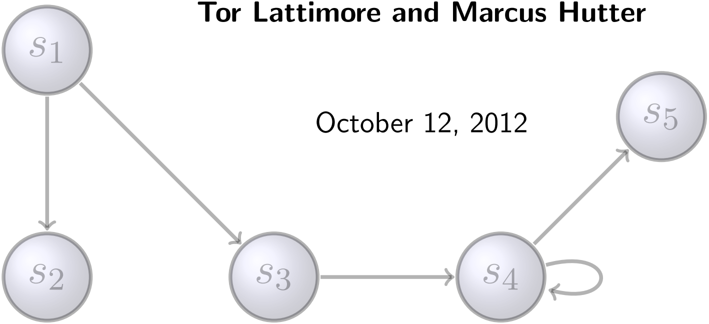
Sample-Complexity of Learning MDPs

Tor Lattimore and Marcus Hutter

October 12, 2012



Reinforcement Learning

- Maximise long-term discounted reward
- Hard because environment is unknown
- We model the environment using finite state Markov Decision Processes with unknown transitions

Markov Decision Processes

Goal: Construct $\mathcal{A}$ with $V_M^{\mathcal{A}}(s_{1:t}) = V_M^*(s_t)$.

Problem 1: M is unknown. $\mathcal{A}$ has to spend some time exploring

Problem 2: The environment is stochastic. $\mathcal{A}$ can be "unlucky"

Notation	
M	(S, A, p, r, γ)
$V_M^*(s_t)$	value of optimal policy
$V_M^{\mathcal{A}}(s_{1:t})$	value of $\mathcal{A}$ in M

Sample Complexity

An algorithm $\mathcal{A}$ is (ϵ, δ)-*correct* with *sample complexity* N if for all $M \in \mathcal{M} := \{(S, A, p, r, \gamma) : p$ transition probabilities$\}$,

$$P\left\{ \sum_{t=1}^{\infty} [\![V_M^*(s_t) - V_M^{\mathcal{A}}(s_{1:t}) > \epsilon]\!] > N \right\} < \delta$$

$\#$ time-steps where $\mathcal{A}$ is not ϵ-optimal

"The probability that I am 'badly' suboptimal for more than N time-steps is at most δ!"

Theorems

UCRLγ is a combination/modification of UCRL2 (Ortner & Auer 2010) and MBIE (Littman et al., 2008)

Theorem (Upper Bound, L & Hutter 2012)

For $0 < \epsilon \leq 1$, UCRLγ is (ϵ, δ)-correct with sample complexity

$$\tilde{O}\left(\frac{T}{\epsilon^2(1-\gamma)^3}\log\frac{1}{\delta}\right),$$

where $T \leq |S|^2|A|$ is the number of non-zero transition probabilities.

Theorem (Lower Bound, L & Hutter 2012)

Every (ϵ, δ)-correct policy has sample complexity at least

$$\tilde{\Omega}\left(\frac{|S||A|}{\epsilon^2(1-\gamma)^3}\log\frac{1}{\delta}\right).$$

History of the Upper Bound

$$\underbrace{\frac{|S|^2|A|}{\epsilon^3(1-\gamma)^6}\log\frac{1}{\delta}}_{\textsc{r-max}\ (2002)}$$

$$\underbrace{\frac{|S||A|}{\epsilon^4(1-\gamma)^7}\log\frac{1}{\delta}}_{\textsc{Delayed Q-Learning}\ (2006)}$$

$$\underbrace{\frac{|S|^2|A|}{\epsilon^3(1-\gamma)^6}\log\frac{1}{\delta}}_{\textsc{mbie}\ (2008)}$$

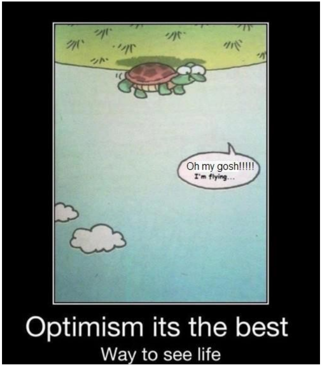
$$\underbrace{\frac{|S||A|}{\epsilon^2(1-\gamma)^6}\log\frac{1}{\delta}}_{\textsc{mormax}\ (2010)}$$

$$\underbrace{\frac{|S||A|}{\epsilon^2(1-\gamma)^4}\log\frac{1}{\delta}}_{\textsc{ucrl}\ (2011)^*}$$

$$\underbrace{\frac{|S|^2|A|}{\epsilon^2(1-\gamma)^3}\log\frac{1}{\delta}}_{\textsc{ucrl}\gamma\ (2012)}$$

*unpublished

Algorithm Sketch



loop
    Compute empiric estimate, $\hat{p}$, of transition matrix p
    Compute confidence interval about $\hat{p}$
    Act according to the most optimistic plausible MDP
end loop

Analysis

- Key component is bounding $|V^{\mathcal{A}}_{\hat{M}}(s_{1:t}) - V^{\mathcal{A}}_{M}(s_{1:t})| < \epsilon$
- Require each state to be visited sufficiently often
- States that are expected to be visited often needed better estimates

Bernstein's, $\sigma^2(s) := \operatorname{Var} V(s'|s)$ and $L := \log 1/\delta$

Discounted future state distribution

Error

$$V^{\mathcal{A}}_{\hat{M}}(s_t) - V^{\mathcal{A}}_{M}(s_t) = \gamma \sum_s \underbrace{w(s)}_{\text{Discounted future state distribution}} \underbrace{(p_s - \hat{p}_s) \cdot \widehat{V}}_{\text{Error}} \overset{\approx}{\lesssim} \sum_s w(s) \sqrt{\frac{|S|\sigma^2(s)L}{n(s)}}$$

$$= \sum_s \sqrt{\frac{L|S|w(s)\sigma^2(s)}{m := n(s)/w(s)}} \overset{\approx}{\lesssim} \sqrt{\frac{L|S|^2}{m} \sum_s w(s)\sigma^2(s)}$$

$\operatorname{Var} \sum_{k=t}^{\infty} \gamma^{k-t} r_k \le \frac{1}{(1-\gamma)^2}$

$$\le \sqrt{\frac{L|S|^2}{m(1-\gamma)^2}}$$

Therefore $n(s) \approx \dfrac{w(s)L|S|^2}{\epsilon^2(1-\gamma)^2}$ visits to state s needed

Analysis

$$n \approx \frac{w(s)L|S|^2}{\epsilon^2(1-\gamma)^2} \qquad\qquad H := \frac{1}{1-\gamma}\log\frac{1}{\epsilon(1-\gamma)}$$

- $w(s)$ is the discounted future state distribution
- Expect to visit s at least $w(s)$ times within H time-steps
- Expect to "know" a state after $\dfrac{L|S|^2 H}{\epsilon^2(1-\gamma)^2}$ time-steps
- Once we "know" all states we are optimal
- Expect sample complexity bounded by $\dfrac{|S|^2|A|H}{\epsilon^2(1-\gamma)^2}\log\dfrac{1}{\delta}$
- Analysis harder since $w(s)$ changes over time and we want results with high probability

Summary

- Upper and lower bounds with (unimprovable) cubic dependence on horizon
- Unfortunately our analysis led to an extra dependence on $|S|$ for dense transition probabilities
- See paper for algorithm and (very) messy details

Questions