Agnostic KWIK learning and efficient approximate reinforcement learning

István Szita    Csaba Szepesvári

Department of Computing Science
University of Alberta

Annual Conference on Learning Theory, 2011
Outline

1 Basic concepts
   - Efficient reinforcement learning
   - The “Knows what it knows” (KWIK) framework

2 Agnostic KWIK learning
   - Definitions
   - Results for several problem classes

3 Summary
Reinforcement learning

- Maximize long-term reward
- but environment is unknown
- agent needs to explore, but exploration is costly
Efficient RL algorithms

- make bounded amount of non-optimal steps\(^1\)
- balance exploration and exploitation
- exist for many environment classes (e.g. MDPs)

\(^{1}\)alternative definitions exist
The “Rmax-construction”:
A general scheme for efficient RL

- keep track of “known” areas → KWIK learner
- assume that unknown areas have maximum reward
- plan optimal path within the known area
- collect new experience when leaving known area
The “Knows what it knows” (KWIK) framework

[Li, Walsh, Littman, 2008]

- Adversary picks a concept
- repeat:
  - Adversary picks query $x$
  - if Learner passes,
    - Adversary gives noisy feedback
    - Learner updates itself
  - if Learner predicts,
    - it has to be accurate
    - otherwise it fails
The Rmax construction with a KWIK learner

\[ \text{KWIK-Rmax}(\text{MDPLearner}, \text{Planner}) \]
\[ \text{MDPLearner}.initialize(...) \]
\[ \text{Planner}.initialize(...) \]
\[ \text{Observe } s_1 \]
\[ \text{for } t := 1, 2, \ldots \text{ do} \]
\[ \quad a_t = \text{Planner.plan}(\text{OPT}(\text{MDPLearner}), s_t) \]
\[ \quad \text{Execute } a_t \text{ and observe } s_{t+1}, r_t \]
\[ \quad \text{if } \text{MDPLearner}.predict(s_t, a_t) = \bot \text{ then} \]
\[ \quad \text{MDPLearner}.learn((s_t, a_t), (\delta_{s_{t+1}}, r_t)) \]

\{Optimistic Wrapper\}
\[ \text{Opt}(\text{MDPLearner}).predict(s, a) \]
\[ \quad \text{if } \text{MDPLearner}.predict(s, a) = \bot \text{ then} \]
\[ \quad \quad \text{return } (\delta_s(\cdot), (1 - \gamma)V_{\max}) \]
\[ \quad \text{else} \]
\[ \quad \quad \text{return } \text{MDPLearner}.predict(s, a) \]
The KWIK-Rmax theorem

[Li, Walsh, Littman, 2008]

Let $G$ be a class of environment models. (e.g. the class of MDPs, factored MDPs, linear MDPs). If we have

- An efficient KWIK-learner for class $G$
- A near-optimal planner for models in $G$

then the KWIK-Rmax algorithm constructed from these is an efficient reinforcement learner on $G$.

but what if the environment is not contained in the class $G$?
The need for agnostic learning

In reinforcement learning, we often need to
- environment is almost a factored MDP, but modeled as an FMDP
- state abstraction (e.g., aggregation) is used, but MDP is uncompressible
- function approximation is used

In such cases, we should not assume that we know the class $\mathcal{G}$ of the environment. We should be agnostic!

“Agnostic” = no knowledge of where the adversary chooses its concept from
Agnostic KWIK learning

- agent does not know the problem class $\mathcal{G}$
- it chooses from another class $\mathcal{H}$
- we assume that an upper bound on their distance is known:

$$D \geq \Delta(\mathcal{G}, \mathcal{H}) \overset{\text{def}}{=} \sup_{(X,Y,g,Z) \in \mathcal{G}} \inf_{h \in \mathcal{H}} \|h - g\|_{\infty}.$$
we cannot guarantee $\epsilon$ accuracy (of course)

interestingly, we cannot guarantee $D + \epsilon$

we require $r \cdot D + \epsilon$
  ▶ $r \geq 1$ is the competitiveness factor
Problems and problem classes

**Definition (Problem)**

A **problem** is a 5-tuple \( G = (\mathcal{X}, \mathcal{Y}, g, Z, \| \cdot \|) \), where

- \( \mathcal{X} \) is the set of inputs,
- \( \mathcal{Y} \subseteq \mathbb{R}^d \) is a measurable set of possible responses,
- \( Z : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y}) \) is the noise distribution (zero-mean)
- \( \| \cdot \| : \mathbb{R}^d \rightarrow \mathbb{R}_+ \) is a semi-norm on \( \mathbb{R}^d \).

**Definition (Problem class)**

A **problem class** \( \mathcal{G} \) is a set of problems.
Agnostic KWIK learner

- $D > 0$: approximation error bound
- $r \geq 1$: competitiveness factor
- $\epsilon \geq 0$: accuracy slack
- $\delta \geq 0$: confidence parameter

A learning agent is agnostic KWIK for $(\epsilon, \delta, r, D)$ if outside of an event of probability at most $\delta$, it holds that

- when it predicts, error is $\leq r \cdot D + \epsilon$
- \# of passes is bounded

Complexity: \# of passes $= f(\epsilon, \delta, D, r)$
Agnostic KWIK-Rmax theorem

Fix $\epsilon > 0$, $r \geq 1$, $0 < \delta \leq 1/2$. If we have

- an $(rD + \epsilon)$-accurate agnostic KWIK learner, with complexity bound $B(\delta)$, and
- a $e_{\text{planner}}$-accurate planner,

then with prob. $1 - 2\delta$, the KWIK-Rmax algorithm makes

$$O \left( \frac{V_{\max}(1 - \gamma)L}{rD + \epsilon} \left\{ B(\delta) + \log \left( \frac{L}{\delta} \right) \right\} \right)$$

mistakes larger than $\frac{5(rD + \epsilon)}{1 - \gamma} + e_{\text{planner}}$, where

$$L = O((1 - \gamma)^{-1} \log(V_{\max}(1 - \gamma)/(rD + \epsilon)))$$

is the $rD + \epsilon$-horizon time.
The agnostic KWIK-Rmax theorem justifies the agnostic KWIK framework!

.. but what can we “agnostic KWIK” learn?
Finite hypothesis class $\mathcal{H}$, deterministic case

- Learner is given $D$ and the hypotheses $f_1, \ldots, f_{|\mathcal{H}|}$;
- does not know the true concept $g$
- for each query $x$, see if there is a prediction $y$ such that $|y - f_i(x)| \leq D$ for all $i$
- if yes, then $y$ is a good prediction! ($2D$-accurate)
- if not, then we have to pass
  - and receive $g(x)$
  - $|y - f_i(x)| > D$ for at least one $f_i$
  - so we can exclude it
Finite hypothesis class $\mathcal{H}$, deterministic case

The previous algorithm
- passes at most $|\mathcal{H}| - 1$ times (for each “i don’t know”, it excludes at least one hypothesis)
- gives $2D$-accurate predictions ($r = 2, \epsilon = 0$)
A sample run of the agnostic KWIK learner
solution is not trivial:

- We cannot exclude a hypothesis by a single sample. We need to take averages.
- If $\sum (y_t - f(x_t))$ is small, $f$ may be still bad (adversary selects over- and underestimating places alternately)
- If $\sum (y_t - f(x_t))$ is large, $f$ is definitely bad
  - but the adversary can prevent us from seeing such a case (for every 1000 small-error $x_t$ it gives one large-error one)
Finite hypothesis class $\mathcal{H}$, noisy problems

If $f_1 < f_2 + 2D$ on some region, then sample average in that region is much closer to one of them. The other one can be excluded.

$\mathbf{f_1} > \mathbf{f_2}$
Finite hypothesis class \( \mathcal{H} \), noisy problems

Algorithm:

- keep a bag of samples for each \( f_i, f_j \)
- for each query \( x \), see if there is a prediction \( y \) such that 
  \[ |y - f_i(x)| < D + \epsilon/2 \] for all \( i \)
- if yes, then \( y \) is a good prediction! (\( 2D + \epsilon \)-accurate)
- if not, then we have to pass
  - and receive \( y' = g(x) + \text{noise} \)
  - \( f_i(x) \ll f_j(x) \) for at least one \( f_i, f_j \)
  - add \((x, y')\) to the corresponding bag
- if \( m \) samples gathered in a bag, calculate sample average
  - one hypothesis can be excluded
<table>
<thead>
<tr>
<th>Hypothesis class</th>
<th>Approx.</th>
<th>Agnostic KWIK</th>
<th>KWIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finite, deterministic</td>
<td>$2D$</td>
<td>$N - 1$</td>
<td>$N - 1$</td>
</tr>
<tr>
<td>Finite, noisy</td>
<td>$2D + \epsilon$</td>
<td>$O \left( \frac{N^2}{\epsilon^2} \log \frac{N}{\delta} \right)$</td>
<td>$O \left( \frac{N}{\epsilon^2} \log \frac{N}{\delta} \right)$</td>
</tr>
<tr>
<td>$d$-dim linear, deterministic</td>
<td>$2D + \epsilon$</td>
<td>$O \left( d! \left( \frac{1}{\epsilon} + 1 \right)^d \right)$</td>
<td>$d + 1$</td>
</tr>
<tr>
<td></td>
<td>$2D$</td>
<td>$\Omega(2^d)$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$d$-dim linear, noisy</td>
<td>$2D + \epsilon$</td>
<td>$O \left( \frac{1}{\epsilon^{2d+2}} \log \frac{1}{\delta \epsilon^d} \right)$</td>
<td>$O \left( \frac{d^3}{\epsilon^4} \log \frac{1}{\delta \epsilon} \right)$</td>
</tr>
</tbody>
</table>
Summary

Agnostic KWIK learning...
- is a new online learning framework
- can be applied to efficient reinforcement learning with non-exact models
- is generally much harder than ordinary KWIK
- proofs and examples in the paper

Open problems:
- agnostic KWIK learner for transition probabilities (essential for agnostic learning of MDPs)
- How to do agnostic RL more efficiently, \textit{without} agnostic KWIK (agnostic KWIK is too restrictive)