Class-based Upper Confidence Reinforcement Learning

Mahsa Asadi, Odalric-Ambrym Maillard

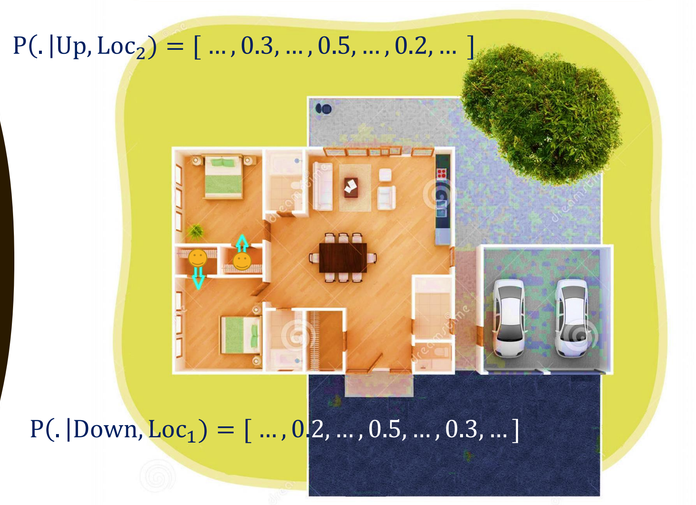
Shiraz University, Inria Lille – Nord Europe

Introduction

Problem: Regret minimization for never-ending single trajectory Reinforcement Learning(RL) problems with unknown MDP structure.

UCRL: To solve the above problem, UCRL tries to estimate MDP dynamics using optimistic and statistically correct models.

Improve UCRL[1] regret bound of $O(DS\sqrt{AT})$ by state action pair aggregation of pairs with similar profile distribution.



https://www.dreamstime.com/royalty-free-stock-photo-house-plan-top-view-interior-cross-section-image38325285

Classes and Mappings

Similar state-action pairs: (s', a') and (s, a) pairs are ϵ −similar for $\epsilon = (\epsilon_p, \epsilon_\mu) \in \mathbb{R}_+^2$ if they have:

- Similar profiles:

$$\left|p(\sigma_{s,a}(.)|s,a) - p(\sigma_{s',a'}(.)|s',a')\right|_1 \leq \epsilon_p$$

- Similar rewards:

$$|\mu(s,a) - \mu(s',a')|_1 \leq \epsilon_r$$

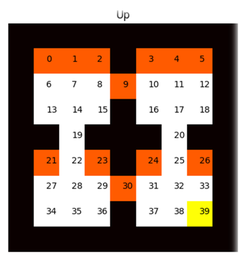
Class(C): 0-Similar pairs are grouped as one class.

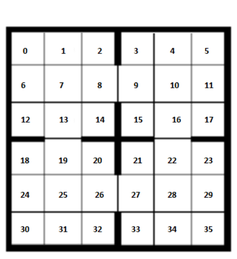
Mapping(σ): an ordering of profile distribution elements s.t.:

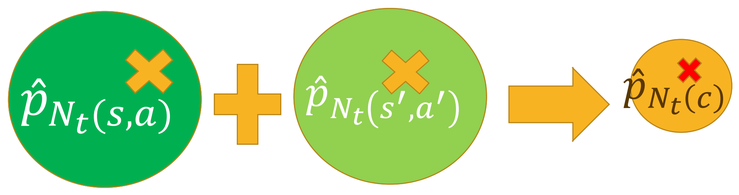
$\sigma_{s.a} : \{1, \dots, S\} \to \sigma_{s,a}\ s.t.\ \sigma_{s,a}(1) \geq \sigma_{s,a}(2) \geq \cdots \geq \sigma_{s,a}(S)$

Algorithms	C	σ
C-UCRL(C, σ)	● Known	● Known
C-UCRL(C)	● Known	Empirical
C-UCRL	Clustering	Empirical

Different Class-based UCRL settings







Confidence bounds and Aggregating samples of two different pairs of same class

C-UCRL(σ, C)

Modified Transition probability estimate:

$$\hat{p}^{\sigma}_{N_t(c)}(.|c) = \frac{\sum_{(s,a)\in c} N_t(s,a) p_{N_t(s,a)}(\sigma_{s,a}(i)|s,a)}{N_t(c)}$$

C-UCRL(C)

Modified Transition probability estimate:

$$\hat{p}^{\hat{\sigma}}_{N_t(c)}(.|c) = \frac{\sum_{(s,a)\in c} N_t(s,a) p_{N_t(s,a)}(\hat{\sigma}_{s,a}(i)|s,a)}{N_t(c)}$$

Non-expensive ordering lemma:

$$\left|p_n(\sigma_n(.)) - p(\sigma(.))\right|_1 \leq \left|p_n(\sigma(.)) - p(\sigma(.))\right|_1$$

C-UCRL

Two cluster centers are merged if with high probability, the samples fall in the statistically correct bound. →See **Confident clustering Algorithm**.

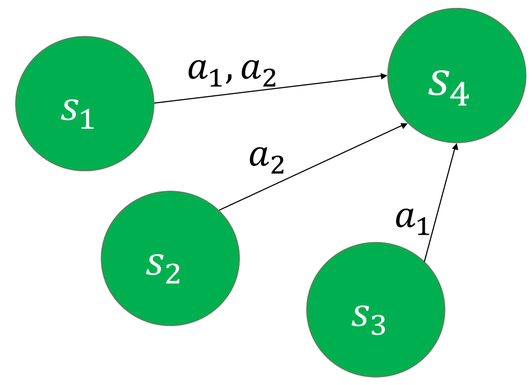
Confident Clustering Algorithm

$C \leftarrow [p^{(1)}, \dots, p^{(S)}]$ {Each sample is it's own cluster center}
$N \leftarrow [n_1, \dots, n_S]$
$size \leftarrow [1, \dots, 1]$ {S-element array of one}
$Changed \leftarrow \mathbf{True}$
While not Converged **and** Changed **do**
  $Changed \leftarrow \mathbf{False}$
  $Ordering \leftarrow argsort(N)$
  for all $i \in Ordering$ **do**
    if $n_i = 0$ **then**
      break
    end if
    $k \leftarrow \mathbf{Near}(i, C)$ {Find the closest cluster to i}
    if $k = -1$ **then**
      continue
    end if
    merge(k, I, C, N, size)
    $Changed \leftarrow \mathbf{True}$
  end for
end while

Stopping Criterion

Instead of the heuristic doubling approach to determine the end of each episode we use statistical testing to check whether our optimistic MDP is still consistent with the observations at subsequent time steps for:

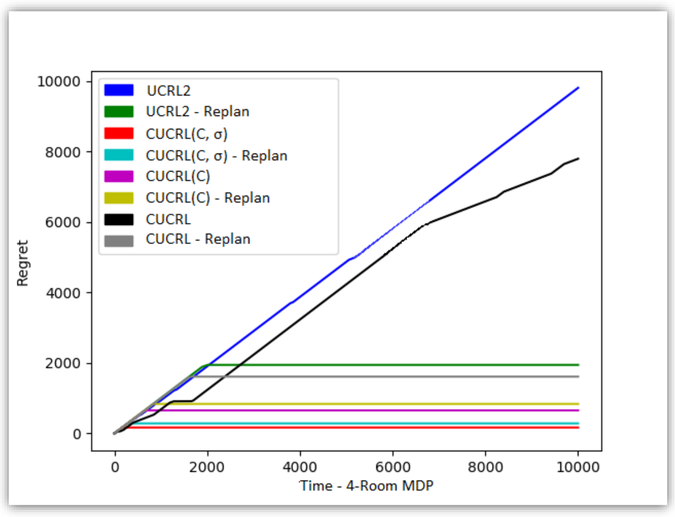
1. $p(s_{t+1}|s_t, a_t)$
2. $p(.|s_t, a_t)$
3. Target state check



Case 3: most of the state-action pairs point to the same state(target)

Empirical results

* One can note that C-UCRL outperforms UCRL in all cases.
* Replan represents the novel stopping criterion.



Comparing different approaches on 4-room MDP problem

Contributions

1. Novel notion of similarity for discrete RL setup
2. Improved regret bound to $O(D\sqrt{KCT})$ for UCRL(C,σ) and UCRL(C) where C:=#Classes, K := transition support size
3. Novel statistically sound clustering algorithm for CUCRL
4. A novel stopping criterion based on statistical testing