Reinforcement Learning in Linear MDPs: Constant Regret and Representation Selection

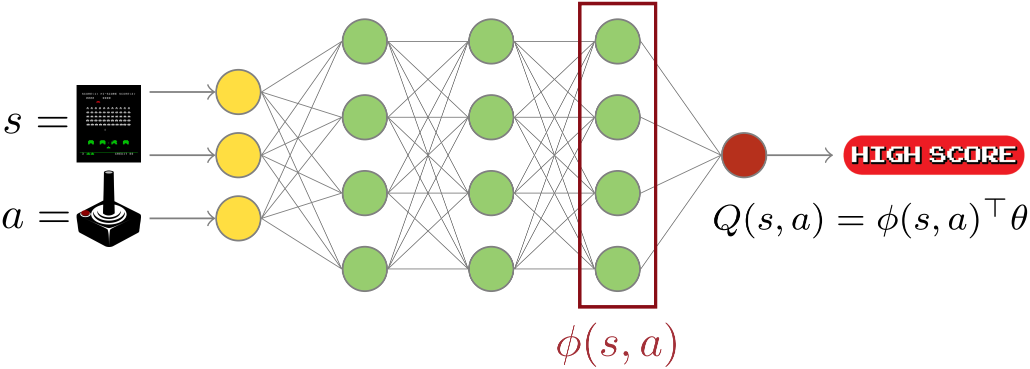
Matteo Papini[1], **Andrea Tirinzoni**[2], **Aldo Pacchiano**[3,†], **Marcello Restelli**[4], **Alessandro Lazaric**[5], **Matteo Pirotta**[5]

[1]Universitat Pompeu Fabra, [2]INRIA Lille [3]Microsoft Research, [4]Politecnico di Milano, [5]Facebook AI Research
[†] work done while at Facebook

Neural Information Processing Systems 2021

FACEBOOK AI

Motivation



Contributions

- Characterization of good representations for RL in linear MDPs
- Constant regret with good representations (LSVI-UCB, ELEANOR)
- Online representation selection (LSVI-LEADER)

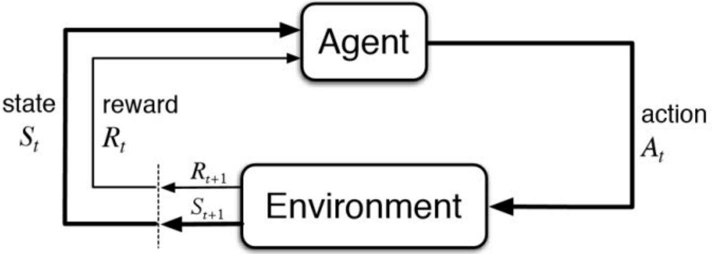
1 Linear Markov Decision Processes

2 Constant Regret with Good Representations

3 Representation Selection

Finite-Horizon Markov Decision Processes (MDPs)

$$(\mathcal{S}, \mathcal{A}, (r_h)_{h=1}^H, (p_h)_{h=1}^H, \mu)$$



- Finite horizon H
- Time-inhomogeneous
- Finite actions
- Possibly infinite states

Reinforcement Learning in Finite-Horizon MDPs

- Policy $\pi = (\pi_h)_{h=1}^{H}$, $\pi_h : \mathcal{S} \to \mathcal{A}$ (deterministic, time-dependent)
- Value function

$$Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim p_h(s,a)} \left[Q_{h+1}^\pi(s', \pi_{h+1}(s')) \right]$$

- Optimal policy

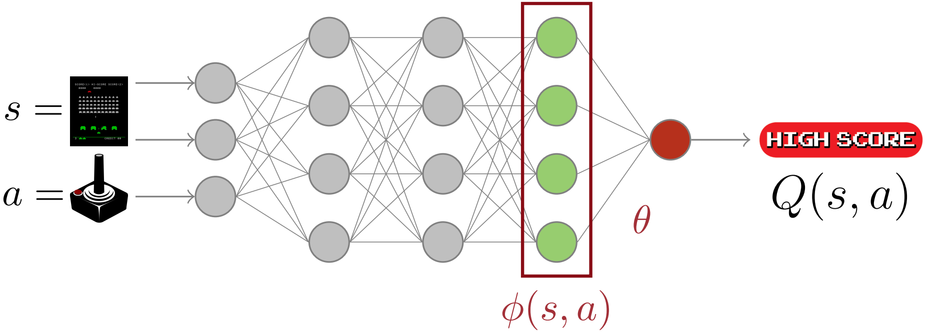
$$\pi^\star = \arg\max_\pi Q^\pi \qquad\qquad Q^\star = Q^{\pi^\star}$$

- Assumption: unique optimal action

$$\left| \arg\max_a \left\{ Q_h^\star(s,a) \right\} \right| = 1$$

Linear MDPs

Linear representation $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d, \qquad d << |\mathcal{S}|$



$$Q^\star(s,a) = \phi(s,a)^\top \theta^\star$$

Not enough! [Weisz et al., 2021]

Low-Rank MDPs

[Yang and Wang, 2019, Jin et al., 2020]

Low-rank MDP

For each $h \in [H]$ there are $\nu_h \in \mathbb{R}^d$ and $\mu_h : \mathcal{S} \to \mathbb{R}^d$ such that

$$r_h(s,a) = \phi_h(s,a)^\top \nu_h \qquad p_h(s'|s,a) = \phi_h(s,a)^\top \mu_h(s')$$

Implies linearly realizable Q-function [Jin et al., 2020]: for each π there is θ^π such that

$$Q_h^\pi(s,a) = \phi_h(s,a)^\top \theta_h^\pi$$

Bellman Closure [Zanette et al., 2020]

Bellman-Closure MDP

For all θ there is θ' such that for all s, a, h:

$$\phi_h(s,a)^\top \theta' = r_h + \mathbb{E}_{s' \sim p_h(s,a)} \left[\max_{a'} \phi_{h+1}(s',a')^\top \theta \right]$$

- Weaker than low-rank
- Linearly realizable *optimal* value function:

$$Q_h^\star(s,a) = \phi_h(s,a)^\top \theta_h^\star$$

Regret Bounds

$$V_h^\pi(s) = \max_a Q_h^\pi(s,a)$$

$$R(K) = \sum_{k=1}^{K} V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)$$

Assumption: positive suboptimality gaps

$$\Delta_h(s,a) = V_h^\star(s) - Q_h^\star(s,a)$$

$$\Delta_{\min} = \min_{s,h,a\neq\pi_h^\star(s)} \Delta_h(s,a) > 0$$

Algorithm (setting)	Minimax	Problem-Dependent Logarithmic	
ELEANOR[1] *(Bellman Closure)*	$\widetilde{O}(\sqrt{d^2H^3T})$ [Zanette et al., 2020]	N/A	**Can we do better?**
LSVI-UCB *(low-rank MDPs)*	$\widetilde{O}(\sqrt{d^3H^3T})$ [Jin et al., 2020]	$O\left(\frac{d^3H^5}{\Delta_{\min}}\log^2(T)\right)$ [He et al., 2020]	**Can we do better?**
Lower Bound	$\Omega(\sqrt{d^2H^2T})$ [Zhou et al., 2020, Remark 5.8]	$\Omega\left(\frac{dH}{\Delta_{\min}}\right)$ [He et al., 2020]	

[1] Computationally intractable!

The UNISOFT Property (UNIversally Spanning Optimal FeaTures)

Inspired by Hao et al. [2020], Papini et al. [2021]

Optimal features $\phi_h^\star(s) = \phi_h(s, \pi_h^\star(s))$

state visitation probability under policy π

$$\operatorname{span}\left\{ \phi_h^\star(s) \mid s \in \operatorname{supp}(\rho_h^\star)\right\} = \operatorname{span}\left\{\phi_h(s,\pi(s)) \mid s \in \operatorname{supp}(\rho_h^\pi) \text{ for some } \pi\right\}$$

- A representation is UNISOFT if optimal features span the whole feature space
- A sufficient condition is (necessary if features span $\mathbb{R}^d$):

$$\lambda_+ = \min_{h\in[H]} \lambda_{\min}\left(\mathbb{E}_{s\sim\rho_h^\star}[\phi_h^\star(s)\phi_h^\star(s)^\top]\right) > 0$$

- In general we can consider the minimum *nonzero* eigenvalue (larger is better)
- $\|\phi(s,a)\| \leq 1 \implies \lambda_+ \leq 1$

UNISOFT is necessary for Constant Regret

Necessity of UNISOFT

Consider any MDP with *linear rewards*:

$$r_h(s,a) = \phi_h(s,a)^\top \nu_h$$

If ϕ is not UNISOFT, no consistent[2] algorithm can achieve constant regret

This applies to low-rank, Bellman closure, and even linear-mixture MDPs (with unknown linear rewards) [Jia et al., 2020, Ayoub et al., 2020, Zhou et al., 2020]

[2] We only ask the algorithm to suffer sublinear regret for all alternative reward parameters

UNISOFT is sufficient for Constant Regret

Regret of LSVI-UCB with UNISOFT

LSVI-UCB achieves CONSTANT regret in low-rank MDPs if and only if the representation is UNISOFT. With probability $1-\delta$:

$$R(K) \lesssim \frac{d^3 H^5}{\Delta_{\min}} \log(dH\tau/\delta)$$

where $\tau \lesssim \dfrac{H^5 d^3}{\lambda_+^3 \Delta^2}$ is a constant independent of K

After τ interactions, the agent has learned the optimal policy

UNISOFT is sufficient for Constant Regret

Algorithm (setting)	Minimax	Problem-Dependent Logarithmic	Constant with UNISOFT (***this work***)[3]
ELEANOR *(Bellman Closure)*	$\widetilde{O}(\sqrt{d^2H^3T})$ [Zanette et al., 2020]	N/A	$\widetilde{O}\left(\frac{d^2H^4}{\Delta_{\min}\lambda_+^{3/2}}\right)$
LSVI-UCB *(low-rank MDPs)*	$\widetilde{O}(\sqrt{d^3H^3T})$ [Jin et al., 2020]	$O\left(\frac{d^3H^5}{\Delta_{\min}}\log^2(T)\right)$ [He et al., 2020]	$\widetilde{O}\left(\frac{d^3H^5}{\Delta_{\min}}\right)$
Lower Bound	$\Omega(\sqrt{d^2H^2T})$	$\Omega\left(\frac{dH}{\Delta_{\min}}\right)$ [He et al., 2020]	N/A

After $k_{\mathcal{A}}$ episodes, the agent $\mathcal{A}$ has learned the optimal policy

[3] Here $\widetilde{O}$ hides terms in $d, H, \Delta_{\min}, \lambda_+, \delta$, but not in T

Representation Selection in Low-Rank MDPs

Typical approach: find an accurate representation in a realizable function class, usually offline [Agarwal et al., 2020, Modi et al., 2021, Lu et al., 2021]

Our setting:

- Agent is given N equivalent linear representations $\phi^1, \ldots, \phi^N$
- Each ϕ^i inducing the same low-rank MDP (no misspecification)
- Possibly different dimension
- Goal: learn as if using the best candidate representation (possibly UNISOFT)

LSVI-LEADER

At each episode k

- For each representation $j \in [N]$, compute an *optimistic* estimate $\overline{Q}_j^k$ of $Q^\star$ using *all* past interaction data

$$\text{Backward induction } (h = H, \dots, 1): \quad Y_i^h = r_i^h + \max_a \min_j \overline{Q}_{j,h+1}^k(s_i^{h+1}, a)$$

$$\text{Least Squares:} \quad \widehat{\theta}_{j,h}^k = (\underbrace{\Lambda_{j,h}^k}_{\text{design matrix}^4})^{-1} \sum_{i=1}^{k-1} \phi_h^j(s_i^h, a_i^h) \underbrace{Y_i^h}_{\text{target}}$$

$$\text{Optimism:} \quad \overline{Q}_{j,h}^k(s,a) = \phi_h^j(s,a)^\top \widehat{\theta}_{j,h}^k + \underbrace{\beta_k \|\phi_h^j\|_{(\Lambda_{j,h}^k)^{-1}}}_{\text{exploration bonus}}$$

- Act greedily w.r.t. the tightest optimistic estimate

$$a_h^k = \arg\max_a \min_j \overline{Q}_{j,h}^k(s,a)$$

[4] $\Lambda_{j,h}^k = \sum_{i=1}^k \phi_h^j(s_i^h, a_i^h)\phi_h^j(s_i^h, a_i^h)^\top$

LSVI-LEADER Achieves Constant Regret

Regret of LSVI-LEADER

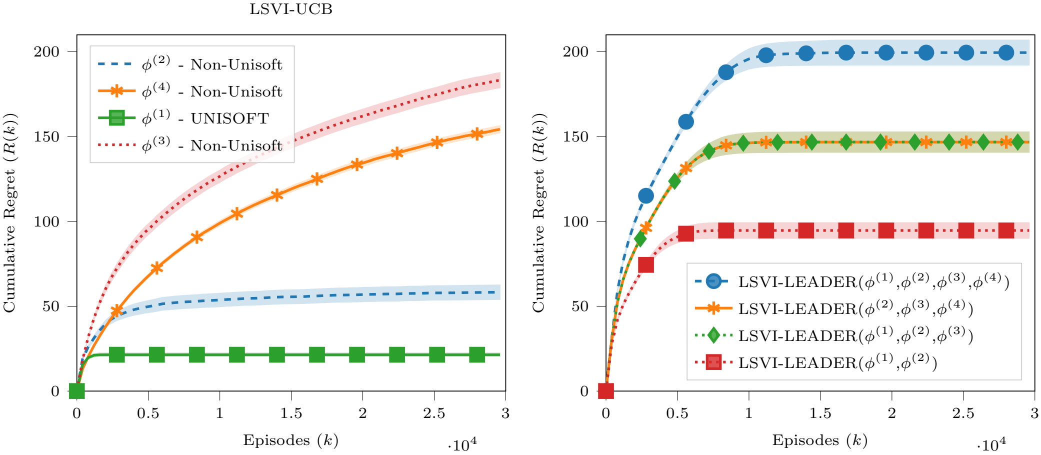
Let $R(K;\phi)$ (an upper bound on) the regret that LSVI suffers by using representation ϕ. The regret of LSVI-LEADER with candidate representations $\phi^j, \ldots, \phi^N$ is

$$R(K) \lesssim \sqrt{N} \min_{\phi \in \Phi} R(K;\phi)$$

where Φ is the set of H^N representations obtained by combining the N candidates *across stages*.

- If one of the candidate representations is UNISOFT, LSVI-LEADER achieves constant regret.
- LSVI-LEADER can combine representations also across *states and actions* and achieve constant regret under a weaker notion of UNISOFT

Empirical Results



Future Work

- Improve the $\sqrt{N}$ factor in LSVI-LEADER ($\log N$ for linear bandits)
- Misspecified representations
- Representation learning for Deep RL
- Multi-task RL [Lu et al., 2021]

Thank you

Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: structural complexity and representation learning of low rank mdps. In *NeurIPS*, 2020.

Alex Ayoub, Zeyu Jia, Csaba Szepesvári, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 463–474. PMLR, 2020.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3536–3545. PMLR, 2020.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *CoRR*, abs/2011.11566, 2020.

Zeyu Jia, Lin Yang, Csaba Szepesvári, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *L4DC*, volume 120 of *Proceedings of Machine Learning Research*, pages 666–686. PMLR, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020.

Rui Lu, Gao Huang, and Simon S. Du. On the power of multitask representation learning in linear MDP. *CoRR*, abs/2106.08053, 2021.

Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8371–8380. PMLR, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pages 1237–1264. PMLR, 2021.

Lin F. Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *CoRR*, abs/1905.10389, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel J. Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *CoRR*, abs/2012.08507, 2020.