Competitive Analysis of the Explore/Exploit Tradeoff

John Langford, Martin Zinkevich, Sham Kakade

You fly to a new continent.

You fly to a new continent.

You are hungry.

You fly to a new continent.

You are hungry.

You walk in search of food.

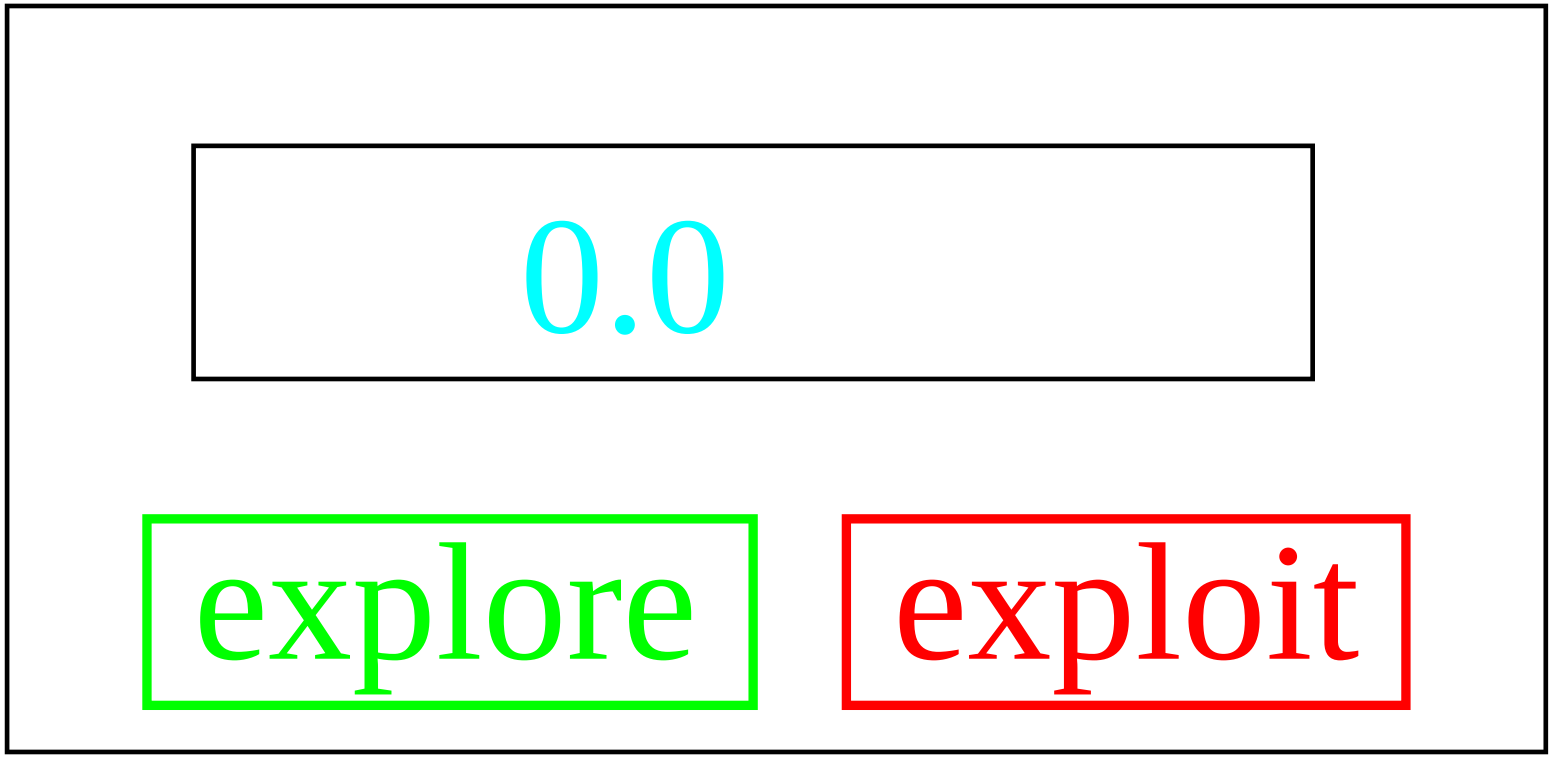
You fly to a new continent.

You are hungry.

You walk in search of food.

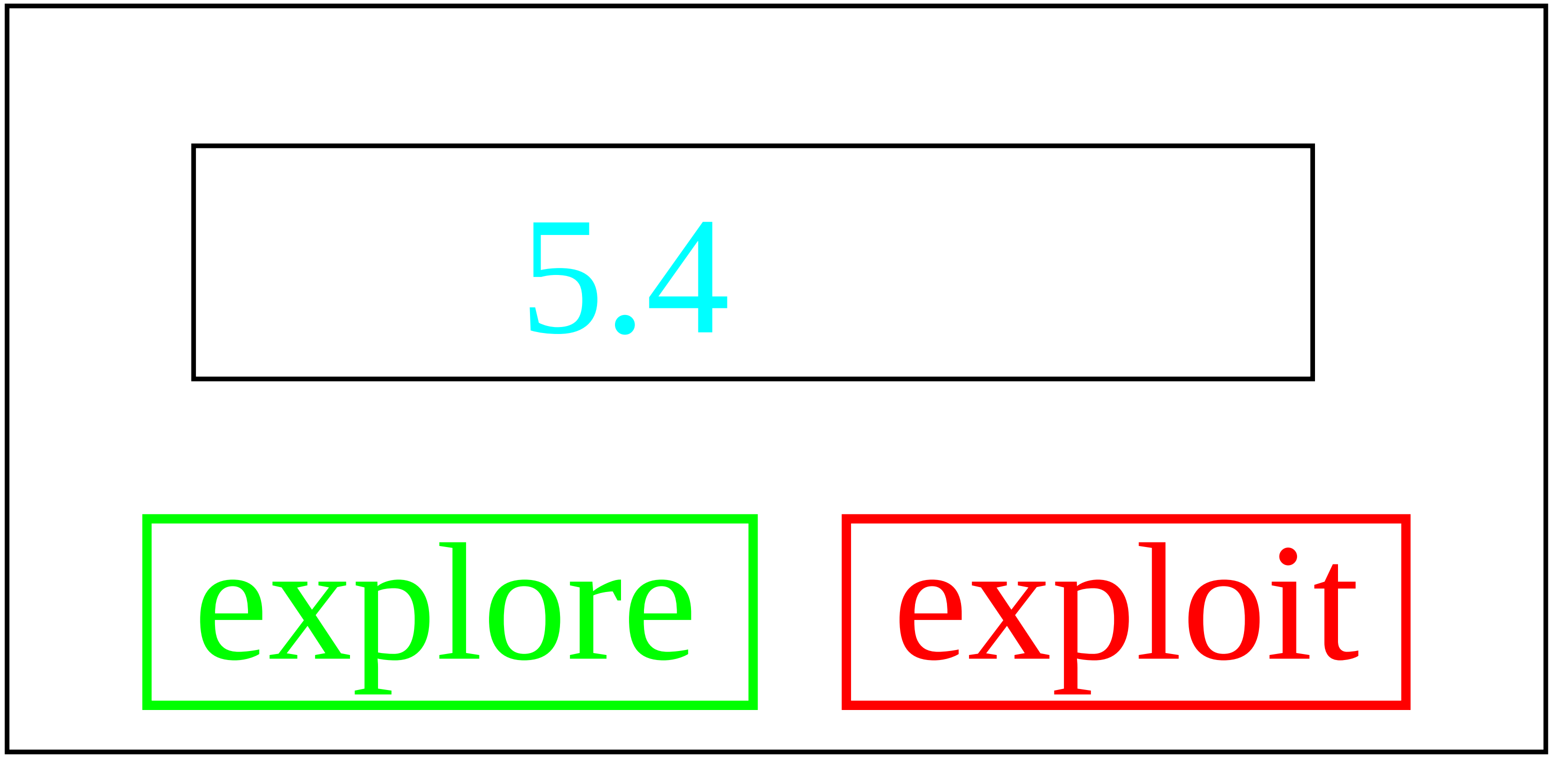
How long should you search before choosing a food and eating?

The Model



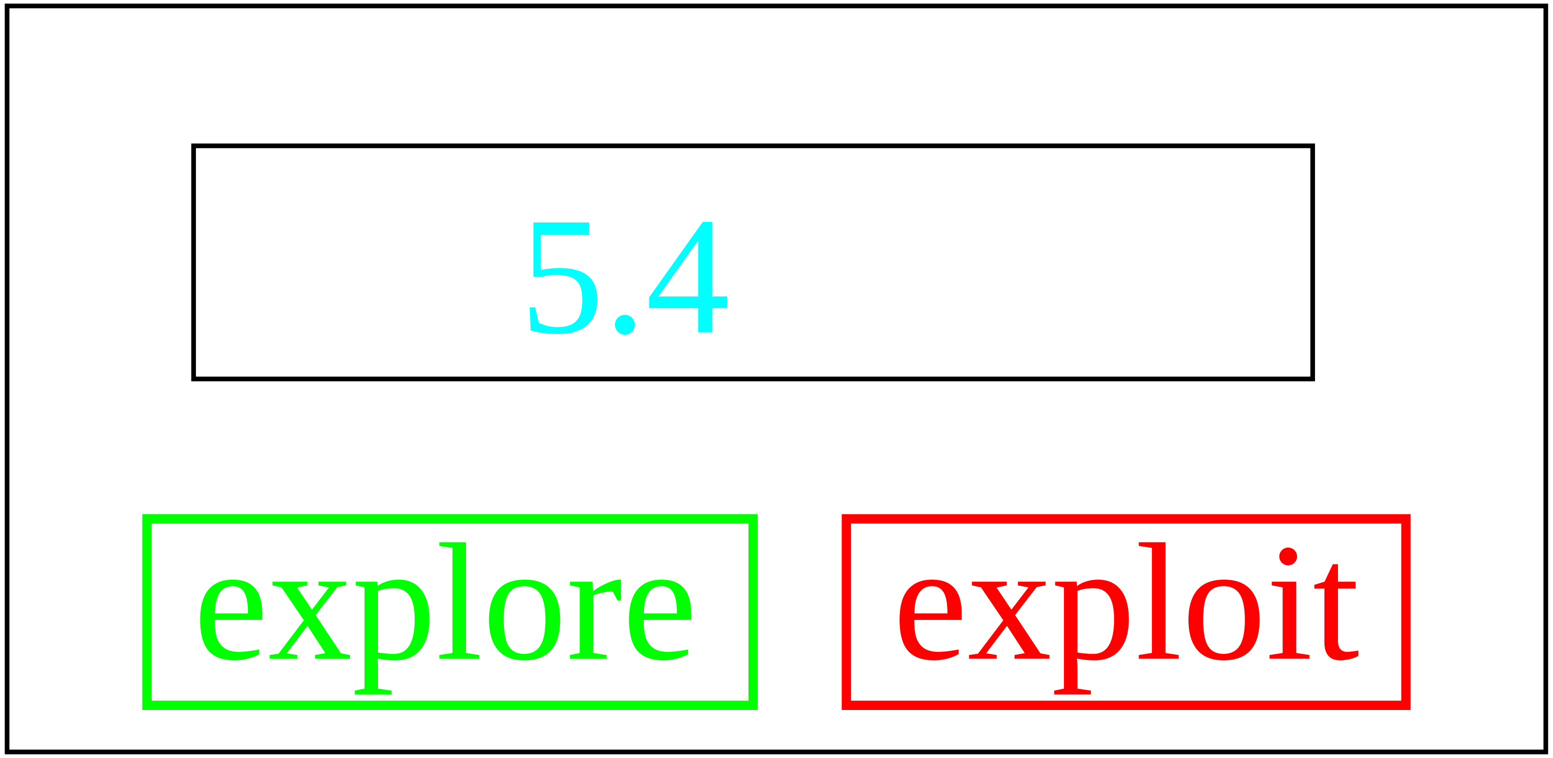
total reward recieved: 0.0

Explore



total reward recieved: 0.0

Exploit



total reward recieved: 5.4

Given: Exploit value

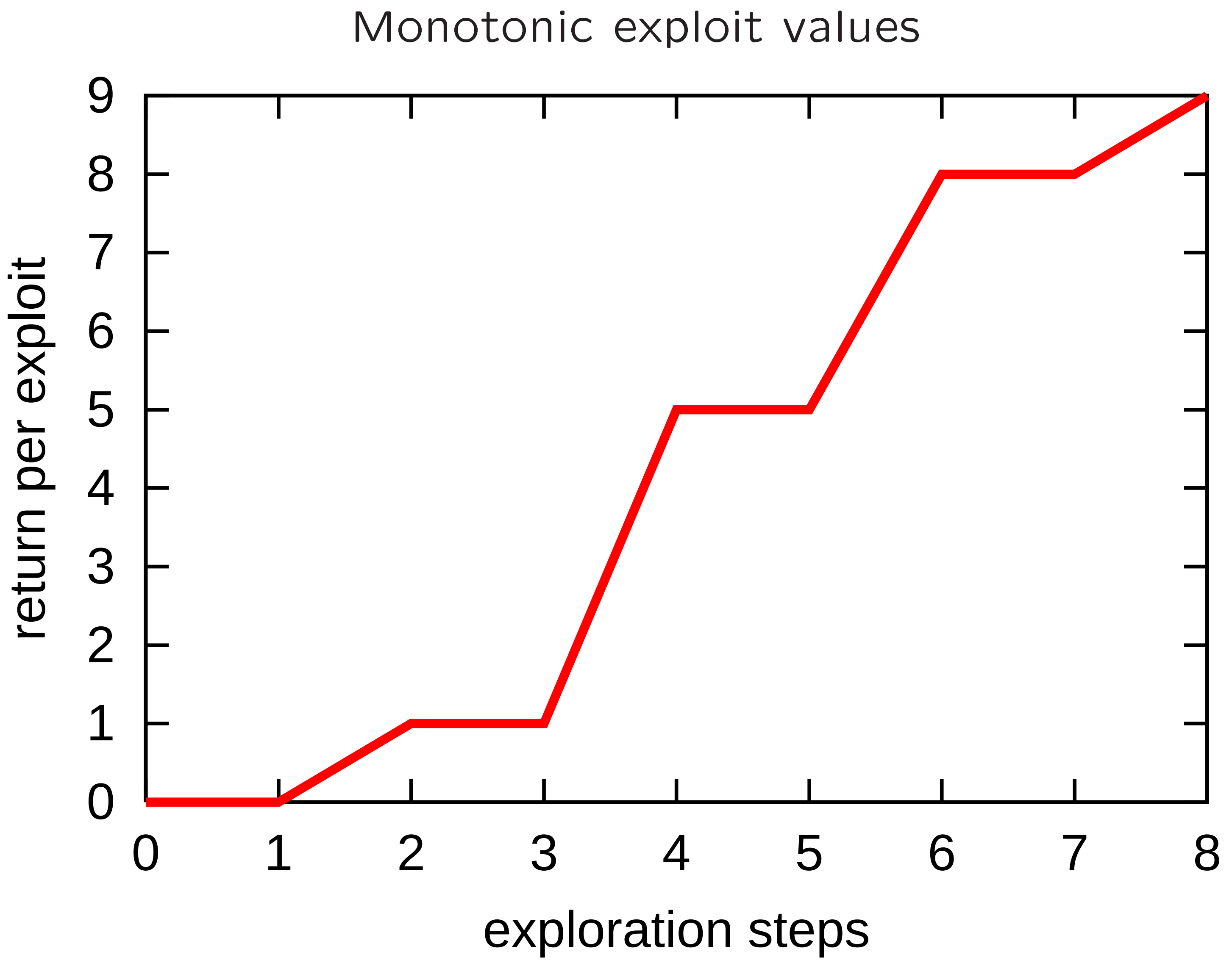
Choose “Explore” or “Exploit”

“Explore” $\rightarrow$ Exploit value increases

“Exploit” $\rightarrow$ Achieve reward

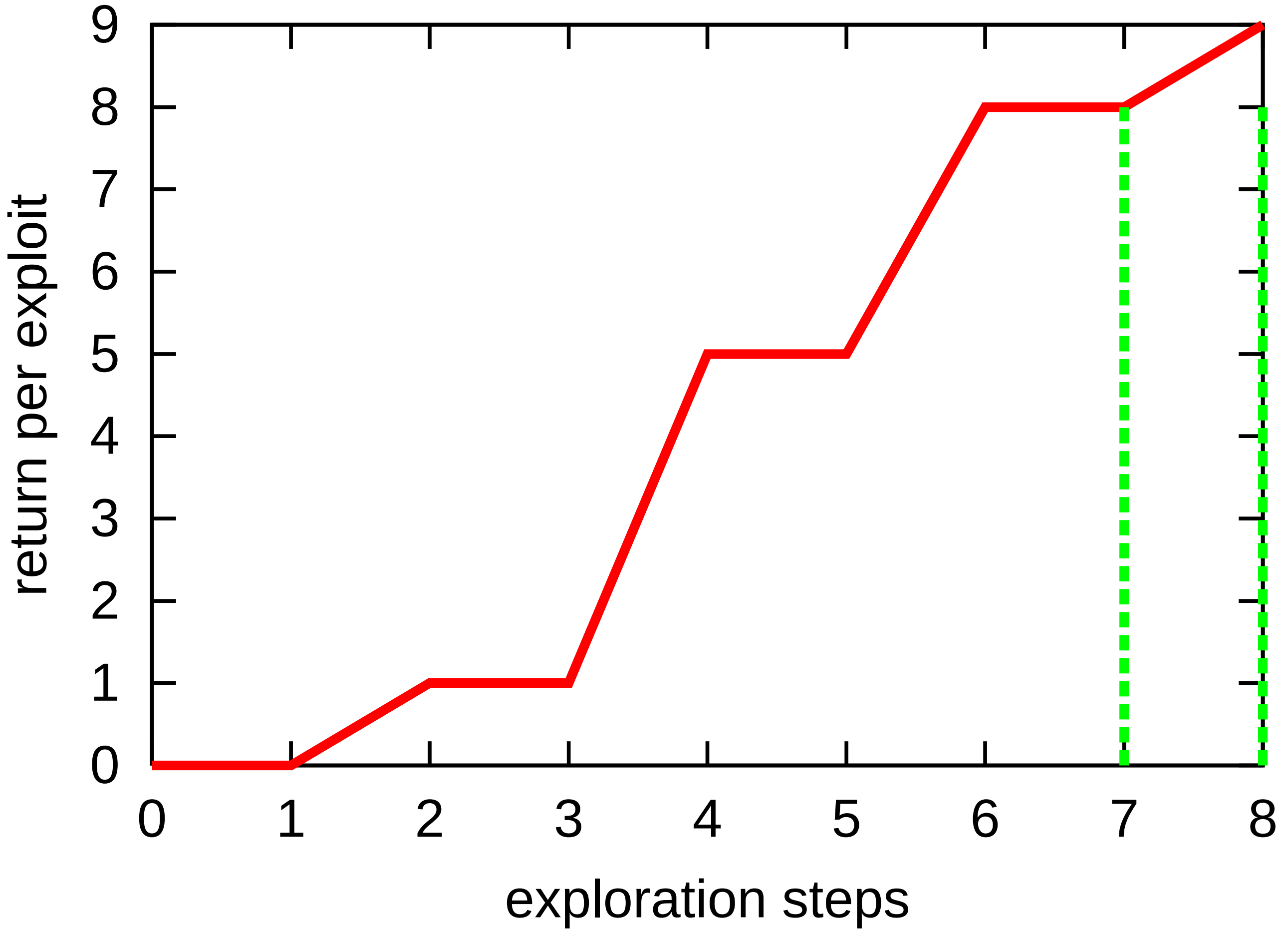
Notes

- Captures notion that optimal policy must explore.
- Idealized: General results $\Rightarrow$ weakly applicable.
- Stateless: Not an MDP model



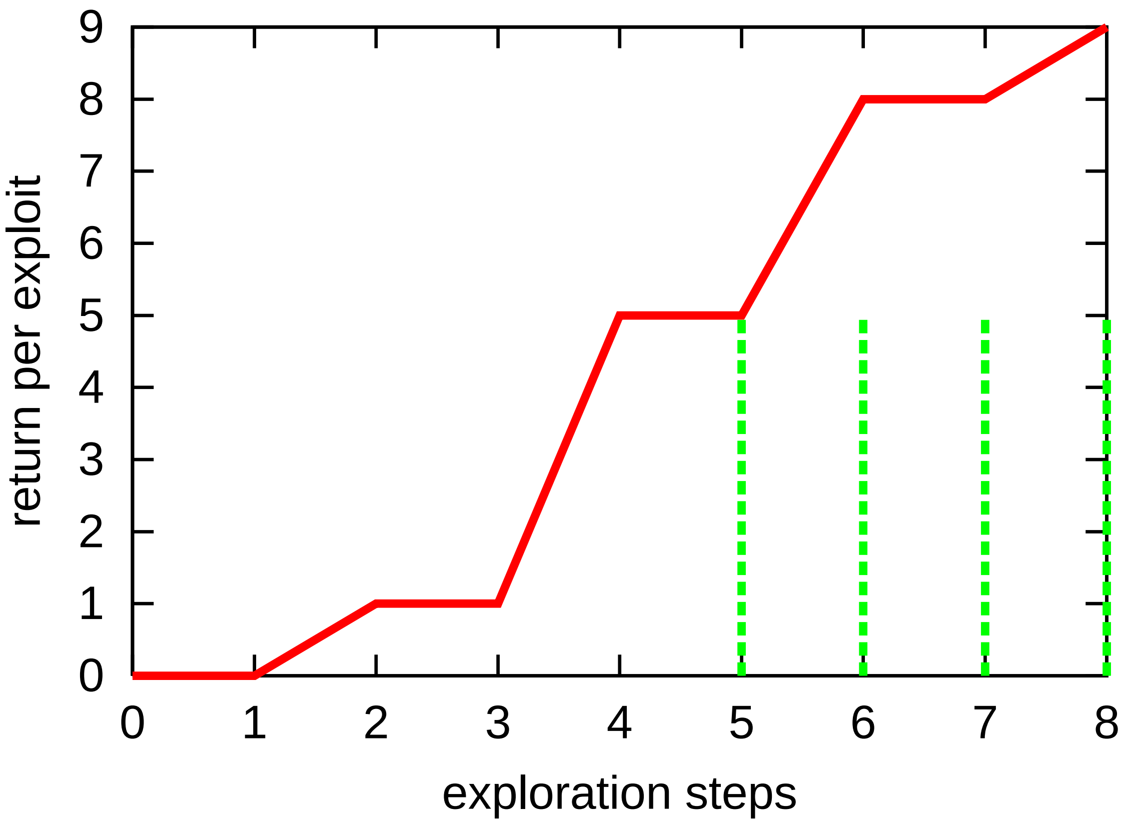
Monotonic exploit values
return per exploit
exploration steps
0
1
2
3
4
5
6
7
8
9

Algorithm Performance



Cumulative: $C_A(s) = 8 + 8 = 16$

Optimal Algorithm



Optimal cumulative: $C_*(s) = 5 + 5 + 5 + 5 = 20$

(note: single switch)

Competitive Analysis

What is the ratio of returns between optimal algorithm and algorithm A?

sequence of exploit values $= s$

$$\frac{C_*(s)}{C_A(s)}$$

Competitive Analysis II

What is the *worst* ratio between optimal algorithm and algorithm A?

$$\max_s \frac{C_*(s)}{C_A(s)}$$

Competitive Analysis III

What is the *worst* ratio between optimal algorithm and the *expected* value of A?

$$\max_s \frac{C_*(s)}{E_r C_{A_r}(s)}$$

Let R_* = exploitation value of optimal algorithm

Quantities:

1. T = the number of time steps

2. $\overline{R}, \underline{R}$ upper and lower bounds on R_*

1. Upper bound $\Rightarrow$ best algorithm we can find

2. Lower bound $\Rightarrow$ no algorithm can do better on all problems *given only the available information.*

Results, Deterministic

	Upper bound	Lower Bound
Nothing	∞	∞
$\overline{R}, \underline{R}$	$\frac{\overline{R}}{\underline{R}}$	$\frac{\overline{R}}{2\underline{R}}$
T	T	T
$T, \overline{R}, \underline{R}$	$\min\left(T, \frac{\overline{R}}{\underline{R}}\right)$	$\min\left(T, \frac{\overline{R}}{2\underline{R}}\right)$

Infinity

Either:

1. The algorithm never exploits (implying ∞ competitive ratio).

2. or the algorithm exploits at time t.

If it exploits at time t, then the tth exploration may reveal an arbitrarily larger reward.

Results, Stochastic

upper bound: $g(x) = 1 + \ln x$

lower bound: $f(x) = \Omega\left(\frac{\log x}{\log\log x}\right)$

	Upper bound	Lower bound
Nothing	$\forall\epsilon : \frac{1+\epsilon}{\epsilon} T^{1+\epsilon}$	$T - 1$
$\overline{R}, \underline{R}$	$g\left(\frac{\overline{R}}{\underline{R}}\right)$	$f\left(\frac{\overline{R}}{\underline{R}}\right)$
T	$g(T)$	$f(T)$
$T, \overline{R}, \underline{R}$	$g\left(\min\left(T, \frac{\overline{R}}{\underline{R}}\right)\right)$	$f\left(\min\left(T, \frac{\overline{R}}{\underline{R}}\right)\right)$

Results, algorithm given T

Consider times $\{T, T-1, T-2, T-4, T-8, ...T-2^i\}$

uniform explore$\rightarrow$exploit switch on these times

Results, algorithm given T

Consider times $\{T, T-1, T-2, T-4, T-8, ...T-2^i\}$

uniform explore$\rightarrow$exploit switch on these times

t_* = optimal exploit time

Results, algorithm given T

Consider times $\{T, T-1, T-2, T-4, T-8, ...T-2^i\}$

uniform explore$\rightarrow$exploit switch on these times

t_* = optimal exploit time

Probability P of switching for $t \in \left\{t_*, \frac{t_*}{2}\right\}$

Results, algorithm given T

Consider times $\{T, T-1, T-2, T-4, T-8, ...T-2^i\}$

uniform explore→exploit switch on these times

$t_* =$ optimal exploit time

Probability P of switching for $t \in \left\{t_*, \frac{t_*}{2}\right\}$

$\Rightarrow R_A \geq R_*$

Results, algorithm given T

Consider times $\{T, T-1, T-2, T-4, T-8, ...T-2^i\}$

uniform explore→exploit switch on these times

t_* = optimal exploit time

Probability P of switching for $t \in \left\{t_*, \frac{t_*}{2}\right\}$

$\Rightarrow R_A \geq R_*$

$\Rightarrow E_r C_A \geq P R_A t \geq P \frac{R_* t_*}{2}$

Implications

1. Optimization of explore-exploit is *much* easier given T or $\overline{R}, \underline{R}$.

2. Worst-case optimization is *much* better with randomized algorithms.

Questions to Think About

1. Can we reduce the variance of the stochastic algorithm?

2. Can we add structure to the model to get results for MDPs?