Sparse Multi-Task Reinforcement Learning

DANIELE.CALANDRIELLO@INRIA.FR, ALESSANDRO.LAZARIC@INRIA.FR, AND MARCELLO.RESTELLI@POLIMI.IT

Contribution

Objective: Exploit ***similarity*** across multiple tasks to ***reduce number of samples*** needed to learn a near-optimal policy.

Results:

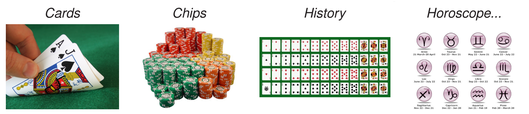
- We define similarity across MDPs in terms of their level of ***sparsity***.
- We introduce ***three algorithms*** obtained by integrating sparse multi-task regression with fitted Q-iteration, including a method the automatically ***learn the most sparse representation***.
- Provable ***sample complexity reduction***.
- Significant ***empirical improvement*** over single-task baselines.

Example

Multiple Tasks to Solve

One Policy Per Task

Task Representation (features)



All solutions use the same (small) set of features!

Multi-Task Reinforcement Learning

Markov decision processes (MDPs) $\mathcal{M}_t = (\mathcal{X}, \mathcal{A}, R_t, P_t)$, $t \in [T]$

- $\mathcal{X}$ is the state space,
- $\mathcal{A}$ is the action space,
- $R_t : \mathcal{X} \times \mathcal{A} \to [0,1]$ is the ***task*** reward,
- $P_t : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$ is the ***task*** dynamics,

Policy $\pi : \mathcal{X} \to \mathcal{A}$ maps states to actions.

Optimal Bellman operator $\mathcal{T}_t : \mathcal{B}(\mathcal{X} \times \mathcal{A}; Q_{\max}) \to \mathcal{B}(\mathcal{X} \times \mathcal{A}; Q_{\max})$

$$\mathcal{T}Q(x,a) = R_t(x,a) + \gamma \sum_y P_t(y|x,a) \max_{a'} Q(y,a')$$

Objective compute the optimal action-value function $Q_t^* = \mathcal{T}_t Q_t^*$ and the optimal policy $\pi_t^*(x) = \max_a Q_t^*(x,a)$ for each ***task*** t.

Linear Fitted Q-Iteration (see e.g., [1])

Linear function space $\mathcal{F} = \{f_w(x,a) = \phi(x)^\mathsf{T} w_a, w_a \in \mathbb{R}^{d_x}\}$, $d = |\mathcal{A}| \cdot d_x$

Set of task samples $\{\mathcal{S}_t = \{x_i\}_{i=1}^{n}\}_{t=1}^{T}$

Initialize $W^0 \leftarrow \mathbf{0}$, $k = 0$
for $k = 1, \dots, K$ **do**
  for $a \leftarrow 1, \dots, |\mathcal{A}|$ **do**
    for $t \leftarrow 1, \dots, T$, $i \leftarrow 1, \dots, n_x$ **do**
      Sample $r_{i,a,t}^k = R_t(x_{i,t}, a)$ and $y_{i,a,t}^k \sim P_t(\cdot|x_{i,t}, a)$
      Compute $z_{i,a,t}^k = r_{i,a,t}^k + \gamma \max_{a'} \widetilde{Q}_t^k(y_{i,a,t}^k, a')$
    end for
    Build datasets $\mathcal{D}_{a,t}^k = \{(x_{i,t}, a), z_{i,a,t}^k\}_{i=1}^n$
    Compute $\widehat{W}_a^k$ ***on*** $\{\mathcal{D}_{a,t}^k\}_{t=1}^T$ ***with MTL regression***
  end for
end for

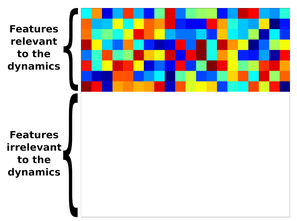
High-dimensional Sparse MDPs

Problem: the regression ***approximation must be accurate***. ***Solution:*** use ***large number of features***, rich feature space captures everything.

Assumption 1 (high-dimensional space) *For any function $f_w \in \mathcal{F}$, the Bellman operator $\mathcal{T}$ can be expressed as*

$$\mathcal{T} f_w(x,a) = R(x,a) + \gamma \mathop{\mathbb{E}}_{x' \sim P(\cdot|x,a)} [f_w(x', \pi_w(x'))] = \psi(x,a)^\mathsf{T} w^R + \gamma \psi(x,a)^\mathsf{T} P_\psi^{\pi_w} w$$

and thus there exists $f_{w'} \in \mathcal{F}$ such that $f_{w'} = \mathcal{T} f_w$ with $w' = w^R + P_\psi^{\pi_w} w$.



⇒ **Linear Model:** at each iteration k, the samples are generated according to a **true** vector w_a^k and perturbed by a zero-mean bounded noise:

$$z_{i,a}^k = \mathcal{T}\widehat{Q}^{k-1}(x_i, a) + \eta_{i,a}^k = \phi(x_i)^\mathsf{T} w_a^k + \eta_{i,a}^k$$

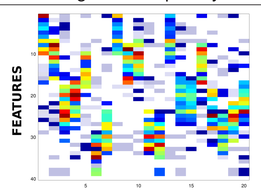
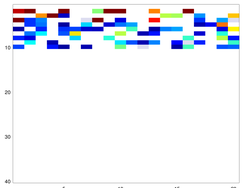
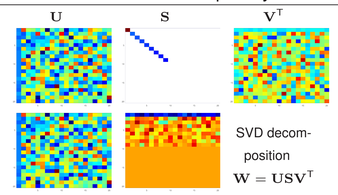
Problem: high-dimensional regression ***requires too many samples***. ***Solution:*** use ***regularization*** to induce ***sparsity***.

Assumption 2 (Sparse MDPs) *For each task $t \in [T]$, there exists a set J_t (the set of useful features), such that for any $i \notin J$, and any policy π the rows $[P_\psi^\pi]^i$ are equal to 0, and there exists a function $f_{w^R} = R$ such that $J(w^R) \subseteq J$.*

⇒ **MDP Sparsity:** for any function $f_w \in \mathcal{F}$, the Bellman image $f_{w'} = \mathcal{T}_t f_w$ is such that $J_t(w') \subseteq J_t$ for any task $t \in [T]$.

Sparse Multi-Task Fitted Q-iteration Algorithms

Sparsity Patterns

Single-Task Sparsity	Shared Sparsity	"Hidden" Shared Sparsity
		
$\bar{s} = \sum s_t / T$	$\widetilde{s} = \bigcup J(s_i)$	$s^* = \mathrm{Rank}(W)$

Multi-Task Regression Problems

LASSO–FQI [2]	GL–FQI [3]	FL–FQI [4]
$\min_{w \in \mathbb{R}^{d_x}} \frac{1}{n_x} \sum_{i=1}^{n_x} \left(\phi(x_i)^\mathsf{T} w - z_{i,a}^k\right)^2 + \lambda \lvert\lvert w \rvert\rvert_1$	$\min_W \sum_{t=1}^T \left\lVert Z_{a,t}^k - \Phi_t w_t \right\rVert_2^2 + \lambda \left\lVert W \right\rVert_{2,1}$	$\min_{U_a \in \boldsymbol{O}^d, A \in \mathbb{R}^{d \times T}} \sum_{t=1}^T \lvert\lvert Z_{a,t}^k - \Phi_t U_a[A]_t \rvert\rvert^2 + \lambda \left\lVert A \right\rVert_{2,1}$ $= \min_W \sum_{t=1}^T \lvert\lvert Z_{a,t}^k - \Phi_t [W]_t \rvert\rvert^2 + \lambda \lvert\lvert W \rvert\rvert_1$

Performance Guarantees

$$\frac{1}{T} \sum_{t=1}^T \left\lVert Q_t^* - Q_t^{\pi_t^K} \right\rVert_{2,\mu}^2 \le \mathcal{O}\left(\frac{1}{(1-\gamma)^4} \left[\bullet\right]\right)$$

LASSO–FQI	GL–FQI	FL–FQI
$\left[\frac{Q_{\max}^2 L^2}{\kappa_{\min}^4(\bar{s})} \frac{\bar{s} \log d}{n} + \gamma^K Q_{\max}^2\right]$	$\left[\frac{L^2 Q_{\max}^2}{\kappa^4(2\tilde{s})} \frac{\tilde{s}}{n} \left(1 + \frac{(\log d)^{3/2+\delta}}{\sqrt{T}}\right) + \gamma^K Q_{\max}^2\right]$	$\left[\frac{Q_{\max}^2 L^4}{\kappa^2} \frac{s^*}{n} \left(1 + \frac{d}{T}\right) + \gamma^K Q_{\max}^2\right]$
+ Only logarithmic dependency on d + Scale with the number of **useful** features $\bar{s}$ − No advantage from multiple tasks	+ When T is large, no dependency on d − $\tilde{s}$ may be larger than $\bar{s}$ and as large as d	+ Learn the most sparse representation + Linear dependency on d may be reduced by number of tasks T − Limited to linear transformations

Assumptions

$$\min \left\{ \frac{\left\lVert \Phi \Delta \right\rVert_2^2}{nT \left\lVert \Delta_J \right\rVert_2^2} : |J| \le s, a \left\lVert \Delta_{J^c} \right\rVert_p \le b \left\lVert \Delta_J \right\rVert_p \right\} \ge \kappa(s)$$

LASSO–FQI: Restricted Eigenvalues	GL–FQI: Multi-Task Restricted Eigenvalue	FL–FQI: Restricted Strong Convexity
P. Bickel, Y. Ritov, and A. B Tsybakov. Simultaneous analysis of lasso and dantzig selector. 2009.	K. Lounici, M. Pontil, S. Van De Geer, A. B Tsybakov. Oracle inequalities and optimal inference under group sparsity. 2011.	S. Negahban, M. Wainwright, et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. 2011.

Experiments: BlackJack

Rules:

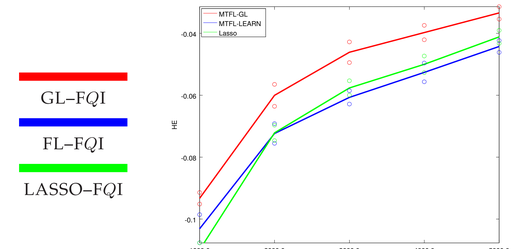
- ♢ Player can ask another card "HIT" or "STAY"
- ♡ If the player goes over 21, he loses, end of game
- ♠ Dealer has to "HIT" until a threshold, then "STAY"
- ♣ If the dealer goes over 21, player wins.
- ♢ If the player has a strictly higher score than the dealer, player wins

Multiple Tasks:

- ♡ Dealer threshold $\{15, 16, 17, 18\}$
- ♠ Number of decks $\{2, 4, 6, 8\}$
- ♣ If the dealer "HIT" when has a soft ace (A=11)

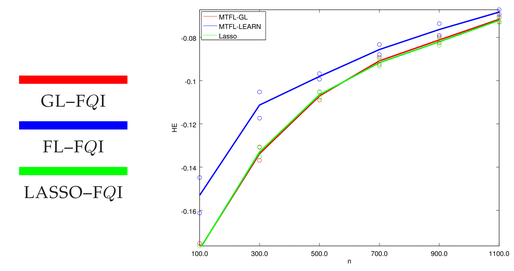
Full Variant Game

- *Actions:* Player can "DOUBLE" his bet after the first two cards
- *Features:* Indicator functions for player's and dealer's hand, and large number of indicator functions for the game history
- *Sparsity:* Most of the history features could be irrelevant.



Reduced Variant Game

- *Actions:* Player cannot "DOUBLE" his bet after the first two cards.
- *Features:* Indicator functions for player's and dealer's hand.
- *Sparsity:* Dense (=non-sparse) representation that can produce correlated tasks (low rank problem).



References & Acknowledgements

[1] Damien Ernst, Pierre Geurts, Louis Wehenkel, and Michael L Littman. Tree-based batch mode reinforcement learning. *JMLR*, 6(4), 2005.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2009.

[3] K. Lounici, M. Pontil, S. Van De Geer, A. B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

ComPLACS

ANR PROJECT FUNDED BY THE ANR

photos from lostvegascasinohire.co.uk, swisscasinos.ch, blackjackinfo.com, thehdwall.com, ruhrpott-poker.de, studieinfo.nl