

Analysis of Optimistic Algorithms for the Exploration/Exploitation Trade-Off

Peter Auer

auer@unileoben.ac.at

University of Leoben, Austria

Hong Kong, June 2008

- ▶ Decision making in the face of uncertainty — and its regret
- ▶ Exploration versus exploitation
- ▶ Optimistic algorithms and their analysis
- ▶ Stochastic bandit problem
- ▶ On-line reinforcement learning

On-line decision making

- ▶ In each step $t = 1, \dots, T$
 - ▶ the agent makes an observation o_t ,
 - ▶ chooses an action a_t ,
 - ▶ and receives some reward r_t .
- ▶ Actions are chosen according to some policy π ,
 $\pi(o_1, a_1, r_1, \dots, o_t) = a_t$.
- ▶ Observations and rewards are generated by an unknown environment M , which might change its state depending on the actions chosen,

$$\begin{aligned}M_t(s_t) &= o_t, \\M_t(s_t, a_t) &= (r_t, s_{t+1}).\end{aligned}$$

Goal of the agent

- ▶ The agent tries to maximize its total reward, without knowing the environment M :

$$R_M^\pi(T) = \sum_{t=1}^T r_t^\pi,$$

where r_t^π are the rewards received by policy π in environment M .

Goal of the agent

- ▶ The agent tries to maximize its total reward, without knowing the environment M :

$$R_M^\pi(T) = \sum_{t=1}^T r_t^\pi,$$

where r_t^π are the rewards received by policy π in environment M .

- ▶ If the environment M were known, an optimal policy $\pi^* = \pi_M^*$ could be used, with maximum expected total reward

$$\mathbb{E} \left[R_M^{\pi^*}(T) \right] = \mathbb{E} \left[\sum_{t=1}^T r_t^* \right],$$

where r_t^* are the rewards received by an optimal policy π^* for environment M .

Uncertainty and regret

- ▶ Because of the uncertainty about the environment M , a policy π suffers some regret Δ_M^π compared to an optimal policy π^* ,

$$\Delta_M^\pi = \Delta_M^{\pi^*, \pi}(T) = R_M^{\pi^*}(T) - R_M^\pi(T).$$

- ▶ We are interested in bounds on the regret of policies π ,

$$\max_{M \in \mathcal{M}, \pi^* \in \Pi} \Delta_M^{\pi^*, \pi}(T), \max_{M \in \mathcal{M}, \pi^* \in \Pi} \mathbb{E} \left[\Delta_M^{\pi^*, \pi}(T) \right].$$

Example: The stochastic bandit problem

- ▶ The environment has only a single state, therefore there are no observations.
- ▶ The agent chooses an arm $a_t \in \{1, \dots, K\}$.
- ▶ The rewards are independent and random with fixed mean for each arm, $\mathbb{E}[r_t] = \mu(a_t)$ (and $r_t \in [0, 1]$).

Example: The stochastic bandit problem

- ▶ The environment has only a single state, therefore there are no observations.
- ▶ The agent chooses an arm $a_t \in \{1, \dots, K\}$.
- ▶ The rewards are independent and random with fixed mean for each arm, $\mathbb{E}[r_t] = \mu(a_t)$ (and $r_t \in [0, 1]$).
- ▶ The optimal policy chooses the best arm, $\pi_M^* \equiv \arg \max_a \mu(a)$.

Example: Reinforcement learning

- ▶ The agent observes the state s_t of the environment from a finite set \mathcal{S} of possible states.
- ▶ The agent chooses an action a_t from a finite set \mathcal{A} .
- ▶ The environment M is any MDP (Markov Decision Process) on \mathcal{S} and \mathcal{A} with random rewards in $[0, 1]$ and means $\bar{r}(s, a)$, and transition probabilities $p(s_{t+1}|s_t, a_t)$.

Example: Reinforcement learning

- ▶ The agent observes the state s_t of the environment from a finite set \mathcal{S} of possible states.
- ▶ The agent chooses an action a_t from a finite set \mathcal{A} .
- ▶ The environment M is any MDP (Markov Decision Process) on \mathcal{S} and \mathcal{A} with random rewards in $[0, 1]$ and means $\bar{r}(s, a)$, and transition probabilities $p(s_{t+1}|s_t, a_t)$.
- ▶ A policy π is compared to an optimal policy π_M^* , $\pi_M^*(t, s_t) = a_t$.

Exploration versus exploitation

The agent faces the following dilemma:

- ▶ either it chooses an action which looks optimal given the past observations and rewards (exploitation),

Exploration versus exploitation

The agent faces the following dilemma:

- ▶ either it chooses an action which looks optimal given the past observations and rewards (exploitation),
- ▶ or it chooses an action which may reveal more information about the environment and which actions are indeed optimal (exploration).

Exploration versus exploitation

The agent faces the following dilemma:

- ▶ either it chooses an action which looks optimal given the past observations and rewards (exploitation),
- ▶ or it chooses an action which may reveal more information about the environment and which actions are indeed optimal (exploration).

E.g. for the stochastic bandit problem, this is

- ▶ either choosing the arm with largest average reward observed so far,
- ▶ or choosing a different arm which is under-explored and potentially might have even larger average reward.

Optimistic algorithms (1)

An optimistic algorithm

- ▶ assumes the best possible environment (consistent with past observations and rewards),
- ▶ and chooses optimal actions in respect to this best possible environment.

Optimistic algorithms (1)

An optimistic algorithm

- ▶ assumes the best possible environment (consistent with past observations and rewards),
- ▶ and chooses optimal actions in respect to this best possible environment.

Consistent environments:

- ▶ Let M^* be the true environment.
- ▶ Consistent environments are environments in

$$\mathcal{M}_t = \mathcal{M}(o_1, a_1, r_1, \dots, o_{t-1}, a_{t-1}, r_{t-1}, o_t) \subseteq \mathcal{M}$$

such that with high probability $M^* \in \mathcal{M}_t$,

$$\mathbb{P}\{\mathcal{M}_t : M^* \in \mathcal{M}_t\} \geq 1 - 1/T^2.$$

Optimistic algorithms (2)

The best consistent environment and the corresponding policy (in respect to expected rewards) for step t could then be chosen as

$$(M_t, \pi_t) = \arg \max_{(M, \pi) \in \mathcal{M}_t \times \Pi} \mathbb{E} \left[\sum_{\tau=t}^T r_{\tau}^{\pi} \right].$$

Optimistic algorithms (2)

The best consistent environment and the corresponding policy (in respect to expected rewards) for step t could then be chosen as

$$(M_t, \pi_t) = \arg \max_{(M, \pi) \in \mathcal{M}_t \times \Pi} \mathbb{E} \left[\sum_{\tau=t}^T r_{\tau}^{\pi} \right].$$

- ▶ This works for simple problems (e.g. the stochastic bandit problem) but needs to be refined for more complicated problems, like reinforcement learning.

Intuition about optimistic algorithms

- ▶ If the optimistically chosen policy receives high rewards — as assumed — then little regret is suffered (exploitation).

Intuition about optimistic algorithms

- ▶ If the optimistically chosen policy receives high rewards — as assumed — then little regret is suffered (exploitation).
- ▶ If less rewards are received, then the true environment is different from optimistically assumed environment, such that something about the environment is learned (exploration).

Intuition about optimistic algorithms

- ▶ If the optimistically chosen policy receives high rewards — as assumed — then little regret is suffered (exploitation).
- ▶ If less rewards are received, then the true environment is different from optimistically assumed environment, such that something about the environment is learned (exploration).
- ▶ Thus optimistic algorithms have the potential to implicitly trade-off exploration and exploitation.

Intuition about optimistic algorithms

- ▶ If the optimistically chosen policy receives high rewards — as assumed — then little regret is suffered (exploitation).
- ▶ If less rewards are received, then the true environment is different from optimistically assumed environment, such that something about the environment is learned (exploration).
- ▶ Thus optimistic algorithms have the potential to implicitly trade-off exploration and exploitation.
- ▶ How can this be quantified?

Regret analysis for the stochastic bandit problem (1)

- ▶ For any arm $a \in \{1, \dots, K\}$, let $N_t(a)$ be the number of times the arm has been tried, and let $R_t(a)$ be the total reward received so far for this arm.
- ▶ The consistent environments can be defined by

$$\left| \bar{r}(a) - \frac{R_t(a)}{N_t(a)} \right| \leq \sqrt{\frac{\log T}{N_t(a)}}$$

for all a , which includes M^* with probability $1 - 2A/T^2$.

Regret analysis for the stochastic bandit problem (1)

- ▶ For any arm $a \in \{1, \dots, K\}$, let $N_t(a)$ be the number of times the arm has been tried, and let $R_t(a)$ be the total reward received so far for this arm.
- ▶ The consistent environments can be defined by

$$\left| \bar{r}(a) - \frac{R_t(a)}{N_t(a)} \right| \leq \sqrt{\frac{\log T}{N_t(a)}}$$

for all a , which includes M^* with probability $1 - 2A/T^2$.

- ▶ The optimistic algorithm chooses

$$a_t = \arg \max_a \left(\frac{R_t(a)}{N_t(a)} + \sqrt{\frac{\log T}{N_t(a)}} \right).$$

Regret analysis for the stochastic bandit problem (2)

This gives the following upper bound on the regret:

$$\begin{aligned} & \mathbb{E} \left[\sum_{t=1}^T (\bar{r}(a^*) - \bar{r}(a_t)) \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \left(\bar{r}(a^*) - \left(\frac{R_t(a_t)}{N_t(a_t)} + \sqrt{\frac{\log T}{N_t(a_t)}} \right) \right) \right] \\ & \quad + \mathbb{E} \left[\sum_{t=1}^T \left(\left(\frac{R_t(a_t)}{N_t(a_t)} + \sqrt{\frac{\log T}{N_t(a_t)}} \right) - \bar{r}(a_t) \right) \right] \\ &\leq 0 + 2\mathbb{E} \left[\sum_{t=1}^T \sqrt{\frac{\log T}{N_t(a_t)}} \right] + 2A \\ &\leq 4\sqrt{TK \log T} + 2A \end{aligned}$$

Regret analysis for RL

- ▶ To keep arguments and notation simple, let the rewards be deterministic, $r(s, a) = \bar{r}(s, a)$.
- ▶ Let $S = |\mathcal{S}|$ be the number of states and $A = |\mathcal{A}|$ be the number of actions.

Regret analysis for RL

- ▶ To keep arguments and notation simple, let the rewards be deterministic, $r(s, a) = \bar{r}(s, a)$.
- ▶ Let $S = |\mathcal{S}|$ be the number of states and $A = |\mathcal{A}|$ be the number of actions.
- ▶ A central notion in our analysis is the **diameter** of an MDP:
- ▶ The diameter $D(M)$ of an MDP M is the smallest D , such that for any pair of states $s, s' \in \mathcal{S}$ there is a stationary policy $\pi_{s,s'} : \mathcal{S} \rightarrow \mathcal{A}$ which, starting at s , reaches s' in at most D steps on average.

- ▶ At time t let $N_t(s, a)$ be the number of times the state/action pair (s, a) has been tried. Let $P_t(s'|s, a)$ be the number of times this has resulted in a transition to s' .
- ▶ Then an MDP is consistent if its transition probabilities satisfy

$$\left\| p(\cdot|s, a) - \frac{P_t(\cdot|s, a)}{N_t(s, a)} \right\|_1 \leq \sqrt{\frac{S \log T}{N_t(s, a)}}.$$

- ▶ The correct M^* is among the consistent MDPs with probability $1 - SA/T^2$.

- ▶ At time t let $N_t(s, a)$ be the number of times the state/action pair (s, a) has been tried. Let $P_t(s'|s, a)$ be the number of times this has resulted in a transition to s' .
- ▶ Then an MDP is consistent if its transition probabilities satisfy

$$\left\| p(\cdot|s, a) - \frac{P_t(\cdot|s, a)}{N_t(s, a)} \right\|_1 \leq \sqrt{\frac{S \log T}{N_t(s, a)}}.$$

- ▶ The correct M^* is among the consistent MDPs with probability $1 - SA/T^2$.
- ▶ For the remaining analysis we extend the set of consistent MDPs by considering infinite action sets with all transition probabilities satisfying the above inequality. (Actually, only the finitely many vertices of the polytopes need to be considered.)

- ▶ At time t an optimistic policy for the remaining steps can be calculated by dynamic programming:

$$R_T(s) = \max_a r(s, a),$$

$$R_t(s) = \max_a \left[r(s, a) + \sum_{s'} p(s'|s, a) \cdot R_{t+1}(s') \right]$$

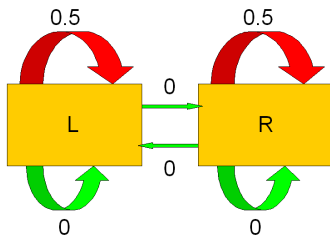
- ▶ At time t an optimistic policy for the remaining steps can be calculated by dynamic programming:

$$R_T(s) = \max_a r(s, a),$$

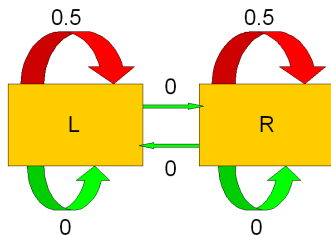
$$R_t(s) = \max_a \left[r(s, a) + \sum_{s'} p(s'|s, a) \cdot R_{t+1}(s') \right]$$

- ▶ This computation takes $O(TAS^2 \log S)$ time.

RL: The optimistic policy cannot be changed too often



RL: The optimistic policy cannot be changed too often



- ▶ If the red action in a state is tried, the optimistic estimate of the state is likely to be reduced.
- ▶ Thus the other state may appear more advantageous, but changing the state (using the green action) is costly.

RL: Splitting up into rounds with fixed policy

- ▶ Let $1 = t_1 < t_2 < \dots < t_m < t_{m+1} = T + 1$ be the (random) times when a new policy is chosen.
- ▶ We choose a new policy when the occurrences of a state/action pair has doubled:

$$t_{i+1} = \min\{t > t_i \mid \exists(s, a) : N_t(s, a) = 2N_{t_i}(s, a)\}.$$

RL: Splitting up into rounds with fixed policy

- ▶ Let $1 = t_1 < t_2 < \dots < t_m < t_{m+1} = T + 1$ be the (random) times when a new policy is chosen.
- ▶ We choose a new policy when the occurrences of a state/action pair has doubled:

$$t_{i+1} = \min\{t > t_i \mid \exists(s, a) : N_t(s, a) = 2N_{t_i}(s, a)\}.$$

- ▶ This bounds the number of rounds by

$$m \leq SA \log_2 T.$$

RL: Splitting up into rounds with fixed policy

- ▶ Let $1 = t_1 < t_2 < \dots < t_m < t_{m+1} = T + 1$ be the (random) times when a new policy is chosen.
- ▶ We choose a new policy when the occurrences of a state/action pair has doubled:

$$t_{i+1} = \min\{t > t_i \mid \exists(s, a) : N_t(s, a) = 2N_{t_i}(s, a)\}.$$

- ▶ This bounds the number of rounds by

$$m \leq SA \log_2 T.$$

- ▶ The regret is given by

$$\sum_{i=1}^m \sum_{t=t_i}^{t_{i+1}-1} (r_t^* - r_t).$$

RL: Regret per round (1)

- ▶ The optimistic policy chosen at time t_i and the corresponding MDP are denoted by π_i and M_i , respectively.
- ▶ The optimal policy and the true MDP are denoted by π^* and M^* .
- ▶ Let $R_M^\pi(t|s)$ be the total reward after t steps of a policy π on MDP M when starting in state s .
- ▶ Let s_t denote the states visited by the actual policy, and let s_t^* be the states which the optimal policy would visit.

RL: Regret per round (2)

Then

$$\begin{aligned} & \sum_{i=1}^m \sum_{t=t_i}^{t_{i+1}-1} (r_t^* - r_t) \\ &= \sum_{i=1}^m \left[R_{M^*}^{\pi^*}(t_{i+1} - t_i | s_{t_i}^*) - R_{M^*}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \end{aligned}$$

RL: Regret per round (2)

Then

$$\begin{aligned} & \sum_{i=1}^m \sum_{t=t_i}^{t_{i+1}-1} (r_t^* - r_t) \\ &= \sum_{i=1}^m \left[R_{M^*}^{\pi^*}(t_{i+1} - t_i | s_{t_i}^*) - R_{M^*}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \\ &= \sum_{i=1}^m \left[R_{M^*}^{\pi^*}(t_{i+1} - t_i | s_{t_i}^*) - R_{M_i}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \\ & \quad + \sum_{i=1}^m \left[R_{M_i}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) - R_{M^*}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \end{aligned}$$

RL: Optimality of (π_i, M_i)

We have

$$\mathbb{E} \left[R_{M^*}^{\pi^*}(t_{i+1} - t_i | s_{t_i}^*) - R_{M_i}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \leq 2D.$$

Otherwise, (π_i, M_i) would not be optimal among the consistent policies/MDPs:

- ▶ The policy which moves from s_{t_i} to $s_{t_i}^*$, then follows π^* for $t_{i+1} - t_i$ steps, then returns to $s_{t_{i+1}}$ and follows π_i , would have higher average reward.

Lemma:

$$\begin{aligned} & \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) - R_{M^*}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \\ & \leq 2D \sum_{s,a} \mathbb{E} \left[\frac{N_{t_{i+1}}(s,a) - N_{t_i}(s,a)}{\sqrt{N_{t_i}(s,a)}} \sqrt{S \log T} \right] \end{aligned}$$

RL: Overestimation of the optimistic policies (1)

Lemma:

$$\begin{aligned} & \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) - R_{M^*}^{\pi_i}(t_{i+1} - t_i | s_{t_i}) \right] \\ & \leq 2D \sum_{s,a} \mathbb{E} \left[\frac{N_{t_{i+1}}(s,a) - N_{t_i}(s,a)}{\sqrt{N_{t_i}(s,a)}} \sqrt{S \log T} \right] \end{aligned}$$

Proof: First observe that for all $t_i \leq t < t_{i+1}$ and s, s' ,

$$\left| \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t | s) \right] - \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t | s') \right] \right| \leq D$$

since otherwise (π_i, M_i) would not be optimal among the consistent policies/MDPs.

RL: Overestimation of the optimistic policies (2)

$$\begin{aligned} & \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t | s) - R_{M^*}^{\pi_i}(t_{i+1} - t | s) \right] \\ &= r_{M_i}(s, a) + \sum_{s'} p_{M_i}(s' | s, a) \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t - 1 | s') \right] \\ &\quad - r_{M^*}(s, a) - \sum_{s'} p_{M^*}(s' | s, a) \mathbb{E} \left[R_{M^*}^{\pi_i}(t_{i+1} - t - 1 | s') \right] \\ &\leq \sum_{s'} p_{M^*}(s' | s, a) \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t - 1 | s') - R_{M^*}^{\pi_i}(t_{i+1} - t - 1 | s') \right] \\ &\quad + \| p_{M_i}(\cdot | s, a) - p_{M^*}(\cdot | s, a) \|_1 \left\| \mathbb{E} \left[R_{M_i}^{\pi_i}(t_{i+1} - t | \cdot) \right] \right\|_\infty \end{aligned}$$

Summing up: Regret bound for UCRL

The (expected) regret is bounded by

$$\begin{aligned} \mathbb{E} & \left[Dm + 2D\sqrt{S \log T} \sum_{i=1}^m \sum_{s,a} \frac{N_{t_{i+1}}(s, a) - N_{t_i}(s, a)}{\sqrt{N_{t_i}(s, a)}} \right] \\ & \leq DAS \log T + 8DS\sqrt{AT \log T} \\ & = O\left(DS\sqrt{AT \log T}\right). \end{aligned}$$

This can be made to hold also with high probability for any T .

Lower bound:

$$\Omega\left(\sqrt{DSAT \log T}\right).$$

Tracking changes

- ▶ For N times we allow the MDP to change completely.
- ▶ If we restart UCRL after $T^{2/3}$ steps, we get $O\left(DS\sqrt{AT^{2/3}\log T}\right)$ regret for each episode where the MDP does not change, and $T^{2/3}$ regret for each episode where it does change.
- ▶ Thus the total regret is

$$O\left(DST^{2/3}\sqrt{A\log T} + NT^{2/3}\right).$$

Relation to other work: PAC-like bounds

- ▶ Our result gives that after

$$T \geq \tilde{\Omega}(D^2 S^2 A / \epsilon^2)$$

steps the average regret of UCRL is at most ϵ .

Relation to other work: PAC-like bounds

- ▶ Our result gives that after

$$T \geq \tilde{\Omega}(D^2 S^2 A / \epsilon^2)$$

steps the average regret of UCRL is at most ϵ .

- ▶ E^3 by Kearns and Singh (1998):
After $\text{poly}(1/\epsilon, S, A, T_{mix}^\epsilon)$ steps the per-trial regret is at most ϵ .
- ▶ Analysis of Rmax by Kakade (2003):
Bound on the number of actions which are not ϵ -optimal:

$$\#\{t : a_t \neq a_t^\epsilon\} = \tilde{O}(S^2 A (T_{mix}^\epsilon / \epsilon)^3)$$

Relation to other work: PAC-like bounds

- ▶ Our result gives that after

$$T \geq \tilde{\Omega}(D^2 S^2 A / \epsilon^2)$$

steps the average regret of UCRL is at most ϵ .

- ▶ E^3 by Kearns and Singh (1998):
After $\text{poly}(1/\epsilon, S, A, T_{mix}^\epsilon)$ steps the per-trial regret is at most ϵ .
- ▶ Analysis of Rmax by Kakade (2003):
Bound on the number of actions which are not ϵ -optimal:

$$\#\{t : a_t \neq a_t^\epsilon\} = \tilde{O}(S^2 A (T_{mix}^\epsilon / \epsilon)^3)$$

- ▶ T_{mix}^ϵ is the number of steps such that the actual per-trial reward of a stationary policy is ϵ -close to the expected per-trial reward.

Relation to other work: $\log T$ bounds

Assume that there is a **gap** g between the average per-trial reward of the optimal and the 2nd best stationary policy.

- ▶ Then the regret of UCRL is bounded by

$$O\left(\frac{D^2 S^2 A}{g} \log T\right).$$

Relation to other work: $\log T$ bounds

Assume that there is a **gap** g between the average per-trial reward of the optimal and the 2nd best stationary policy.

- ▶ Then the regret of UCRL is bounded by

$$O\left(\frac{D^2 S^2 A}{g} \log T\right).$$

- ▶ Burnetas, Katehakis, 1997, Tewari, Bartlett, 2007:

$$O\left(\frac{D_{\max}^2 SA}{g} \log T\right)$$

Relation to other work: $\log T$ bounds

Assume that there is a **gap** g between the average per-trial reward of the optimal and the 2nd best stationary policy.

- ▶ Then the regret of UCRL is bounded by

$$O\left(\frac{D^2 S^2 A}{g} \log T\right).$$

- ▶ Burnetas, Katehakis, 1997, Tewari, Bartlett, 2007:

$$O\left(\frac{D_{\max}^2 SA}{g} \log T\right)$$

- ▶ Main difference (recall $D = \max_{s_1, s_2} \min_{\pi} \mathbb{E}[T(s_2|\pi, s_1)]$):

$$D_{\max} = \max_{s_1, s_2} \max_{\pi} \mathbb{E}[T(s_2|\pi, s_1)]$$