On Oracle-Efficient PAC Reinforcement Learning with Rich Observations

Christoph Dann
Carnegie Mellon University

Nan Jiang
ILLINOIS

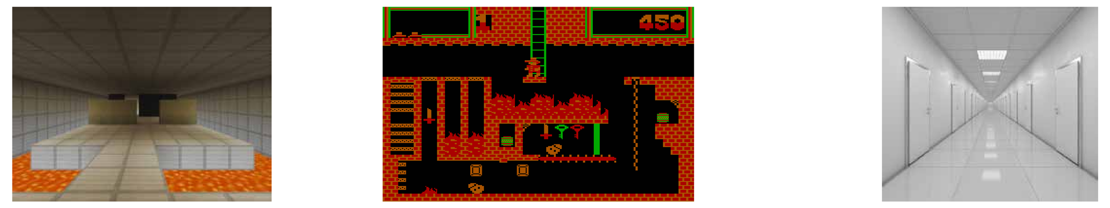
Akshay Krishnamurthy
Microsoft Research

Alekh Agarwal
Microsoft Research

John Langford
Microsoft Research

Robert E. Schapire
Microsoft Research

Motivation



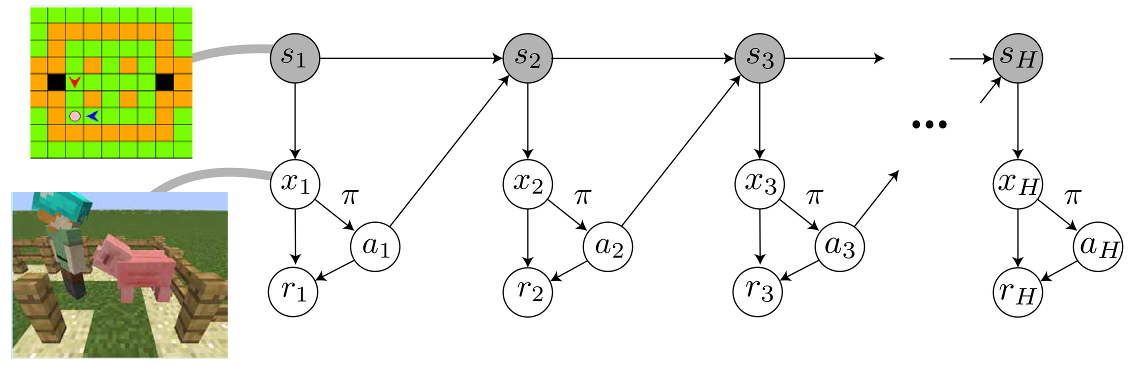
Algorithm	*Provably Sample Efficient*	*Handles Rich Observations*	*Computationally Tractable*	*Stochastic Environments*
UCRL / RMax / UCBVI / UBEV/ PSRL	✓	✗	✓	✓
Deep RL with ε-greedy (DQN, Policy Gradient...)	✗	✓	✓	✓
LSVEE	✓	✓	✗	✓
OLIVE	✓	✓	✗	✓
OCP	✓	✓	✓	✗

✓ : requires deterministic latent state transitions

Is there a reinforcement learning algorithm that can be implemented in polynomial time and learn efficiently and reliably with rich stochastic observations?

VALOR (this work)	✓	✓	✓	✓

Setting: Contextual Decision Processes with Deterministic Hidden State Transitions



Markov / Reactivity: previous hidden state identifiable from observation but mapping $x \mapsto s$ unkown

Deterministic transitions: hidden state is a deterministic function of previous state and action (but observations are stochastic!)

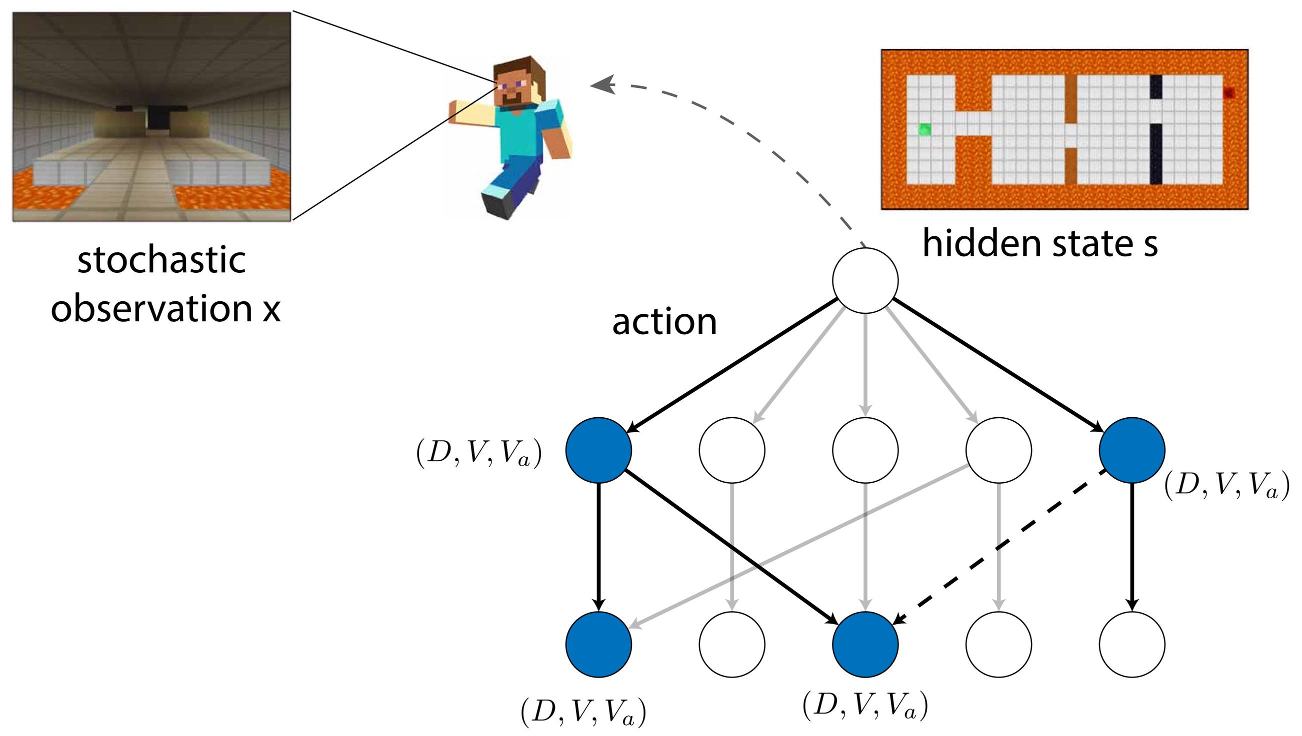
Optimal value function: reward to go for every state / observation when acting optimally

$$V^\star(x) = \max_{a \in \mathcal{A}} \mathbb{E}[r_h + V^\star(s_{h+1}) | x_h = x]$$
$$V^\star(s) = \mathbb{E}[V^\star(x_h) | s_h = s]$$

New Algorithm: VALOR

Main Ideas of VALOR (VAlues stored LOcally for RL):

- **Learn values of hidden states by depth-first search** (as in LSVEE)
- **Store values** and observation distributions of hidden states **explicitly**
- **Prune search tree** by checking consensus among all value functions that agree with stored values



Function Approximation and Oracles

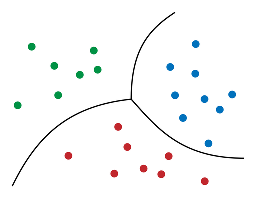
Learning with realizable function classes: to handle rich observations (e.g. images), our algorithm assumes access to:

- Class of policies $\Pi \subseteq \mathcal{X} \to \mathcal{A}$, with $\pi^\star \in \Pi$
- Class of value functions $\mathcal{G} \subseteq \mathcal{X} \to \mathbb{R}$, with $V^\star \in \mathcal{G}$

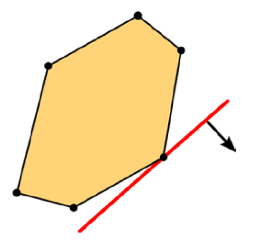
Access function classes only through standard optimization oracles for computational tractability:

can be reduced to binary classification!

- **Cost-Sensitive Classification (CSC)** oracle on policies
  Given sequence $(x^{(i)}, c^{(i)})_{i \in [n]}$ of observations $x^{(i)} \in \mathcal{X}$ and cost $c^{(i)} \in \mathbb{R}^K$ return a policy with approximately minimal average cost
  $$\min_{\pi \in \Pi} n^{-1} \sum_{i=1}^{n} c^{(i)}(\pi(x^{(i)}))$$



- **Linear Programming (LP)** oracle on value functions
  Given objective $o(g)$ and constraints $h_j(g)$ linear in g, that is, of the form $\sum_{i=1}^n \alpha_i g(x^i)$, return a value function that approximately optimizes
  $$\arg\max_{g \in \mathcal{G}} o(g) \quad \text{such that} \quad h_j(g) \le c_j \quad \forall j$$



Learning values in depth-first manner :

```
dfslearn(path p):
    for all actions a:
```

Solve: $V_{opt} = \max_{g \in \mathcal{G}} \mathbb{E}_{p \circ a}[g(x)] \qquad V_{pes} = \min_{g \in \mathcal{G}} \mathbb{E}_{p \circ a}[g(x)]$ — Linear Programs

such that: $\forall (D, V) \in \mathcal{D}_{h+1} : \; V - \phi_{h+1} \le \mathbb{E}_D[g(x)] \le V + \phi_{h+1}$

`If` **consensus** ($V_{opt} - V_{pes}$ small):

$V_a = \frac{V_{opt} + V_{pes}}{2}$ — Successor already learned

`Else:`

$V_a =$ `dfslearn(p∘a)` — Successor not learned yet

Learn p by storing a new entry in $\mathcal{D}$ with:

1. Data set of observations, uniformly chosen actions and immediate rewards at current path $(x, a, r) \sim D$
2. Values of all successor states $\{V_a\}_{a \in \mathcal{A}}$
3. Value of hidden state $V \leftarrow \max_{\pi \in \Pi} \mathbb{E}_D[(r + V_a)\mathbf{1}\{\pi(x) = a\}]$ — Cost-Sensitive Classification

Return V

Compute global policy from learned values:

$$\hat{\pi} \leftarrow \arg\max_{\pi \in \Pi} \sum_{(D, V, V_a) \in \mathcal{D}} \mathbb{E}_D[(r + V_a)\mathbf{1}\{\pi(x) = a\}]$$

Cost-Sensitive Classification

Theoretical Analysis of VALOR

Sample Efficiency:

Theorem: If $\pi^\star \in \Pi$ and $V^\star \in \mathcal{G}$, VALOR returns an ε-optimal policy with probability at least $1 - \delta$ after collecting at most

#hidden states / epsiode length / #actions

$$\tilde{O}\left(\frac{M^3 H^8 K}{\epsilon^5} \log(|\mathcal{G}||\Pi|/\delta)\right) \text{ trajectories.}$$

Oracle Efficiency:

Theorem: If $\pi^\star \in \Pi$ and $V^\star \in \mathcal{G}$, VALOR is oracle efficient with probability at least $1 - \delta$, that is, it can be implemented with at most

$$O\left(\frac{MKH^2}{\epsilon} \log \frac{MH}{\delta}\right) \text{ Linear Program (LS) oracle calls and}$$

$$O\left(\frac{MH^2}{\epsilon} \log \frac{MH}{\delta}\right) \text{ Cost-sensitive Classification (CSC) oracle calls,}$$

each of which only needs to be accurate up $\epsilon_{sub} = O\left(\frac{\epsilon^2}{MH^3}\right)$.

Additional Result: OLIVE is Oracle-Inefficient

OLIVE is the only algorithm that is known to be provably sample-efficient in contextual decision processes with stochastic state transitions.

Theorem: OLIVE is not oracle-efficient, that is, it cannot be implemented with polynomially many calls to LP, CSC and least-squares oracles.

After having collected data sets $D_1, \dots D_k \in \mathcal{D}$ of transitions with previous policies, OLIVE chooses the next policy to execute by solving:

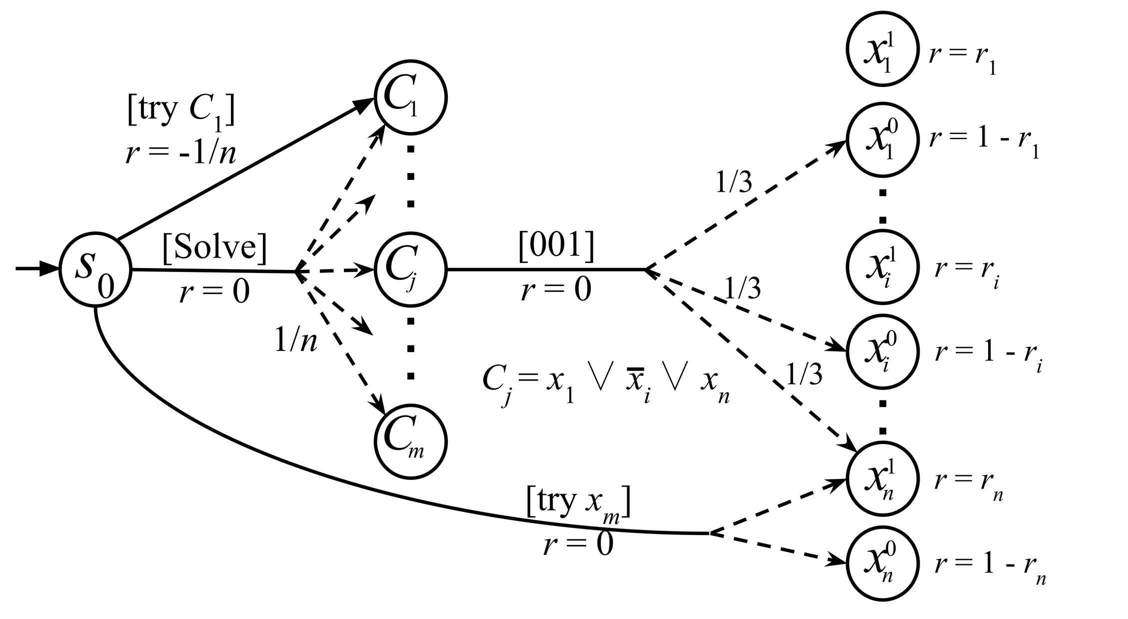
Value of initial observations

$$\hat{g}_k, \hat{\pi}_k = \operatorname*{argmax}_{g \in \mathcal{G}, \pi \in \Pi} \hat{\mathbf{E}}_{D_0}[g(x)] \quad \text{s.t.} \tag{1}$$
$$\underbrace{\forall\, D_i \in \mathcal{D}:}_{\text{On all data so far}} |\hat{\mathbf{E}}_{D_i}[\mathbf{1}\underbrace{\{a = \pi(x)\}}_{\text{For chosen policy}}\underbrace{(g(x) - r - g(x'))}_{\text{Bellman residual}}]| \le \phi$$

has to be small

Reducing 3-SAT to OLIVE:

Given 3-SAT instance with variables $x_1, \dots x_n$, and clauses $C_1, \dots, C_m$ of the form $C_j = \bar{x}_k \lor x_i \lor x_l$, construct MDPs $M \in \mathcal{M}$ determined up to terminal rewards



$\hat{\pi}_k(s_0) =$ `[Solve]` $\Leftrightarrow \max_{M \in \mathcal{M}, \pi \in \Pi} V^\pi(s_0) = 1 \Leftrightarrow$ formula satisfiable

References

- Jiang, Nan, Krishnamurthy, Akshay, Agarwal, Alekh, Langford, John, and Schapire, Robert E. Contextual decision processes with low Bellman rank are PAC-learnable [ICML 17]
- Krishnamurthy, Akshay, Agarwal, Alekh, and Langford, John. PAC reinforcement learning with rich observations [NIPS 16]
- Wen, Zheng and Van Roy, Benjamin. Efficient exploration and value function generalization in deterministic systems [NIPS 13]
- Azar, Mohammad Gheshlaghi, Osband, Ian, and Munos, Remi. Minimax regret bounds for reinforcement learning [ICML 17]
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A., Veness, Joel, Bellemare, Marc G., Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K., Ostrovski, Georg, Petersen, Stig, Beattie, Charles, Sadik, Amir, Antonoglou, Ioannis, King, Helen, Kumaran, Dharshan, Wierstra, Daan, Legg, Shane, and Hassabis, Demis. Human-level control through deep rein- forcement learning [Nature 15]