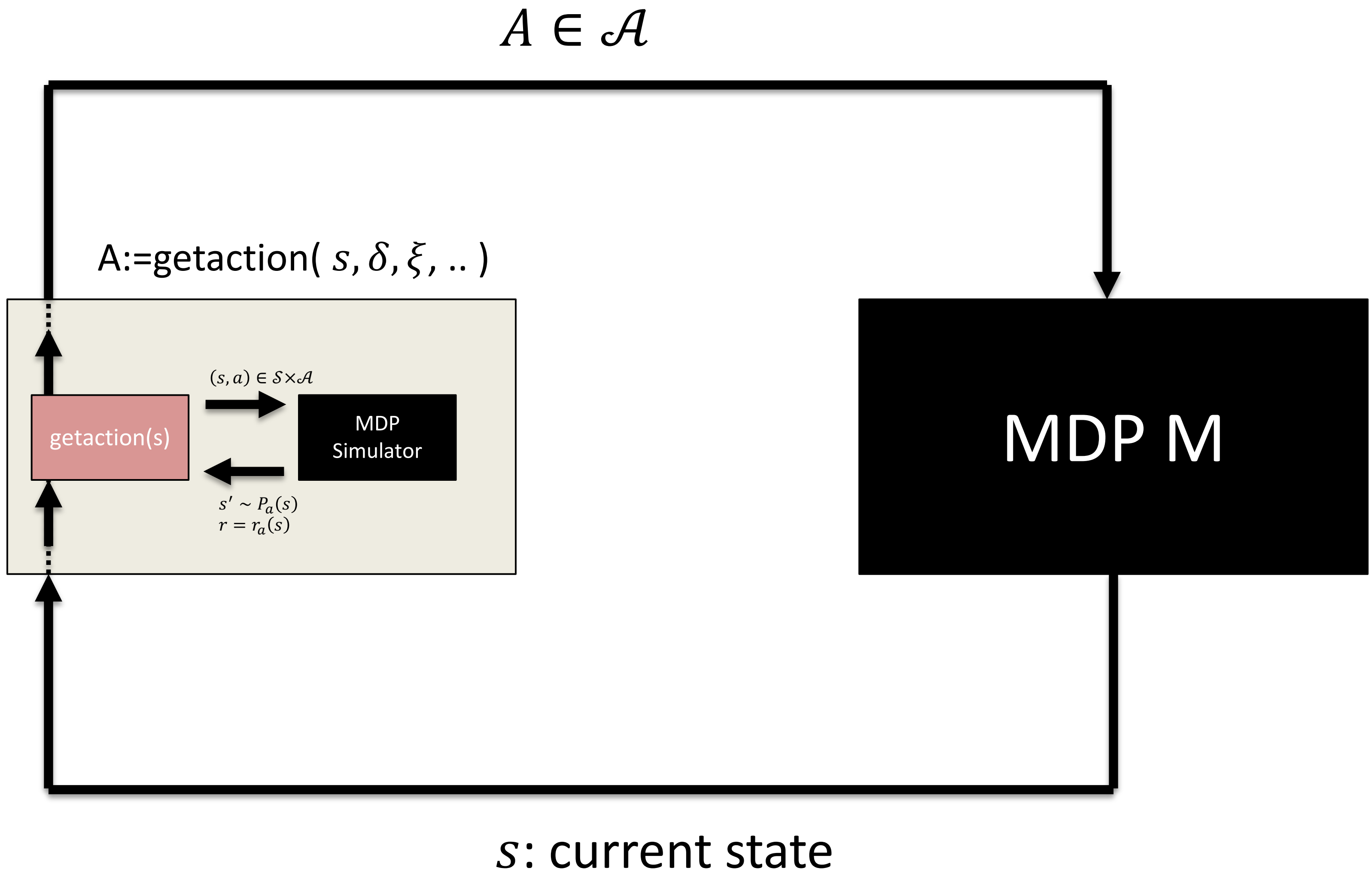
Function approximation

- Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. 2020. “Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?” ICLR and arXiv:1910.03016.
- Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. 2020. “Learning with Good Feature Representations in Bandits and in RL with a Generative Model.” ICML and arXiv:1911.07676.
- Roshan Shariff and Csaba Szepesvári. 2020. “Efficient Planning in Large MDPs with Weak Linear Function Approximation”. In NeurIPS 2020 and arXiv:2007.06184
- Gellért Weisz, Philip Amortila, Csaba Szepesvári. 2020. Exponential Lower Bounds for Planning in MDPs With Linearly-Realizable Optimal Action-Value Functions, To appear at ALT and also arXiv:2010.01374
- Gellért Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, Csaba Szepesvári. 2021. On Query-efficient Planning in MDPs under Linear Realizability of the Optimal State-value Function, arXiv:2102.02049
- Z. Wen and B. Van Roy. 2017. "Efficient Reinforcement Learning in Deterministic Systems with Value Function Generalization", *Mathematics of Operations Research*, 42(3):762–782. [arXiv]
- Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In Advances in Neural Information Processing Systems, pages 8060–8070, 2019b.

Function approximation/2

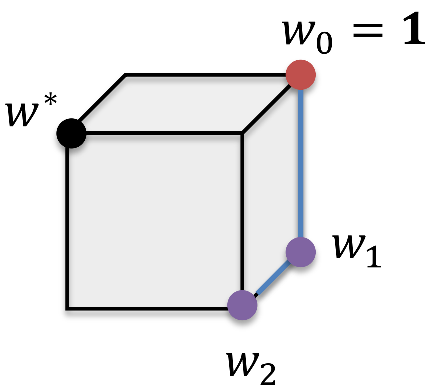
- Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In International Conference on Machine Learning, pages 1704–1713. PMLR, 2017.
- Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Conference on Learning Theory, pages 2137–2143, 2020.
- Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In ICML, pages 6995–7004, 2019.
- Richard Bellman, Robert Kalaba and Bella Kotkin. 1963. Polynomial Approximation--A New Computational Technique in Dynamic Programming: Allocation Processes. Mathematics of Computation, 17 (82): 155-161
- Daniel, James W. 1976. “Splines and Efficiency in Dynamic Programming.” *Journal of Mathematical Analysis and Applications* 54 (2): 402–7.
- Schweitzer, Paul J., and Abraham Seidmann. 1985. “Generalized Polynomial Approximations in Markovian Decision Processes.” *Journal of Mathematical Analysis and Applications* 110 (2): 568–82.

Online planning (R97, KMS02)



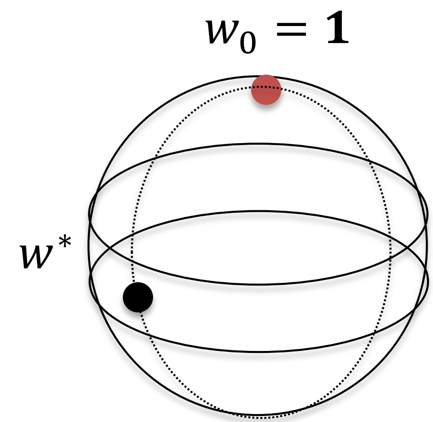
Objective: $v^\pi \geq v^* - \delta \mathbf{1}$ w.p. 1–ξ

The semi-bandit



p: dimension
K: #steps
Want: if both large, game is hard!
$w^*, w_i \in \{-1,1\}^p$

$$p/4 \leq h(w^*, \mathbf{1}) \leq 3p/4$$



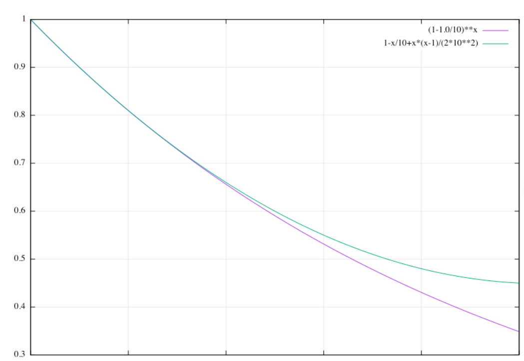
$k \leq K$ rounds, maximize f_{w^*}

$$f_{w^*}(w_{1:k}) = g\big(h(w_1, w_2)\big) \cdots g\big(h(w_{k-1}, w_k)\big)\, g\big(h(w_k, w^*)\big)$$

$$h(w_{i-1}, w_i) \geq \frac{p}{4}, 1 \leq i \leq k, \quad w_0 \quad = \mathbf{1}$$

stop as early as possible

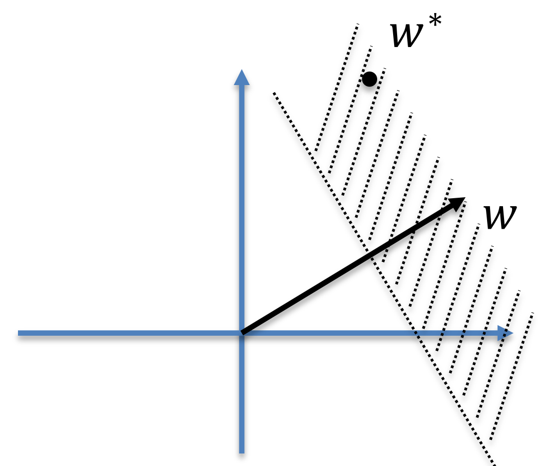
$$g(x) = 1 - \frac{x}{p} + \frac{(x-1)x}{2p^2}$$



Interaction

- Choose $w_{1:k}$ with some $1 \leq k \leq K$ (# rounds)
- Done? If yes, $k \coloneqq 8$, payoff is $R = f_{w^*}(w_{1:i})$ with smallest i such that $h(w_i, w^*) < p/4$, N: # queries before this round
- If not done then receive feedback:
  1. $h(w_{k-1}, w^*) < p/4$? ($w_0 \coloneqq \mathbf{1}$)
  2. $h(w_k, w^*) < p/4$?
  3. $Z \sim \mathrm{Ber}(f_{w^*}(w_{1:k}))$ if ($k = K$ or $h(w_k, w^*) < p/4$) else $Z = 0$

$$h(w, w') = 0.5\,(p - \langle w, w' \rangle)$$
$$h(w, w^*) < p/4 \;\Leftrightarrow \langle w, w^* \rangle > p/4$$



The lower bound

$\mathcal{A}$ is sound if for any $w^* \in W^*$,

$$\mathbb{E}_{w^*}^{\mathcal{A}}[R] \geq \max_{w_{1:8} \text{ adm.}} f_{w^*}\left(w_{1:i^*(w_{1:8})}\right) - 0.01$$

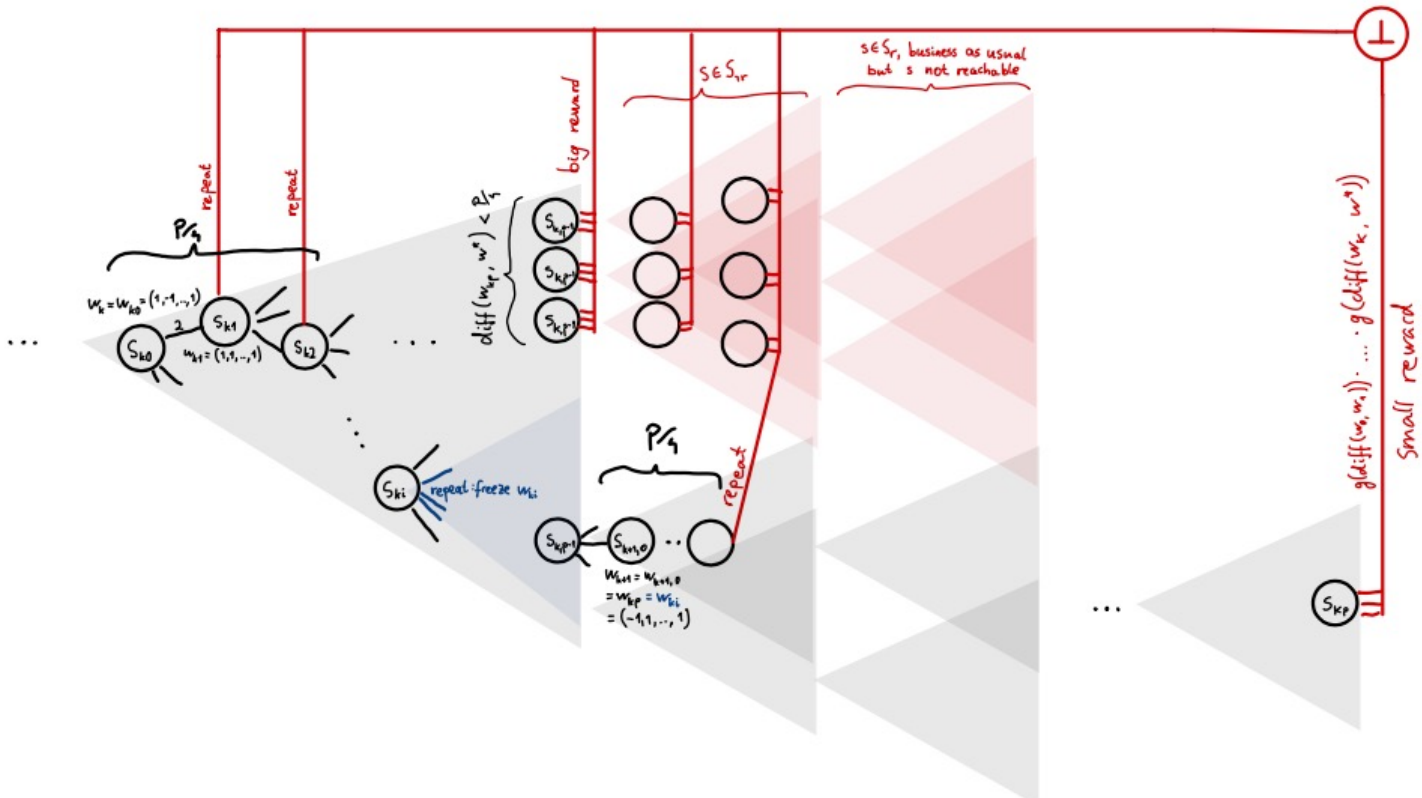
Theorem: If $\mathcal{A}$ is sound then $\max_{w^* \in W^*} \mathbb{E}_{w^*}^{\mathcal{A}}[N] = 2^{\Omega(p \wedge K)}$

Idea: Planner only gets info only when hits $B(w^*, \frac{p}{4})$. Chance of hitting this is $\exp\left(-\frac{p}{8}\right) \Rightarrow$ many queries are needed

Why this f_{w^*}? Helps with MDP realizability + large gap

MDP definition

- $H \approx Kp, A = p \approx d^{1/4} \wedge H^{1/2}$.
- Actions: flipping components



Robert Kalaba

Robert E. Kalaba, an applied mathematician associated with USC for almost half a century and internationally renowned for his analytical and computational solutions to problems in physics, engineering, operations analysis and biology.

A professor of biomedical engineering, electrical engineering and economics, Kalaba was an engineering lecturer at USC from 1956 to 1971.

He became a research associate in biomathematics in 1966 and a visiting professor of electrical engineering in the biomedical engineering program of the USC Viterbi School of Engineering in 1969.

In 1974, he became a full professor at USC with appointments in biomedical engineering, electrical engineering and economics.

1926—2004

https://news.usc.edu/24478/USC-Professor-of-Biomedical-Engineering-Dies/

Bella Kotkin → Bella Manel Greenfield

Bella Manel was born in New York City. A pioneering woman in mathematics, she earned her PhD in 1939 from New York University under the supervision of Richard Courant. She worked for Ramo-Wooldridge (now TRW) and at the Rand Corporation with Richard Bellman. Later, she taught mathematics at the College of Notre Dame (now Notre Dame de Namur University) in Belmont, California, and at UCLA. The Bella Manel Prize for outstanding graduate work by a woman or minority was established at NYU's Courant Institute in 1995.

October 13, 1915-
April 03, 2010

Spoke Hungarian?

https://www.wikidata.org/wiki/Q102188233