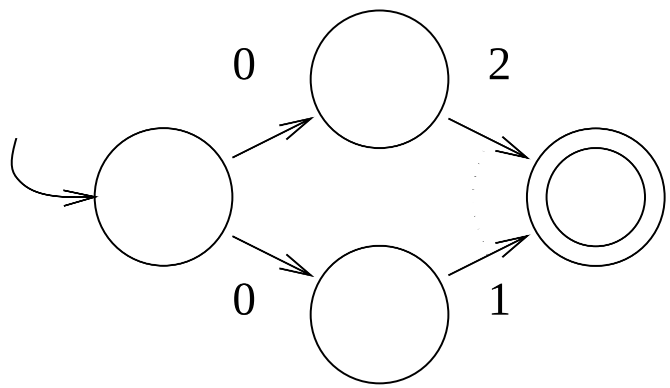
Reinforcement Learning with Function Approximation Converges to a Region

Geoff Gordon
ggordon@cs.cmu.edu
ggordon@burning-glass.com

December, 2000

RL Example



True action values	A possible approximation
$Q(\text{left,up}) = 2$	2
$Q(\text{left,down}) = 1$	1
$Q(\text{top}) = 2$	1.1
$Q(\text{bottom}) = 1$	1.1
$Q(\text{goal}) = 0$	0

One view of RL in large problems

Given experience with an MDP

Find an approximate value function

- say, from a linear subspace of all value fns

Do shallow lookahead to determine policy

Note: good recent work on learning policies w/o explicit values

Convergence of RL Algorithms

Want: fast, no local optima, good def'n of optimum

Can't always get what you want

Many algorithms defined as update rules

- no notion of minimizing an objective
- no measure of how good a candidate is
- no guarantee that fixed points exist

Counterexamples

Counterexamples exist for:

- offline value iteration
- Q-learning
- SARSA
- online value iteration

That is, MDPs exist for which these update rules produce a sequence that doesn't converge

Not all counterexamples created equal

Q and offline VI vs. SARSA and online VI

Divergence vs. oscillation

Real difference or gap in our understanding?

If real, heuristic explanation why SARSA and online VI are more successful

Is this difference real?

Yes.

Is this difference real?

If policy changes only between episodes, SARSA and online VI can never diverge

Instead, converge to a region (but not necessarily a point)

Theorem

For any MDP M ... there is a bounded region R such that the sequence of weight vectors produced by the SARSA(0) algorithm on M converges to R w. p. 1.

And similarly for $V(0)$

Proof intuition I

Policy fixed in each episode

If policy stayed fixed in all episodes, would converge

Expected update is 2-norm contraction

Proof intuition II

All fixed points are in some bounded region

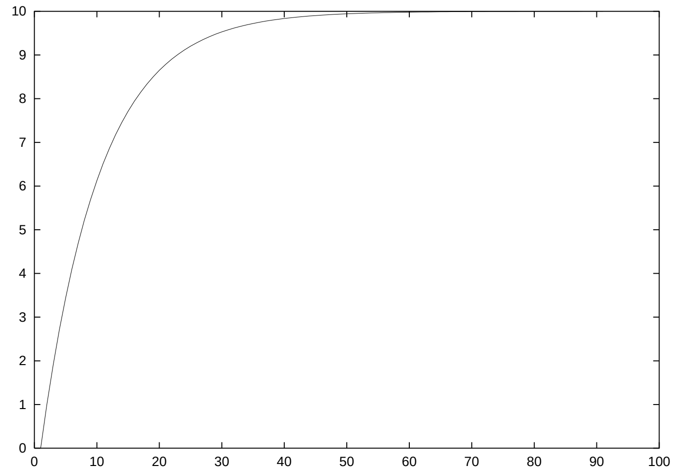
Far enough from that region, all fixed points equivalent

Go towards one $\Rightarrow$ go towards all

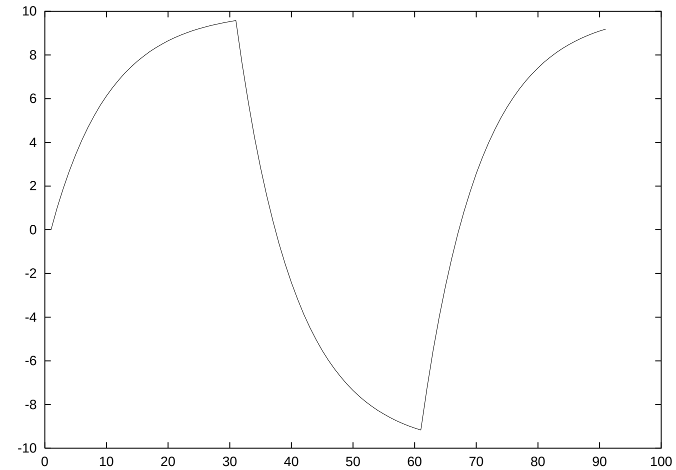
Proof intuition III

Update is 2-norm contraction plus bounded perturbation

Like $x_{t+1} \leftarrow .9x_t + k_t$



k_t constant, sequence converges



k_t bounded, sequence can oscillate

What now?

Just proved SARSA, online VI can't diverge

Already know they don't always converge

Only question left: how big are the oscillations

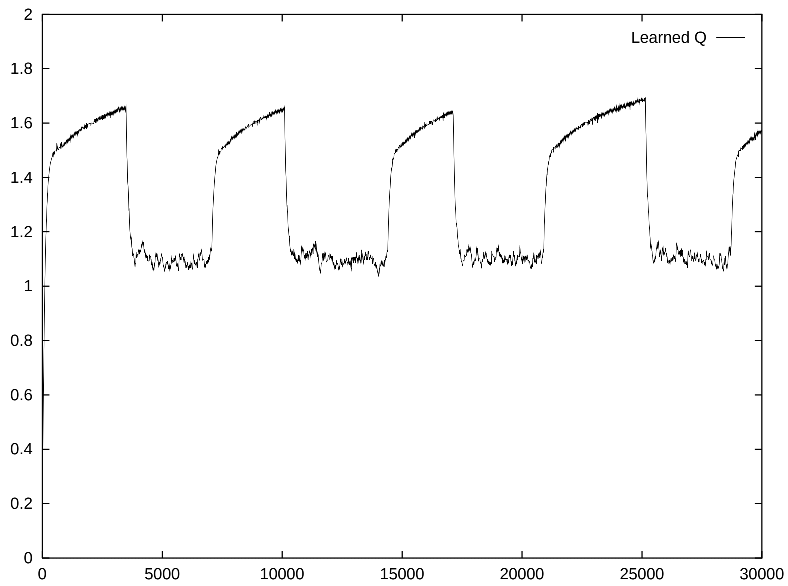
How big is R?

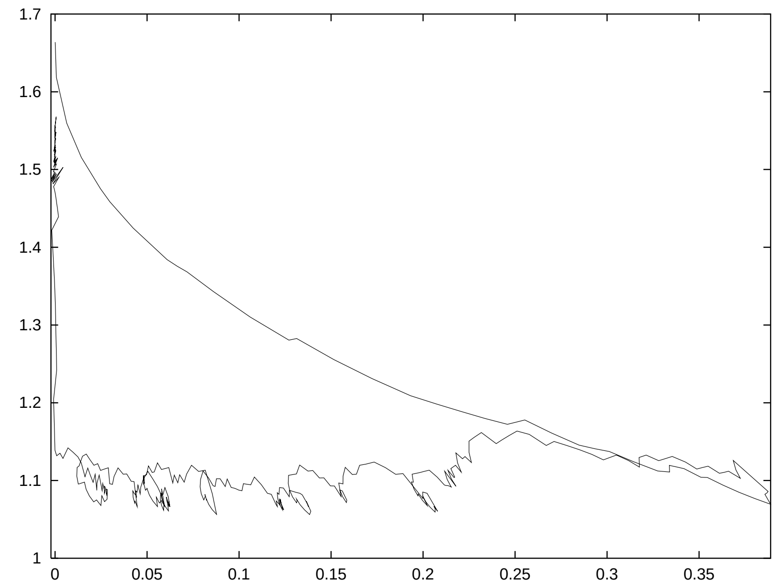
Current bound is loose: R contains all fixed points for all policies, then must be larger by a factor depending on slowest mixing time

If we could limit the number of policies, could tighten bound

Perhaps once we get near the true value function, only a few policies can be greedy

SARSA Counterexample

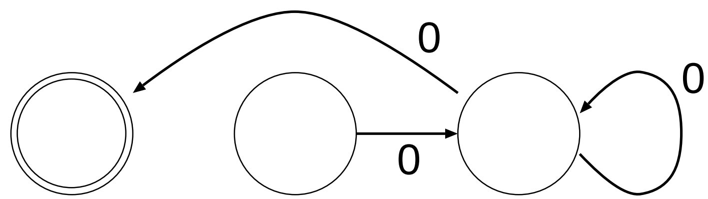


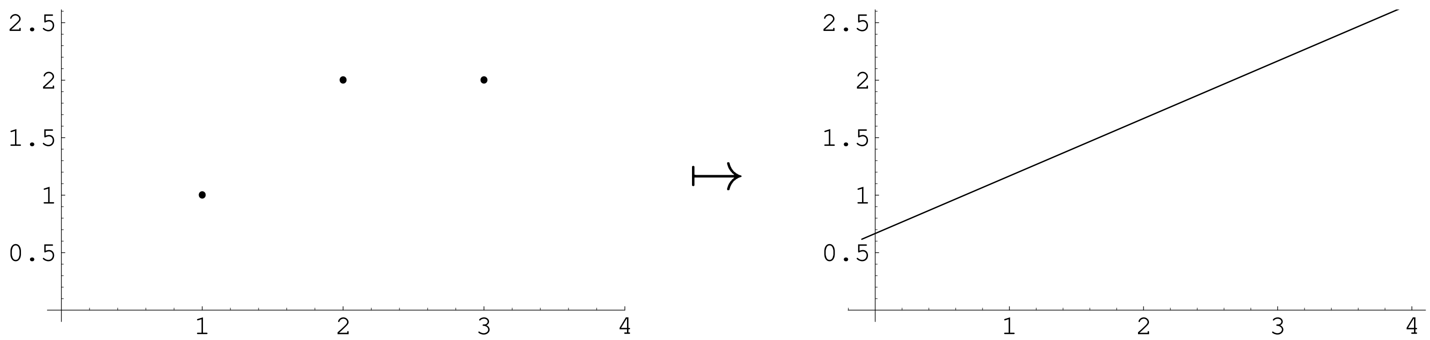


Same counterexample works for online VI

See [Gordon, 1996]

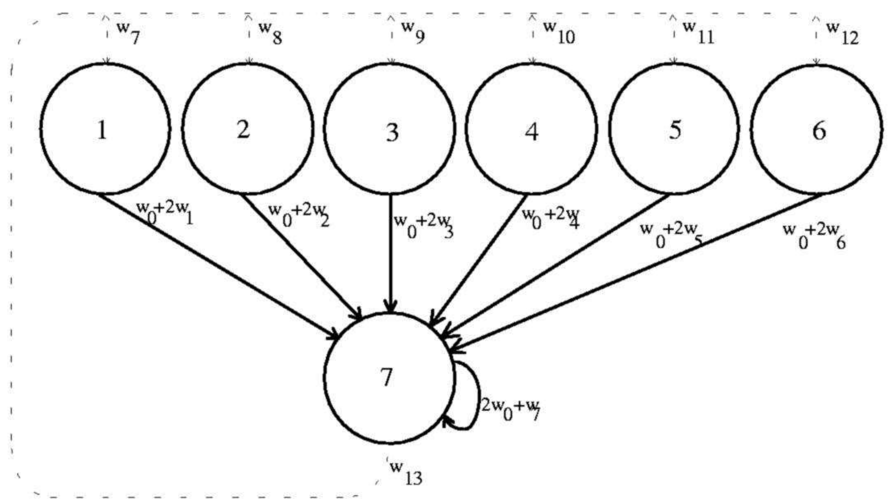
Value Iteration Counterexample





See [Gordon, 1995]

Q-Learning Counterexample



See [Baird, 1995]

Significance

Strongest convergence result for these algorithms

Good first step towards even stronger results

See also [de Farias & Van Roy, 2000]