Optimistic Reinforcement Learning

(with lots of equations)

Tor Lattimore, Marcus Hutter and Peter Sunehag

October 3, 2012

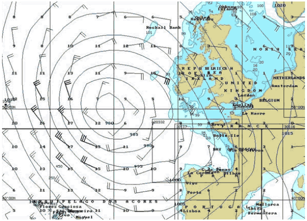
Planning

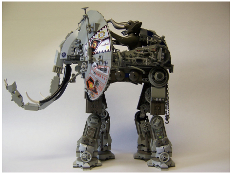
Finding the best policy in a *known* world.

Reinforcement Learning

Finding the best policy in an *unknown* world.

Reinforcement Learning





How Good is My Algorithm?

- Model class, $\mathcal{M}$
- Calculate the expected number of mistakes you make in each possible model[1]
- Worst-case result is a measure of ability

[1] Sometimes expectation is replaced with "with high probability"

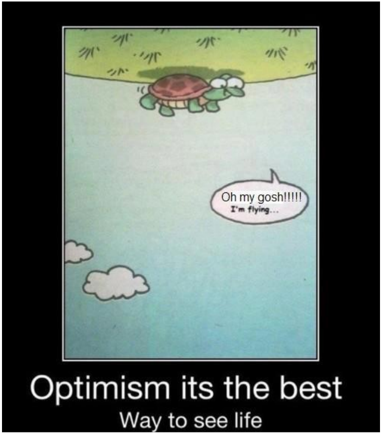
How to Make a Good Algorithm?

The optimism principle

- Think of the **plausable** world you'd most like to be in.
- Act as if you're in that world.

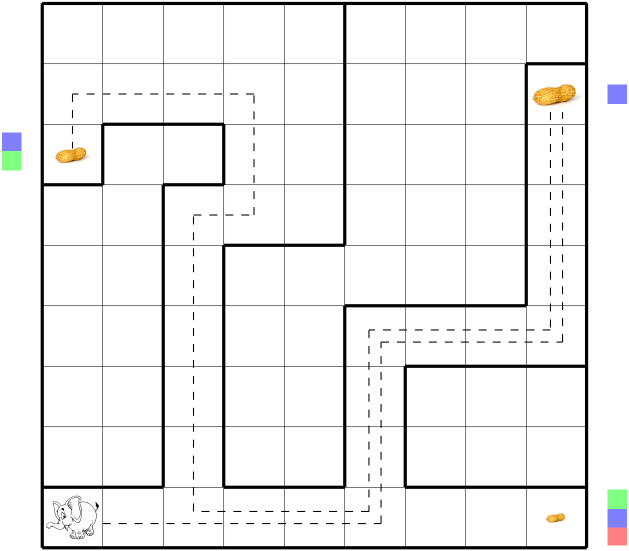
Why it works

- If you're right then your actions are **optimal**.
- If you're wrong then you can **discard** that world.



Example - Grid World

Three possible worlds

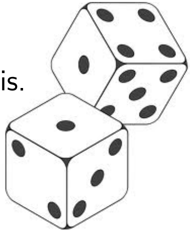


Stochastic Case

- Never know anything for sure. Seems hard.
- Eliminate environments when they become very unlikely (implausable).
- Take the bound you proved in deterministic case and multiply it by

$$\frac{1}{\epsilon^2} \log \frac{1}{\delta}$$

Claim proof is too long for the paper. It probably is.



Theory

Theorem

If $\mathcal{M}$ is a class of N arbitrary environments where values are discounted geometrically. Then with probability at least $1-\delta$ an algorithm (loosely) based on the optimism principle makes at most

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

ϵ-errors.

- Matching lower bound
- Compact classes
- Counter-example in non-compact case

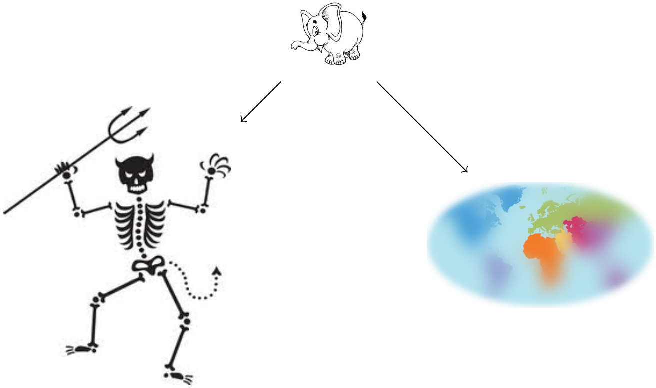
$\Big($ $\frac{1}{1-\gamma}$ is essentially the diameter, so the heuristic on the previous slide works $\Big)$

The Best Algorithm (?)

- We know model class, $\mathcal{M}$
- Want to minimise the maximum number of errors
- Search through all algorithms and choose the best one![2]
- This is a horrible idea

[2] ~~Totally incomputable~~ Analysis of computation complexity an interesting direction for future research

Hell is Bad

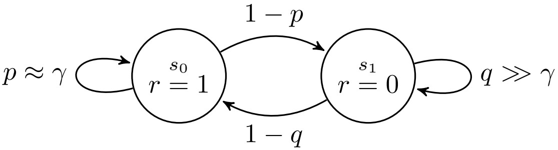


- Best to go directly to hell
- Therefore **don't** optimise for sample-complexity bounds only

Summary and Questions?

1. Optimism is a good principle for reinforcement learning if *uniform* optimality properties are desired
2. We proved sample-complexity bounds for very general environment classes (see paper)
3. Blindly optimising for sample-complexity is not smart

Example



Let $\hat{p}$ be the empiric estimate of p from n samples.

$$|V(s_0) - \widehat{V}(s_0)| \approx \frac{|\hat{p} - p|}{(1-\gamma)^2} \overset{?}{<} \epsilon$$

Bound	**Estimate**	**n**
Hoeffding	$|\hat{p} - p| \lesssim \sqrt{L/n}$	$n > \frac{L}{\epsilon^2(1-\gamma)^4}$
Bernstein	$|\hat{p} - p| \lesssim \sqrt{p(1-p)L/n}$	$n > \frac{Lp(1-p)}{\epsilon^2(1-\gamma)^4} \approx \frac{L}{\epsilon^2(1-\gamma)^3}$

$$L = \log \frac{1}{\delta}$$

Concentration Inequalities

Theorem (Markov's inequality)

Let X be an arbitrary random variable and $\epsilon > 0$ then

$$P\left\{|X| \geq \epsilon\right\} \leq \frac{\mathbf{E}|X|}{\epsilon}$$

Theorem (Chebyshev's inequality)

Let X be an arbitrary random variable and $\epsilon > 0$ then

$$P\left\{|X - \mathbf{E}X| \geq \epsilon\right\} \leq \frac{\mathrm{Var}\, X}{\epsilon^2}$$

Corollary

Let $X_1 \cdots X_n$ be i.i.d with $|X_i| < c$ and mean μ then

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right\} \leq \frac{c^2}{n\epsilon^2}$$

Concentration Inequalities

Theorem (Hoeffding-Azuma Inequality)

Let $X_1 \cdots X_n$ be independent r.v's with $X_i \in [a_i, b_i]$ with probability 1. If $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ then

$$P\left\{\left|\bar{X} - \mathbf{E}[\bar{X}]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Corollary

If $X_1 \cdots X_n$ are Bernoulli with parameter p then

$$P\left\{|p - \hat{p}| \geq \epsilon\right\} \leq 2\exp\left(-2\epsilon^2 n\right)$$

Concentration Inequalities

Theorem (Bernstein's Inequality)

Let $X_1 \cdots X_n$ be independent with means μ_i and variances σ_i^2. If $|X_i| \leq c$ w.p.1 then

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2 n}{2\sigma^2 + 2c\epsilon/3}\right)$$

where $\mu := \frac{1}{n}\sum_{i=1}^{n}\mu_i$ *and* $\sigma^2 := \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$.

Corollary

If $X_1 \cdots X_n$ are i.i.d Bernoulli with parameter p then

$$P\left\{|p - \hat{p}| \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2 n}{2p(1-p) + 2\epsilon/3}\right)$$

Confidence Intervals

We say CI is a *confidence interval* at level $1-\delta$ if

$$P\{|p-\hat{p}| \geq CI\} \leq \delta$$

Different bounds lead to different confidence intervals.

Name	**Probability Bound**	**Confidence Interval**
Chebyshev's	$\frac{1}{n\epsilon^2}$	$\sqrt{\frac{1}{n\delta}}$
Hoeffding's	$\exp(-2\epsilon^2 n)$	$\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$
Bernstein's	$\exp(-\frac{\epsilon^2 n}{2p(1-p)+2\epsilon/3})$	$\frac{2}{3n}\log\frac{2}{\delta} + \sqrt{\frac{2p(1-p)}{n}\log\frac{2}{\delta}}$

FAQ

1. What if your data isn't independent? Most results can be extended to Martingales. Beautiful paper by McDiarmid and more recently by Seldin et al (2011).
2. Are they tight? They can be.
3. How do I prove these bounds? A variety of methods. Often a Markov inequality on a cleverly chosen r.v is enough.
4. Can I eliminate the dependence on c? Yes, amazingly, but not with an unbiased estimator. See Catoni (2009).

Chernoff Bound

Theorem (Chernoff)

Let $X_1 \cdots X_n$ be Bernoulli r.v's with parameter p then

$$P\{\hat{p} \geq q\} \leq \exp(-nD(q,p))$$

Proof.

Let $x \in \mathcal{B}^n$ be a sequence where the number of successes, k satisfies $k \geq nq$.

$$\frac{P_q(x)}{P_p(x)} = \frac{q^k(1-q)^{n-k}}{p^k(1-p)^k} \geq \frac{q^{nq}(1-q)^{n-nq}}{p^{nq}(1-p)^{n-nq}} = \exp(nD(q,p))$$

Let S be the set of all such x then

$$P_p(S) \leq P_q(S)\exp(-nD(q,p)) \leq \exp(-nD(q,p))$$

as required. □

Bandits

Definition (Bandit)

Let A be a set of actions then a bandit is a vector $p \in [0,1]^{|A|}$.

At each time-step an agent chooses an action a and receives reward 1 with probability $p(a)$ and reward 0 otherwise.

Definition

The best arm is $a^* := \arg\max_a p(a)$.

Definition (Policy)

A policy is a function $\pi : \{0,1\}^* \to A$

Bandit Sample Complexity

Question. Can we construct an algorithm where the number of mistakes is bounded high probability?

Definition (Bandit Sample Complexity)

A policy π has sample complexity N if

$$P\left\{\sum_{t=1}^{\infty}[\![p(a^*) - p(a_t) > \epsilon]\!] > N\right\} < \delta$$

for all $|A|$-armed bandits.

A Naive Bandit Learner

Naive Bandit Learner

1: $L := \log \frac{2|A|}{\delta}$ and $m := \frac{2L}{\epsilon^2}$
2: Pull each arm m times for $r(a)$ accumulated reward.
3: $\hat{p}(a) := r(a)/m$
4: **loop**
5: Pull arm $\hat{a}^* := \arg\max \hat{p}(a)$

Sample Complexity

Theorem

The naive bandit learner has sample complexity of $O\left(\frac{2|A|}{\epsilon^2}\log\frac{2|A|}{\delta}\right)$.

Proof.

1. By Hoeffding's bound $|\hat{p}(a) - p(a)| \leq \sqrt{\frac{L}{2m}} \leq \epsilon/2$ with probability $1 - \delta/|A|$.
2. By the union bound this holds for all a with probability at least $1 - \delta$.

Let $t \geq m|A|$ then with probability at least $1 - \delta$

$$\begin{aligned} p(a^*) - p(a_t) &\leq \hat{p}(a^*) - p(a_t) + \epsilon/2 \\ &\leq \hat{p}(a_t) - p(a_t) + \epsilon/2 \\ &\leq \epsilon \end{aligned}$$

$\square$

Remarks

Bandit specialists would be very unexcited about the Naive Bandit Learner for a few reasons:

1. Although it has a uniform (and optimal) sample-complexity bound, it achieves this bound on all bandits, even easy ones.
2. It has a linear (hopeless) regret bound.
3. The algorithm depends on ϵ and δ. Many modern bandits algorithms have optimal sample-complexity bounds with no dependence on ϵ/δ.
4. It only works for stationary discrete bandits.