Optimistic Agents are Asymptotically Optimal

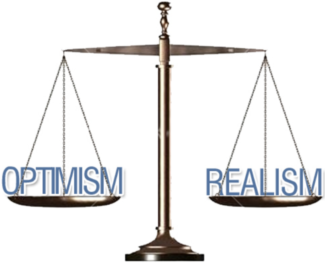
Peter Sunehag and Marcus Hutter

2012

Optimism in Reinforcement Learning (MDPs)

Optimism has been extensively used in many different ways as an exploration technique for Markov Decision Processes

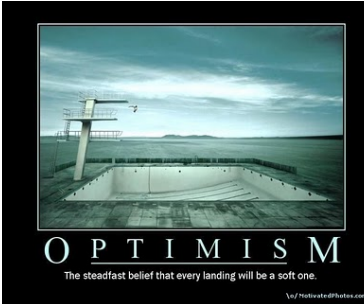
Choose the plausible model in which one can achieve the highest expected return (used by Strehl and Littman in MBIE (discounted) and by Auer and Ortner in UCRL (undiscounted))



Disasters

To choose the most optimistic plausible
model is wildly optimistic to begin
with when everything is equally plausible

Our results do not rely on any restrictions
(Markov, Ergodicity,...) on our environments
except on the size of the class considered.
The results are, however, primarily
interesting if the agent cannot destroy itself.



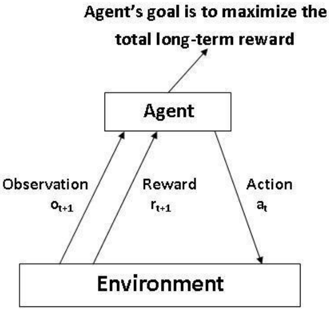
General Reinforcement Learning

An environment $\nu(h_t, a_t) = (o_t, r_t)$
where $h_t = a_1 o_1 r_1, ..., a_t o_t r_t$.

Maximize the discounted reward sum (return) $\sum_{i=t}^{\infty} r_i \gamma^{t-i}$ where $\gamma \in (0, 1)$

A policy is a function $\pi(h_t) = a_t$

$V^{\pi}_{\nu}(h_t) =$ expected return (in ν)
achieved by following policy π after h_t



π is asymptotically optimal if for the true environment μ
$\lim_{t\to\infty}(\max_{\tilde{\pi}} V^{\tilde{\pi}}_{\mu}(h_t) - V^{\pi}_{\mu}(h_t)) = 0$

Optimistic Agent for Deterministic Environments

- Consider a finite class $\mathcal{M} = \{\nu_1, ..., \nu_m\}$
- We will define a policy π^0 such that there is a T such that $V^\pi_\nu(h_t)$ is maximal (given the past) when $t \geq T$ as long as $\nu \in \mathcal{M}$
- Let $\mathcal{M}_t$ be the set of environments that remains consistent at time t
- We will choose the most optimistic hypothesis and policy and act according to it until it is contradicted. As long as the outcome is consistent with the optimistic prediction the return is optimal, even if the environment is wrong.

Algorithm 1 (π°):

- 1. Choose $(\pi^*, \nu^*) \in \arg\max_{\pi \in \Pi, \nu \in \mathcal{M}_{t-1}} V^\pi_\nu(h_{t-1})$,
- 2. Act according to π^* until contradicted and then go back to 1.

Asymptotic Optimality

Theorem (Optimality, Finite Deterministic Class)

If we use Algorithm 1 (π°) in an environment $\mu \in \mathcal{M}$, then there is $T < \infty$ such that

$$V_\mu^{\pi^\circ}(h_t) \; = \; \max_\pi V_\mu^\pi(h_t) \; \forall t \geq T.$$

Proof.

(sketch) After a finite amount of time all environments that will be excluded have been excluded. Due to time consistency of geometric discounting, the optimistic policy remains optimistic and since the optimistic environment remains consistent under this policy there is no better policy, even if the environment is wrong.

■

Finite Error Bound

Theorem (Finite error bound)

Following π° (Algorithm 1),

$$V_\mu^{\pi^\circ}(h_t) \geq \max_{\pi\in\Pi} V_\mu^{\pi}(h_t) - \epsilon,\ 0 < \epsilon < 1/(1-\gamma)$$

for all but at most $|\mathcal{M}|\frac{\log\epsilon(1-\gamma)}{\gamma-1}$ time steps t.

Proof.

Each time an environment is contradicted π° could have been ϵ-suboptimal for at most $\frac{\log\epsilon(1-\gamma)}{\log\gamma}$ time steps. ∎

- We can prove asymptotic optinality also for agents that reevaluate its choice of optimistic hypothesis at every time step. We call these agents liberal and the agent that keep their hypothesis until contradiction conservative.

Stochastic Environments

Class of stochastic environment $\mathcal{M}$

- We need a new exclusion criteria: Use threshold on likelihood ratio
- Exclude ν if $\frac{\nu(h_t|a_{1:t})}{\max_{\tilde{\nu}\in\mathcal{M}} \tilde{\nu}(h_t|a_{1:t})} < z$
- Reevaluate choice of optimistic environment at each time step, because optimism can end without exclusion

Optimality, Stochastic Case

Theorem (Optimality, Finite Stochastic Class)

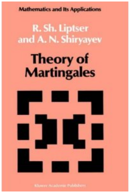
Define π° *by using Algorithm 2 with any threshold* $z \in (0, 1)$ *and a finite class* $\mathcal{M}$ *of stochastic environments containing the true environment* μ*, then with probability* $1 - z|\mathcal{M} - 1|$ *there exists, for every* $\epsilon > 0$*, a number* $T < \infty$ *such that*

$$V_\mu^{\pi^\circ}(h_t) \ > \ \max_\pi V_\mu^\pi(h_t) - \epsilon \, \forall t \geq T.$$

Remark

For a different exclusion criteria Lattimore, Hutter and Sunehag (submitted) prove sample complexity bounds.

Extending the proofs



Mathematics and Its Applications

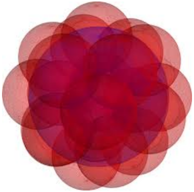
R. Sh. Liptser
and A. N. Shiryayev

Theory of
Martingales

Kluwer Academic Publishers

- The martingale convergence theorem tells us that the likelihood ratio's converge
- A slightly extended Blackwell-Dubins Theorem tells us that the limit is strictly larger than 0 if and only if the environments merge (in total variation) under the policy followed
- Hence the environments that do not merge will be excluded and there is a finite amount of time after which every environment that will be excluded have been excluded
- For every $\epsilon_1 > 0$ there is a time after which all the environments are within a sufficiently small total variation ball to make the value functions differ by less than ϵ_1
- From optimality we conclude near optimality

Infinite Compact Classes



- The simplest way to extend to the compact class is to choose an accuracy $\epsilon > 0$ in advance and then we use the centers of finitely many sufficiently small balls that cover the space of environments
- If we want the same asymptotic optimality as in the finite case, we need to have balls that decrease at the right speed and we let $\mathcal{M}_t$ consists of such (confidence) balls around every non-excluded environment
- There are non-compact countable classes for which asymptotic optimality is impossible to achieve, but we can achieve weak asymptotic optimality by introducing environments slowly into the class our optimistic agent works with.

Conclusions

- Sunehag and Hutter (AGI'2012) provides an axiomatic system for optimistic agents
- Weakens one of the rationality axioms in a way that break symmetry. We refer to the complying agents as optimistically rational
- In comparison to the rational/bayesian agents we have asymptotic optimality guarantees
- Bayesian agents are, however, best on average with respect to the prior (which can have a big impact, even for relatively uninformative priors)
- Different sense of optimality leads to different agents. One is not optimal in every sense