3.1 Guarantees for policy gradient methods

What kinds of problems are policy gradient methods good at solving?

For example, consider a very simple MDP in which

outline

sparse reward setting

e.g. l/r, move to end; random policy $1/(2^n)$

no rewards early on, so no gradients

possible solutions

if simulator: use better starting

imitation learning (today)

exploration - ucb-vi

reward shaping

guarantees for pg

sl works in many settings

want to show that some benefits extend to rl

eg sample efficiency needed for softmax (log linear) policy

- eg under npg

what features do we need for good learning? (approximation error between ground truth and our function class)

hopefully samples poly(dim(), 1/eps)

need some coverage over state space

—

but convergence guarantees are hard

imitation learning

eg how humans learn by imitating experts

access to expert demonstrations

use sl to create a policy

input: senses output: action

setting

unknown reward function

assume expert has good policy

goal is to learn a policy as good as expert

—

BC

e.g. maximum likelihood (stochastic)

or classification error (deterministic)

or squared error for continuous actions

—

theorem: il is almost as easy as sl