Reinforcement Learning in Rich-Observation MDPs using Spectral Methods

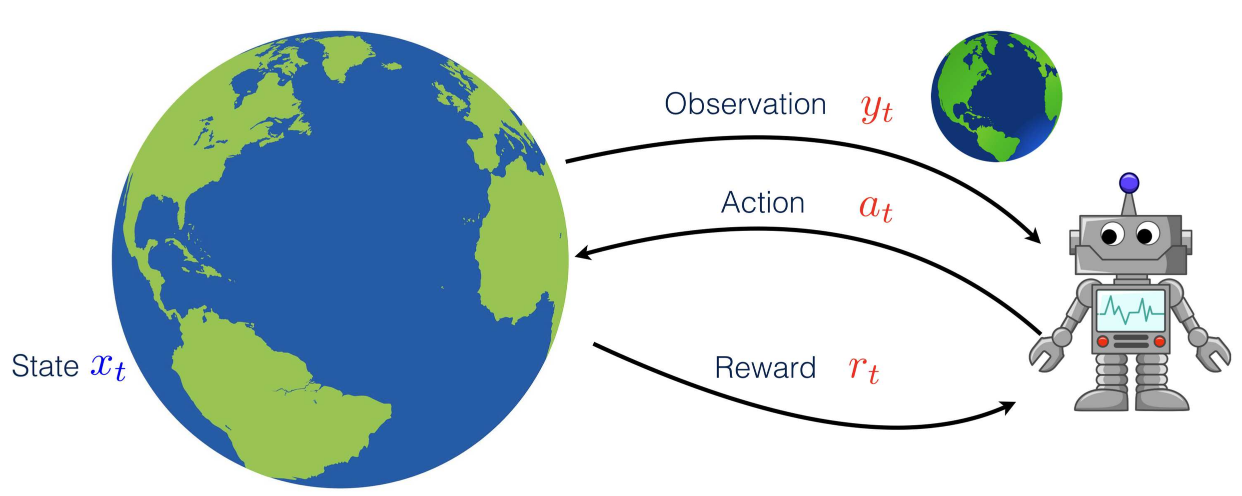
Kamyar Azizzadenesheli* Alessandro Lazaric† Animashree Anandkumar*

*University of California, Irvine (UCI) †Institut National de Recherche en Informatique et en Automatique, (INRIA)

Reinforcement Learning

Agent-Environment interactions under uncertainty:

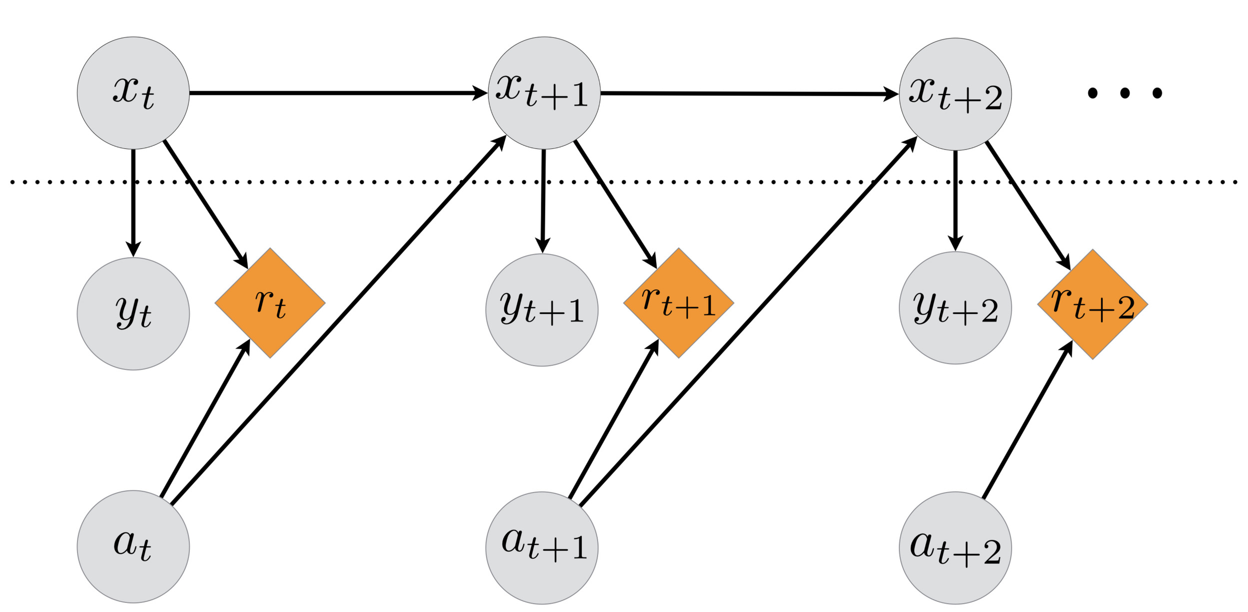
- Policy $\pi(a|y) : \mathcal{Y} \rightarrow \mathcal{A}$.
- Goal: $\max_{\pi} \eta_\pi = \max_{\pi} \lim_{N\rightarrow\infty} \frac{1}{N} \mathbb{E}_\pi \sum_t^N r_t$
- No prior knowledge
- Learning (Exploring)
- Planning (Exploiting)
- Undiscounted average reward

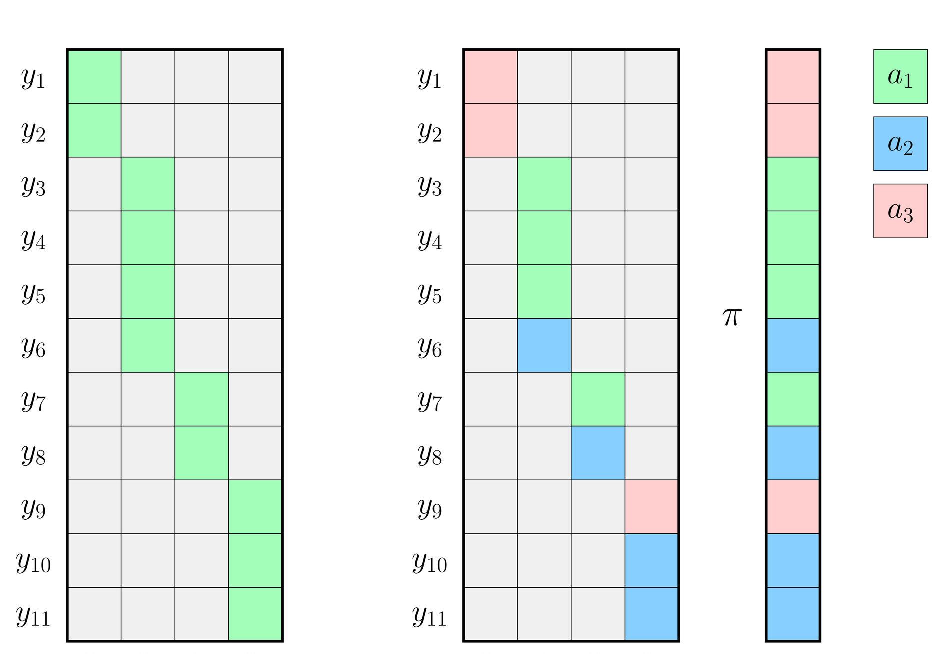


Large MDPs

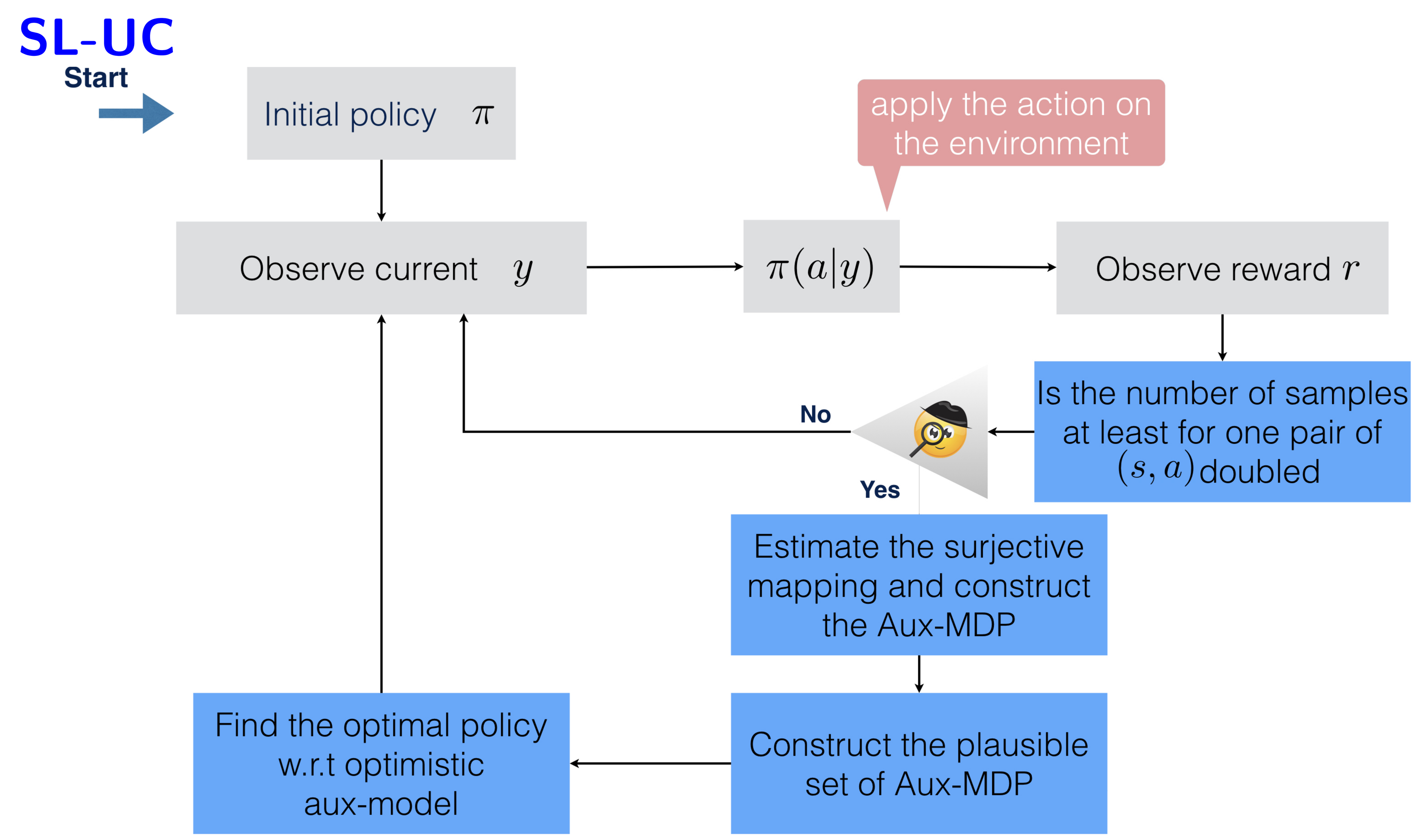
Structured MDPs

- Rich-Observation MDP (ROMDP)
- Injective mapping from $x \rightarrow y$
- Known mapping $\rightarrow$ Regret$(T) = \tilde{\mathcal{O}}\left(D_\mathcal{X} X\sqrt{AT}\right)$
- No Prior knowledge $\rightarrow$ Learn the mapping





SL-UC



Spectral Methods

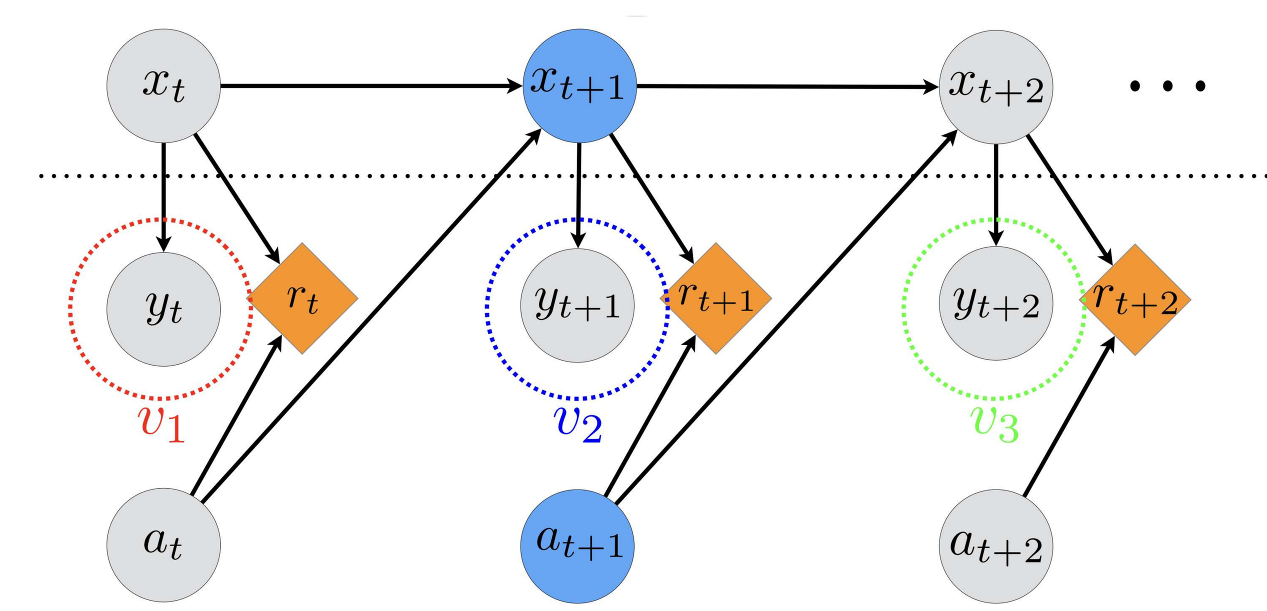
Tensor Decomposition:
Multiview model condition on middle action and middle state

Tensor Moments

- $v_t \perp v_{t+1} \perp v_{t+2} | x_{t+1}, a_{t+1}$
- $V_i^{(l)} = \mathbb{P}\left(\vec{y}_i | x_2, a_2 = l\right), \rightarrow V_1^{(l)}, V_2^{(l)}, V_3^{(l)} \in \mathbb{R}^{Y\times X}$

$$\boxed{\mathbb{E}[v_1 \otimes v_2 \otimes v_3 | a_2 = l] = \sum_j \omega_\pi^{(l)} \cdot [V_1^{(l)}]_{:,j} \otimes [V_2^{(l)}]_{:,j} \otimes [V_3^{(l)}]_{:,j}}$$

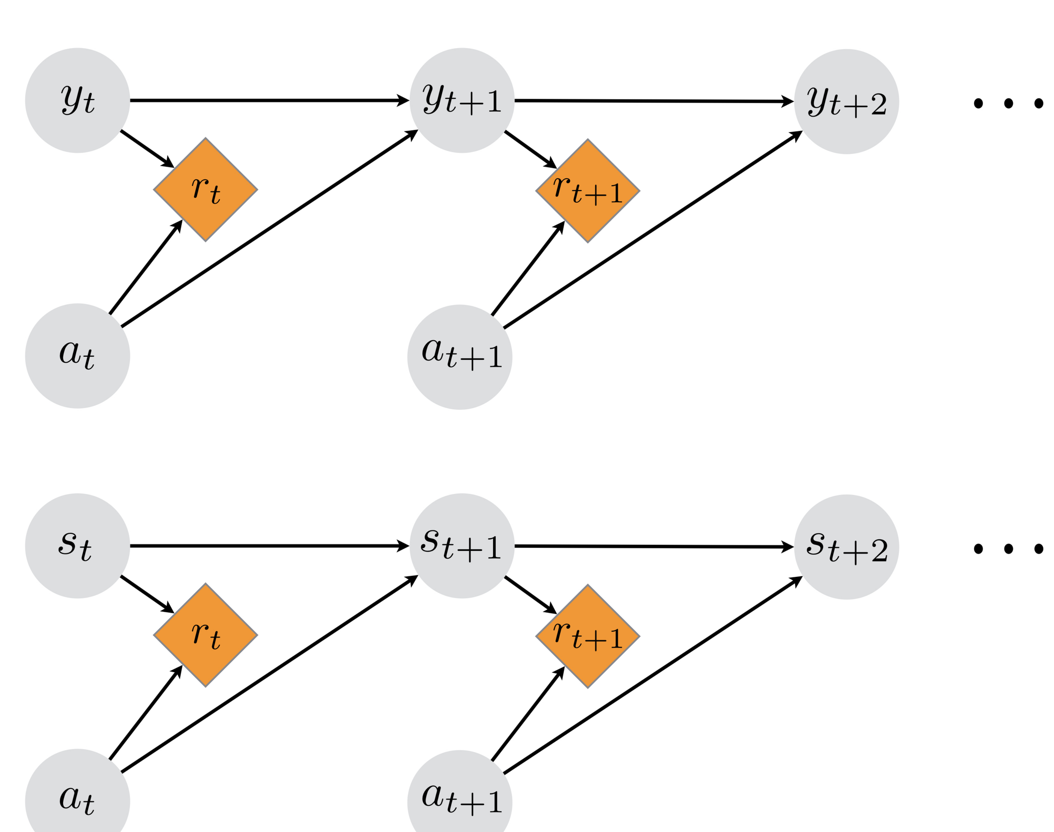
Multiview Model

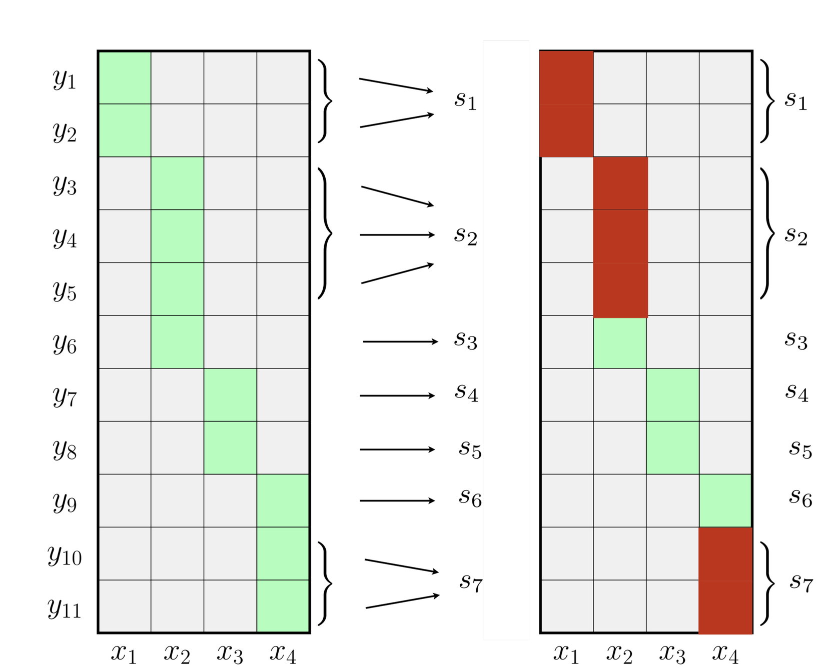


Parameter Learning

Second and Third order moments given middle action **Confidence intervals**

$$\left.\begin{aligned} M_2^{(l)} &= \sum_x \omega^{(l)}(x)[V_1^{(l)}]_{:,x} \otimes [V_3^{(l)}]_{:,x} \\ M_3^{(l)} &= \sum_x \omega^{(l)}(x)[V_1^{(l)}]_{:,x} \otimes [V_3^{(l)}]_{:,x} \otimes [V_2^{(l)}]_{:,x} \end{aligned}\right\} \Rightarrow \|\hat{O}^{(l)}(:,i) - O^{(l)}(:,i)\|_2 = \mathcal{O}\left(\sqrt{\frac{\log(Y/\delta)}{T_i}}\right).$$



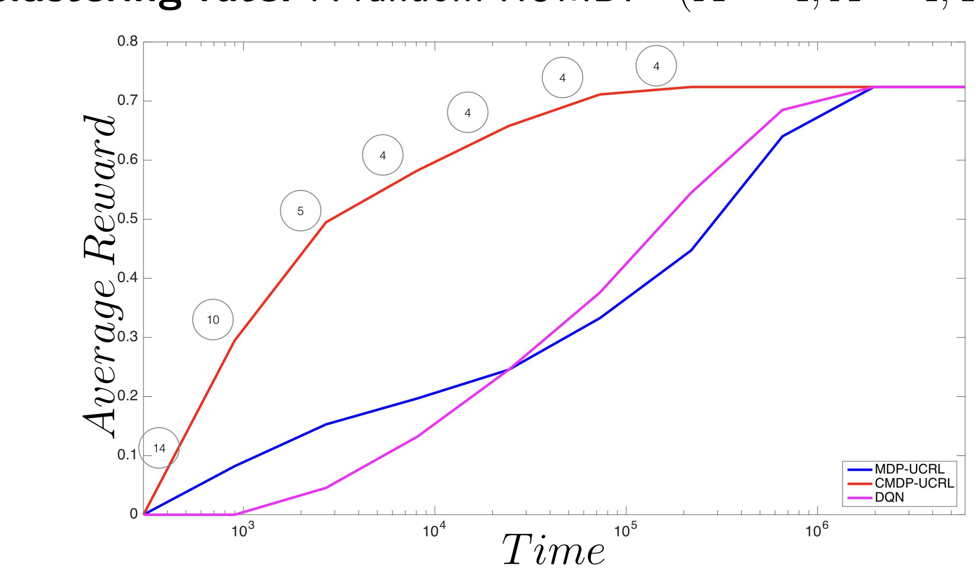


Results

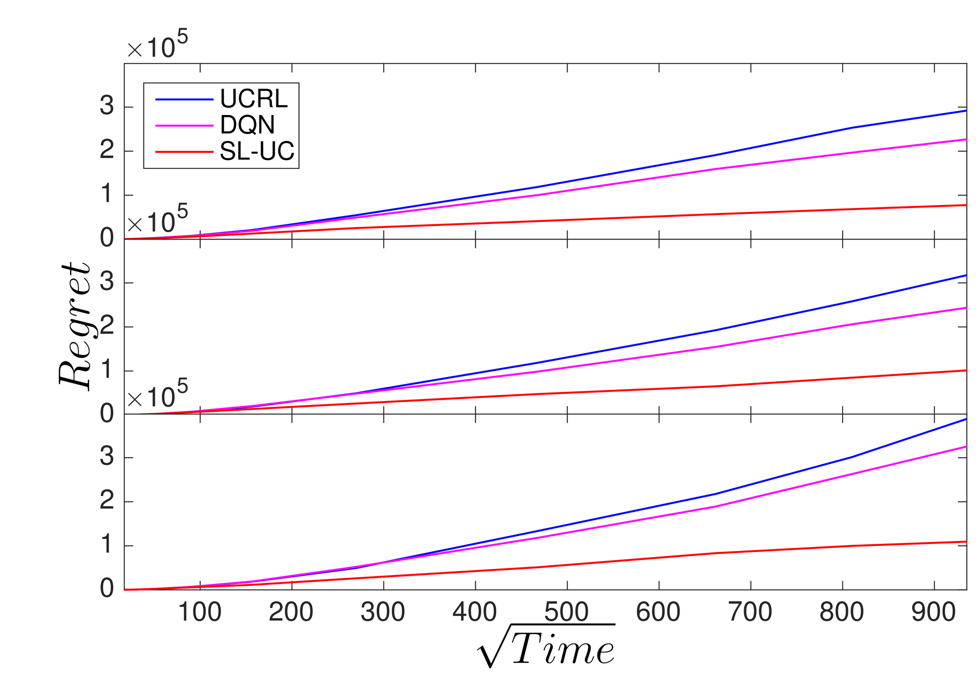
Theorem: *SL-UC* achieves a regret of
$$\boxed{\text{Regret}(T) = \tilde{\mathcal{O}}\left(D_\mathcal{X} X\sqrt{AT}\right)}$$

- Observation independent regret,
- Optimal regret (UCRL)
  Regret$(T) = \tilde{\mathcal{O}}\left(D_\mathcal{Y} Y\sqrt{AT}\right)$,
- Per epoch computation reduction $\mathcal{O}(Y^3) \rightarrow \mathcal{O}(X^3)$,
- Linearly reducing the number of epochs.

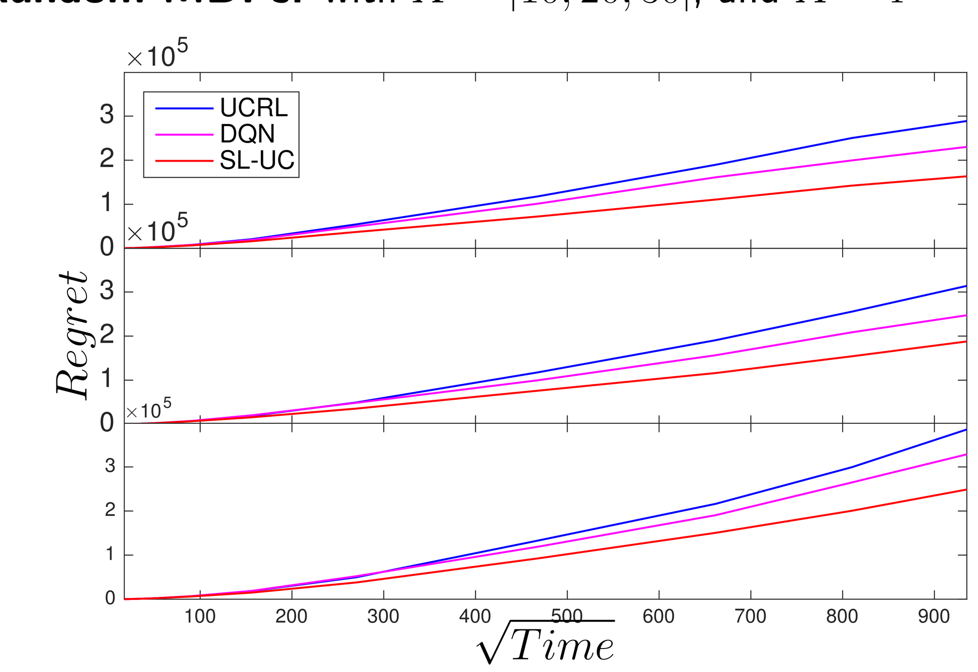
Clustering rate: A random ROMDP $(X = 4, A = 4, Y = 20)$



Random ROMDPs: $X = 5$, $A = 4$, and $[Y = 10, 20, 30]$



Random MDPs: with $X = [10, 20, 30]$, and $A = 4$



- UCRL: The Optimal algorithm
- DQN: 3 hidden layers, 30 hyperbolic tangent unites at each layer with RMSprop update

Gridworld[Johnsonet al., 2016]

*University of California, Irvine (UCI) †Institut National de Recherche en Informatique et en Automatique, (Inria)

Email: kazizzad@uci.edu

WWW: http://newport.eecs.uci.edu/anandkumar/#lab