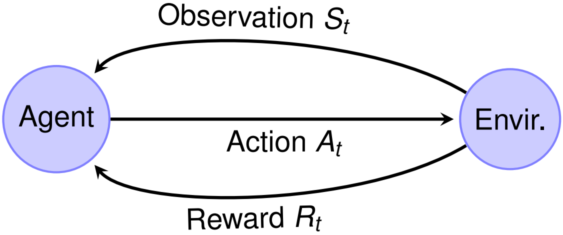
Optimism in Reinforcement Learning and Kullback-Leibler Divergence

Sarah Filippi, Olivier Cappé, Aurélien Garivier

Our Goal: Model-Based Online Reinforcement Learning

Assuming a finite state-space finite action-space Markov Decision Process (MDP)



with unknown

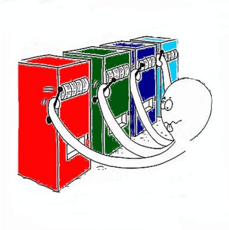
Transition $P(s'; s, a) = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$

Reward $r(s, a) = \mathbb{E}(R_t | S_t = s, A_t = a)$

- Implement an on-policy strategy for controlling the agent
- Doing “almost as good” (in terms of cumulated rewards) as an oracle agent that knows the optimal policy

A Simpler Case: Multi-Armed Bandit Models

In Multi-Armed Bandits (MAB), there is only one state s_0



- Actions A_t do not influence the state of the environment
- The reward $\mu(a) = r(s_0, a)$ is the only unknown

The oracle agent always plays an action a^* that has largest expected reward

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}} \mu(a)$$

and the loss wrt. the oracle agent can be measured by the regret

$$\text{Regret}(n) = \sum_{t=1}^{n} \mu(a^*) - \mu(A_t)$$

Upper Confidence Bound (UCB)

In MAB problems, the optimism in the face of uncertainty heuristic [Lai & Robins, 85; Agrawal, 95] has been very successful

The UCB (Upper Confidence Bound) algorithm [Auer *et al*, 02] plays the action A_t such that

$$A_t = \underset{a \in \mathrm{A}}{\operatorname{argmax}} \underbrace{\hat{\mu}_t(a)}_{\text{greedy action}} + \underbrace{\sqrt{\frac{\alpha \log(t)}{N_t(a)}}}_{\text{exploration bonus}}$$

- $N_t(a)$ is the number of times arm a has been played before time t
- $\hat{\mu}_t(a) = N_t(a)^{-1} \sum_{i=1}^{t-1} R_i \mathbb{1}\{A_i = a\}$ is the empirical estimate of $\mu(a)$

It achieves an expected regret that only grows logarithmically with n

Upper Confidence Reinforcement learning

In MDPs, [Auer et al, 07–10; Tewari & Bartlett, 07–08] propose to replace the upper confidence bound of UCB by an optimistic MDP (P^*, r^*) whose average reward $\rho^* = \lim n^{-1} \sum_{t=0}^{n-1} \mathbb{E}_{\pi^*}(R_t)$ and bias vector $h^*(s)$ satisfy an extended version of Bellman's optimality equations

$$\forall s,\ h^*(s) + \rho^* = \max_{P,r \in \mathcal{C}_t^P \times \mathcal{C}_t^r} \max_{a \in \mathrm{A}} \left(r(s,a) + \sum_{s' \in \mathrm{S}} P(s';s,a) h^*(s') \right)$$

$$\forall s,\ \pi^*(s) = \operatorname*{argmax}_{a \in \mathrm{A}} \left(r^*(s,a) + \sum_{s' \in \mathrm{S}} P^*(s';s,a) h^*(s') \right)$$

where $\mathcal{C}_t^R$ and $\mathcal{C}_t^r$ are confidence sets for P and r, respectively

Extended Value Iteration

The bias vector h^* is determined (up to a constant) as the limit of extended value iterations

- While $\mathrm{span}(V_{k+1} - V_k) > \varepsilon$,

$$\forall s\ ,\ \ V_{k+1}(s) = \max_{a \in \mathrm{A}} \left(\max_{r \in \mathcal{C}_t^r} r(s,a) + \max_{P \in \mathcal{C}_t^P} \sum_{s' \in \mathrm{S}} P(s'; s,a) V_k(s') \right)$$

Issues Not Discussed Here

Influence of the termination tolerance ε Ignored in our work, analyzed in [Auer et al, 07–10]

Convergence of extended value iterations Considered (for L^1 neighborhoods) by [Auer et al, 07–10; Tewari & Bartlett, 07–08] and for KL neighborhoods in the discounted case by [Nilim & EL Ghaoui, 05]

Persistence of policies In MDPs it is not possible to continuously change the policy as in MABs. We used the episodic construction of [Auer et al, 07–10] in which the optimistic policy is recomputed at times that approximately follow a geometric progression with ratio 2

Definition of the Confidence Set

[Auer et al, 07–10; Tewari & Bartlett, 07–08] consider rectangular confidence sets of the form

$$\forall (s,a),\ \left\| \hat{P}_t(.;s,a) - P(.;s,a) \right\|_1 \leq \delta_P$$
$$\forall (s,a),\ |\hat{r}_t(s,a) - r(s,a)| \leq \delta_R$$

where $\hat{P}_t(s';s,a) = N_t(s,a)^{-1} \sum_{i=0}^{t-1} \mathbb{1}\{S_{i+1} = s', S_i = s, A_i = a\}$ and $\hat{r}_t(s,a) = N_t(s,a)^{-1} \sum_{i=0}^{t-1} R_i \mathbb{1}\{A_i = a\}$ are the empirical estimates of P and r at time t

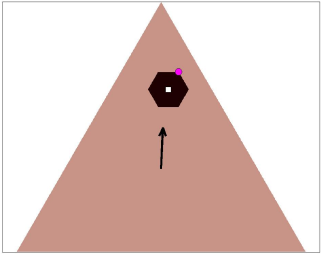
The probabilities of violating the confidence sets are controlled by the Hoeffding inequality for $\hat{r}_t(s,a)$ and by the bound of [Weissman *et al*, 03] for $\hat{P}(.;s,a)$

How Does L^1 Extended Value Iteration Operates?

For each state and action pair, one must solve a problem of the form

$$q^* = \underset{q:\|p-q\|_1 \leq \delta}{\operatorname{argmax}} \; q' V$$

where p is the empirical estimate of the transition probabilities and V is the current estimate of the bias vector



- inflate p_i (if possible) by a total amount of δ for indices i that maximize V_i
- reduce p_i (as much as needed) for indices i where V_i is the smallest

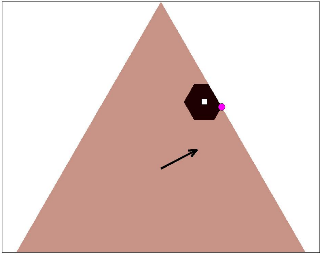
$\Rightarrow$ easy both to implement an interpret, but...

How Does L^1 Extended Value Iteration Operates?

For each state and action pair, one must solve a problem of the form

$$q^* = \underset{q:\|p-q\|_1 \le \delta}{\text{argmax}} \; q' V$$

where p is the empirical estimate of the transition probabilities and V is the current estimate of the bias vector

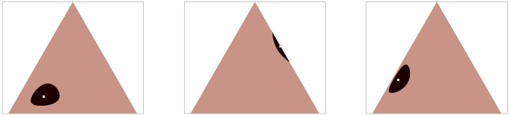


- inflate p_i (if possible) by a total amount of δ for indices i that maximize V_i
- reduce p_i (as much as needed) for indices i where V_i is the smallest

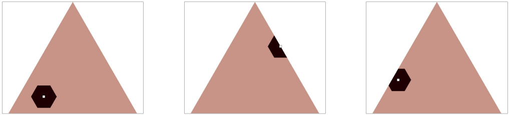
$\Rightarrow$ easy both to implement an interpret, but...

Our Proposal: Kullback-Leibler URCL

The role played by the KL divergence in large deviations of multinomial experiments suggests that the proper confidence neighborhoods are



rather than

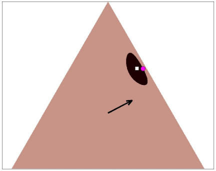


$\Rightarrow$ Use KL rather than L^1 constraints!

Solving KL-Extended Value Maximization

For each state-action pair, one must solve a linear program under KL constraint

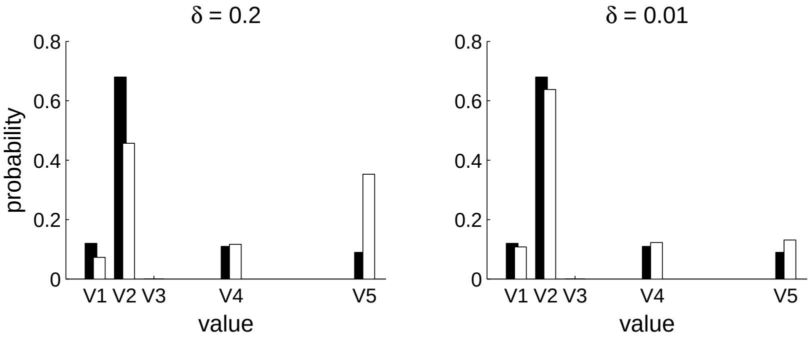
$$q^* = \underset{q : KL(p;q) \leq \delta}{\operatorname{argmax}} \; q' V$$



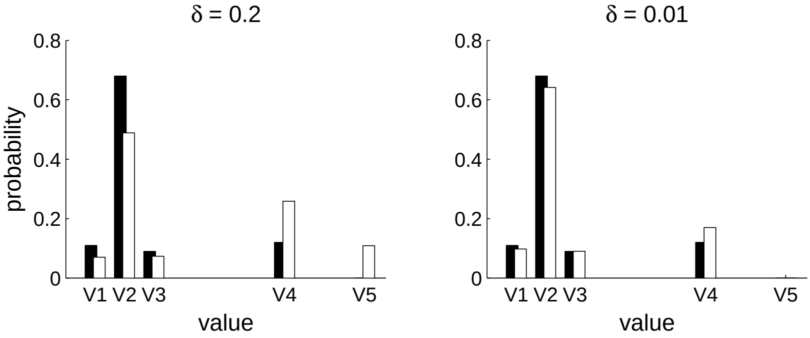
The solution is given by an explicit non-linear transformation of p which is fully controlled by the solution ν to the equation $f(\nu) = \delta$, where f is the one-dimensional decreasing stricly convex function on $(\max_{i:p_i>0} V_i, \infty)$ defined by

$$f(\nu) = \sum_i p_i \log(\nu - V_i) + \log\left(\sum_i \frac{p_i}{\nu - V_i}\right)$$

KL-LP's Rule I
"Bigger rewards gets more likely"



KL-LP's Rule II
"You can't get to heaven when δ is too small"



Regret Bound

Adapting the proof of [Auer et al, 07–10] it is possible to show that KL-UCRL achieves logarithmic regret in communicating MDPs (as does UCRL)

Main arguments of the proof

- Pinsker's inequality $\|p - q\|_1 \leq \sqrt{2KL(p; q)}$
- Bound of [Garivier & Leonardi, 10]

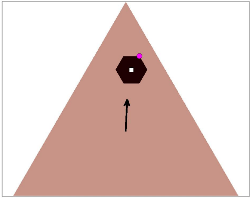
$$\mathbb{P}\left(\forall t \leq n,\ KL(\hat{p}_t; p) > \frac{\delta}{t}\right) \leq 2e(\delta \log(n) + |\mathrm{S}|)e^{-\delta/|\mathrm{S}|}$$

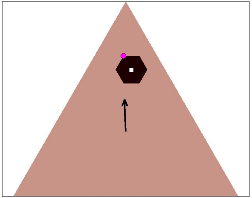
$\square$

In simulations however (benchmark and random sparsely connected environments), KL-UCRL performs significantly better than UCRL

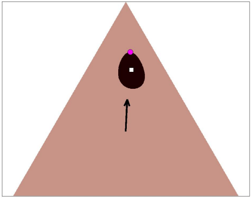
Discussion: Continuity of the optimistic MDP

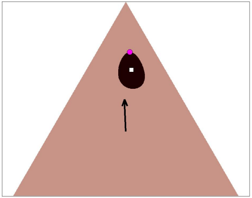
L^1 Neighborhoods





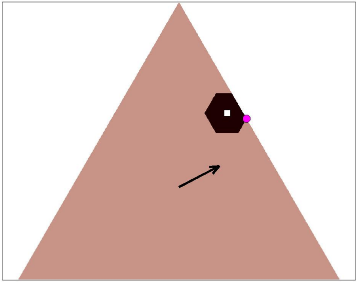
KL Neighborhoods



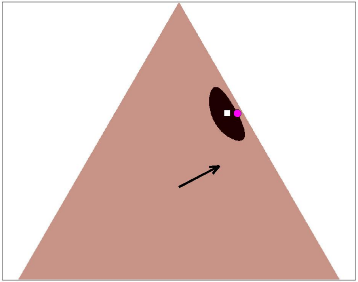


Discussion: Compatibility with observed transitions

L^1 Neighborhoods

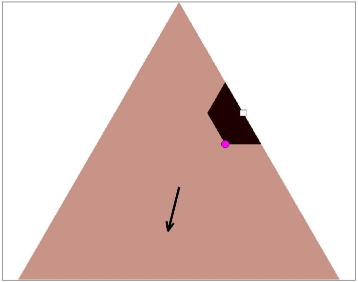


KL Neighborhoods

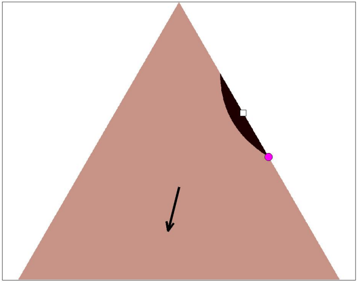


Discussion: Tradeoff between the attraction towards the best state and the statistical evidence that it may not be reachable from all states

L^1 Neighborhoods



KL Neighborhoods



Thank you!