Autonomous Exploration in Reinforcement Learning

Peter Auer[1] Shiau Hong Lim[1] Chris Watkins[2]

[1]Montanuniversität Leoben, Austria

[2]Royal Holloway University of London, UK

(CompLACS, EU-FP7)

NIPS Workshop 2011
New Frontiers in Model Order Selection

Motivation

- Learning in animals and infants:
  - explore their environments with (apparently) autonomous purpose,
  - appropriate to their current level of skills,
  - incrementally over time.

Evaluation of autonomous exploration algorithms

- Problem: No predefined task to evaluate performance
- No commonly agreed criterion

Evaluation of autonomous exploration algorithms

- Problem: No predefined task to evaluate performance
- No commonly agreed criterion
- Early work by Schmidhuber (1991, 2006)
  - information gain as intrinsic reward
  - "choose actions to increase accuracy of future predictions"

Evaluation of autonomous exploration algorithms

- Problem: No predefined task to evaluate performance
- No commonly agreed criterion
- Early work by Schmidhuber (1991, 2006)
  - information gain as intrinsic reward
  - "choose actions to increase accuracy of future predictions"
- Oudeyer, Kaplan, Hafner (2007): *Intrinsical Motivation Systems for Autonomous Mental Development*
  - behavioral complexity
    (internal complexity, information complexity,
    psychological complexity [Piaget])

Evaluation of autonomous exploration algorithms

- Problem: No predefined task to evaluate performance
- No commonly agreed criterion
- Early work by Schmidhuber (1991, 2006)
  - information gain as intrinsic reward
  - "choose actions to increase accuracy of future predictions"
- Oudeyer, Kaplan, Hafner (2007): *Intrinsical Motivation Systems for Autonomous Mental Development*
  - behavioral complexity
    (internal complexity, information complexity,
    psychological complexity [Piaget])
- Criticism: Not everything (complex) is interesting and needs to be predicted.
- Our approach: Improved utility instead of improved prediction

Learning to navigate

- MDP without rewards,
- (infinite) discrete state space,
- unknown transition probabilities P,
- start state s_0,
- RESET-action to s_0,
- known set of actions.

- Objective: explore vicinity of s_0 and find routes to all reachable states.
- Question: how many exploration steps are necessary?

Reaching all states that are reachable in L steps

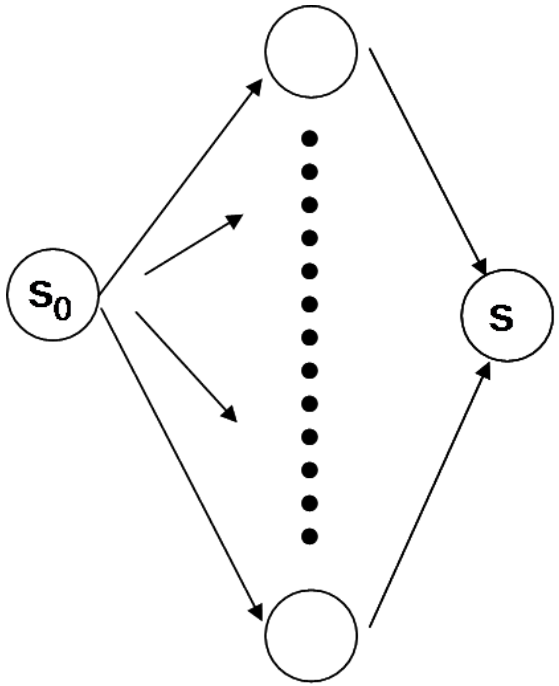
Let $\tau(s|\pi)$ be the average time for policy π to reach state s when starting in s_0. Let $\tau^*(s) = \min_\pi \tau(s|\pi)$.

- Q: How many exploration steps are sufficient, such that for all s with $\tau^*(s) \leq L$ the learner finds a policy π_s with $\tau(s|\pi_s) \leq 2L$?

Reaching all states that are reachable in L steps

Let $\tau(s|\pi)$ be the average time for policy π to reach state s when starting in s_0. Let $\tau^*(s) = \min_\pi \tau(s|\pi)$.

- Q: How many exploration steps are sufficient, such that for all s with $\tau^*(s) \leq L$ the learner finds a policy π_s with $\tau(s|\pi_s) \leq 2L$?
- A: ∞



Excluding unreachable states

Let $\pi^{(S)}$ be a policy restricted to set S, i.e. $\pi^{(S)}(s) = \mathrm{RESET}$ for $s \notin S$.

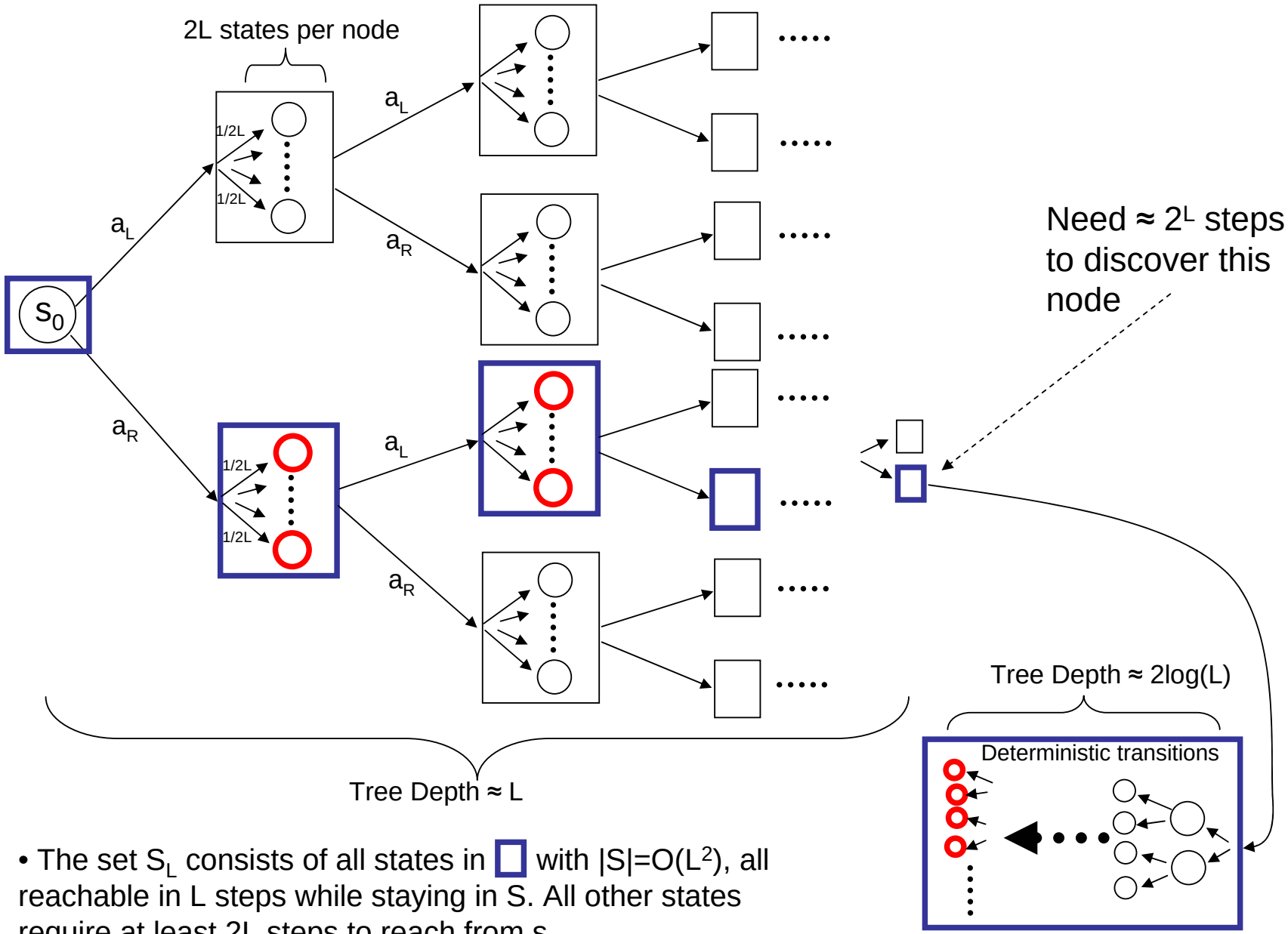
- Q: Let S be a maximal set such that for all $s \in S$, $\min_\pi \tau(s|\pi^{(S)}) \leq L$. How many exploration steps are sufficient, such that for all $s \in S$ the learner finds a policy π_s with $\tau(s|\pi_s) \leq 2L$ (S is unknown)?

Excluding unreachable states

Let $\pi^{(S)}$ be a policy restricted to set S, i.e. $\pi^{(S)}(s) = \mathrm{RESET}$ for $s \notin S$.

- Q: Let S be a maximal set such that for all $s \in S$, $\min_\pi \tau(s|\pi^{(S)}) \leq L$. How many exploration steps are sufficient, such that for all $s \in S$ the learner finds a policy π_s with $\tau(s|\pi_s) \leq 2L$ (S is unknown)?
- A: exponentially many in $\sqrt{L}$.

Excluding unreachable states (counterexample)



• The set S_L consists of all states in ▢ with $|S|=O(L^2)$, all reachable in L steps while staying in S. All other states require at least 2L steps to reach from s_0.

Excluding intermediate states

- Let $\prec$ be an (unknown) partial order on the state space.
- Let $S_{\prec s} := \{s' \in S : s' \prec s\}$.
- Let S be an (unknown) maximal set such that for all $s \in S$, $\min_\pi \tau(s|\pi^{(S_{\prec s})}) \leq L$.

Excluding intermediate states

- Let $\prec$ be an (unknown) partial order on the state space.
- Let $S_{\prec s} := \{s' \in S : s' \prec s\}$.
- Let S be an (unknown) maximal set such that for all $s \in S$, $\min_\pi \tau(s|\pi^{(S_{\prec s})}) \leq L$.

- Then policies π_s can be learned incrementally for all states $s \in S' \supseteq S$ such that $\tau(s|\pi_s^{(S'_{\prec s})}) \leq 2L$, using an optimistic algorithm and $\tilde{O}\left(L^3|S'|A\right)$ exploration steps.

Reinforcement learning

Markov decision process (MDP):

- In each step $t = 0, 1, \ldots$
  - the agent observes state s_t,
  - chooses an action a_t,
  - receives reward $r_t = r(s_t, a_t) \in [0, 1]$,
  - and the state changes according to transition probability $p(s_{t+1}|s_t, a_t)$.
- Actions are chosen according to some policy π.

Reinforcement learning

Markov decision process (MDP):

- In each step $t = 0, 1, \ldots$
  - the agent observes state s_t,
  - chooses an action a_t,
  - receives reward $r_t = r(s_t, a_t) \in [0, 1]$,
  - and the state changes according to transition probability $p(s_{t+1}|s_t, a_t)$.
- Actions are chosen according to some policy π.

(The reward function $r(\cdot, \cdot) \in [0, 1]$ is known, the transition probabilities are unknown.)

Discounted and undiscounted rewards

An optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$ maximizes

either $V_\gamma(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$ for some discount factor $\gamma \in [0, 1)$,

or $\rho := \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t\right]$.

Discounted and undiscounted rewards

An optimal policy $\pi^* : \mathcal{S} \to \mathcal{A}$ maximizes

either $V_\gamma(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$ for some discount factor $\gamma \in [0, 1)$,

or $\rho := \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t\right]$.

Undiscounted rewards allow for an immediate notion of regret:

$$\Delta(T) := \mathbb{E}\left[\sum_{t=0}^{T-1} [r(s_t^*, a_t^*) - r(s_t, a_t)]\right] .$$

PAC-MDP bounds for discounted rewards

- #steps t where $V_\gamma(s_t) < V^*_\gamma(s_t) - \epsilon$,
- or #steps t where $Q^*_\gamma(s_t, a_t) < V^*_\gamma(s_t) - \epsilon$.

Q-function of the optimal policy:

$Q^*_\gamma(s,a) := r(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V^*_\gamma(s')$.

PAC-MDP bounds for discounted rewards

- #steps t where $V_\gamma(s_t) < V^*_\gamma(s_t) - \epsilon$,
- or #steps t where $Q^*_\gamma(s_t, a_t) < V^*_\gamma(s_t) - \epsilon$.

Q-function of the optimal policy:
$Q^*_\gamma(s,a) := r(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V^*_\gamma(s')$.

Corresponding notions of regret:

$$\Delta^V_\gamma(T) := \mathbb{E}\left[\sum_{t=0}^{T-1} \left[V^*_\gamma(s_t) - V_\gamma(s_t)\right]\right]$$

$$\Delta^Q_\gamma(T) := \mathbb{E}\left[\sum_{t=0}^{T-1} \left[V^*_\gamma(s_t) - Q^*_\gamma(s_t, a_t)\right]\right]$$

Regret bounds

UCRL2 (Jaksch at al, 2010)

$$\Delta(T) = \tilde{O}\left(DS\sqrt{AT}\right)$$

The diameter D

For any pair of states s, s' there is a policy $\pi_{s,s'} : \mathcal{S} \to \mathcal{A}$ which, starting at s, reaches s' in at most D steps on average.

Discounted UCRL

$$\Delta_\gamma^Q(T) = \tilde{O}\left(\frac{V_{\max}\sqrt{SAT}}{1-\gamma}\right)$$

Corollary

$$\Delta_\gamma^V(T) = \tilde{O}\left(\frac{V_{\max}\sqrt{SAT}}{(1-\gamma)^2}\right)$$

PAC-MDP bounds from regret bounds

Corollary

$$\Delta_\gamma^V(T) = \tilde{O}\left(\frac{V_{\max}\sqrt{SAT}}{(1-\gamma)^2}\right)$$

Theorem (Szita & Szepesvari, 2010)

For any MDP, $V_\gamma(s_t) < V_\gamma^*(s_t) - \epsilon$ at most

$$\tilde{O}\left(\frac{SAV_{\max}^2}{\epsilon^2(1-\gamma)^4}\right)$$

times.

(By $\Delta(T) \approx \epsilon T$.)

Optimistic policies for regret bounds in RL

An optimistic policy

- assumes the best possible *consistent MDP* (consistent with past observations),
- and chooses an optimal policy in respect to this best possible MDP.
- Since — in contrast to bandit problems — rewards may be delayed in RL, the optimistic policy needs to be followed for some time.

Optimistic policies for regret bounds in RL

An optimistic policy

- assumes the best possible *consistent MDP* (consistent with past observations),
- and chooses an optimal policy in respect to this best possible MDP.
- Since — in contrast to bandit problems — rewards may be delayed in RL, the optimistic policy needs to be followed for some time.

Set of consistent MDPs $\mathcal{M}_t$:

- With high probability, $\mathcal{M}_t$ contains the true MDP M^*.
- For good regret bounds, $\mathcal{M}_t$ should be small.

Intuition about optimistic policies

- If the optimistically chosen policy receives high rewards — as assumed — then little regret is suffered (exploitation).
- If less reward is received, then the true environment is different from optimistically assumed environment, such that something about the environment is learned (exploration).
- Thus optimistic policies potentially trade-off exploration and exploitation.

Consistent MDPs

- Let $N(s,a)$ be the plays of state/action pair (s,a) so far. An MDP $\tilde{M}$ is consistent if for all (s,a) the following holds.

- UCRL2:

$$||\tilde{\mathbf{p}}(\cdot|s,a) - \hat{\mathbf{p}}(\cdot|s,a)||_1 \leq \sqrt{\frac{S\log(1/\delta)}{N(s,a)}}.$$

Consistent MDPs

- Let $N(s,a)$ be the plays of state/action pair (s,a) so far. An MDP $\tilde{M}$ is consistent if for all (s,a) the following holds.

- UCRL2:

$$||\tilde{\mathbf{p}}(\cdot|s,a) - \hat{\mathbf{p}}(\cdot|s,a)||_1 \leq \sqrt{\frac{S\log(1/\delta)}{N(s,a)}}.$$

- Discounted UCRL:
  - $||\tilde{\mathbf{p}}(\cdot|s,a) - \hat{\mathbf{p}}(\cdot|s,a)||_\infty \leq \sqrt{\frac{\log(1/\delta)}{N_i(s,a)}}$
  - $\left|[\tilde{\mathbf{p}}(\cdot|s,a) - \hat{\mathbf{p}}(\cdot|s,a)]\tilde{\mathbf{V}}^*_\gamma(\cdot)\right| \leq V_{\max}\sqrt{\frac{\log(1/\delta)}{N_i(s,a)}}$

Main quantities in the proof

- UCRL2:

$$
\begin{aligned}
& \mathbf{n}(\circ)\,[\tilde{\rho} - \mathbf{r}(\circ)] \\
&= \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{I}]\,\tilde{\mathbf{V}}_0(\cdot) \\
&= \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) + \mathbf{n}(\circ)\,[\mathbf{p}(\cdot|\circ) - \mathbf{I}]\,\tilde{\mathbf{V}}_0(\cdot) \\
&\approx \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) \\
&= \mathbf{n}\,[\tilde{\mathbf{p}}(\cdot|\circ) - \hat{\mathbf{p}}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) + \mathbf{n}\,[\hat{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot)
\end{aligned}
$$

Main quantities in the proof

- UCRL2:

$$
\begin{aligned}
& \mathbf{n}(\circ)\,[\tilde{\rho} - \mathbf{r}(\circ)] \\
&= \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{I}]\,\tilde{\mathbf{V}}_0(\cdot) \\
&= \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) + \mathbf{n}(\circ)\,[\mathbf{p}(\cdot|\circ) - \mathbf{I}]\,\tilde{\mathbf{V}}_0(\cdot) \\
&\approx \mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) \\
&= \mathbf{n}\,[\tilde{\mathbf{p}}(\cdot|\circ) - \hat{\mathbf{p}}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot) + \mathbf{n}\,[\hat{\mathbf{p}}(\cdot|\circ) - \mathbf{p}(\cdot|\circ)]\,\tilde{\mathbf{V}}_0(\cdot)
\end{aligned}
$$

- Discounted UCRL:

$$
\begin{aligned}
& \mathbf{n}(\circ)\,[\tilde{\mathbf{V}}(\circ) - \mathbf{Q}^*(\circ)] \\
&= \gamma\mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ)\tilde{\mathbf{V}}(\cdot) - \mathbf{p}(\cdot|\circ)\mathbf{V}^*(\cdot)] \\
&= \gamma\mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \hat{\mathbf{p}}(\cdot|\circ)]\,\tilde{\mathbf{V}}(\cdot) \\
&\quad + \gamma\mathbf{n}(\circ)\,[\hat{\mathbf{p}}(\cdot|\circ)\tilde{\mathbf{V}}(\cdot) - \mathbf{p}(\cdot|\circ)\mathbf{V}^*(\cdot)] \\
&\approx \gamma\mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \hat{\mathbf{p}}(\cdot|\circ)]\,\tilde{\mathbf{V}}(\cdot) + \gamma\mathbf{n}(\circ)\,\mathbf{p}(\cdot|\circ)\,[\tilde{\mathbf{V}}(\cdot) - \mathbf{V}^*(\cdot)] \\
&\lesssim \gamma\mathbf{n}(\circ)\,[\tilde{\mathbf{p}}(\cdot|\circ) - \hat{\mathbf{p}}(\cdot|\circ)]\,\tilde{\mathbf{V}}(\cdot) + \gamma\mathbf{n}(\circ)\,[\tilde{\mathbf{V}}(\circ) - \mathbf{Q}^*(\circ)]
\end{aligned}
$$

Summing over episodes (discounted UCRL)

$$
\begin{aligned}
\text{Regret} &\leq \sum_i \mathbf{n}_i(\circ)\,[\tilde{\mathbf{V}}_i(\circ) - \mathbf{Q}_i^*(\circ)] \\
&\lesssim \frac{V_{\max}}{1-\gamma} \sum_{i,s,a} \frac{\mathbf{n}_i(s,a)}{\sqrt{N_i(s,a)}} \\
&\lesssim \frac{V_{\max}\sqrt{SAT}}{1-\gamma}
\end{aligned}
$$

Optimistic algorithm for autonomous exploration

- Maintains a set of known states S and a set of unknown states U.
- Initially $S = \{s_0\}$ and $U = \{s_I\}$ where s_I is an imaginary state.

- Iterate:
  - Optimistically choose $s \in U$ and π_s (defined only on S) with minimal $\tilde{\tau}(s|\pi_s)$.
  - If $\tilde{\tau}(s|\pi_s) > L$ then stop.
  - Repeat $K = C \log(|S|/\delta)$ times:
    - Run π_s for $2L$ steps, then RESET.
  - If some state $s' \notin S$ is reached in at least $K/3$ of the repetitions, then $S \leftarrow S \cup \{s'\}$, $U \leftarrow U \setminus \{s'\}$, and let π_s be the corresponding policy.

Analysis (1)

Lemma

If $s \in S$, then s is reached by the corresponding policy in $8L$ steps.

Let $s' \notin S$ be such that s' can be reached in L steps by a policy on S, and let $M_{s'}$ be a modified MDP where s' is absorbing and gives reward 1.

Lemma

There is a policy for $M_{s'}$ that gives total expected reward $\geq L$ after $2L$ steps.

Lemma

If no state is added to S in an iteration of the optimistic algorithm, then π_s suffers total regret $\geq LK/3$ after K repetitions of length $2L$ on any $M_{s'}$.

Analysis (2)

Let $\mathbf{n}(s,a)$ be the plays of state/action pair (s,a) in the current iteration.

Lemma

The total regret of π_s after K repetitions of length $2L$ on $M_{s'}$, is bounded by

$$\tilde{O}\left(\sum_{s,a} L \frac{\mathbf{n}(s,a)}{\sqrt{\mathbf{N}(s,a)/L}}\right) .$$

Lemma

The total regret in unsuccessful iterations is $\tilde{O}\left(\sqrt{L^3|S'|AT}\right)$ where T is the number of steps in these iterations.

Theorem

$T = \tilde{O}\left(L^3|S'|A\right)$.

Analysis (3): Consistent MDPs

We are calculating an optimistic $2L$-step policy π_s for $s \in U$.

For all (s,a), divide the $N(s,a)$ observations into $2L$ chunks of equal size, and let $\hat{\mathbf{p}}_i(\cdot|s,a)$, $t = 1, \ldots, 2L$, be the corresponding estimates.

- $R_0(s') = 0$.
- $R_{t+1}(s) = t + 1$.
- $R_{t+1}(s') = \max_a \left[\hat{\mathbf{p}}_t(\cdot|s',a) \cdot R_t(\cdot) + \tilde{O}\left(\frac{L}{\sqrt{N(s',a)/L}} \right) \right]$.

Summary

- Autonomous exploration is interesting
- Evaluation of autonomous exploration is difficult
- Analysis of a concrete model of autonomous(?) exploration
  - Learning to navigate: reasonably general setting
  - Difficult cases cannot be learned
  - Our algorithm: incremental optimistic exploration
  - Performance analysis through regret analysis
- Imitation is also an interesting learning mode.

(Why) is autonomous exploration useful?