PEGASUS: A Policy Search Method for Large MDPs and POMDPs

Andrew Ng and Michael Jordan
UC Berkeley

Presented at UAI 2000, Stanford

Summary

- An algorithm for searching a space of policies of MDPs and POMDPs.
- Can be applied to large (even infinite) state and action spaces.
- This is a direct policy search method.
- Requires a *deterministic* simulative model.

Main idea

- Transform a stochastic (PO)MDP into a deterministic one and use the latter for evaluating and improving policies.
- Deterministic simulative model: $g : S \times A \times [0,1]^d \to S'$.
- Draw a sample $\{s_0^{(1)}, s_0^{(1)}, \ldots, s_0^{(m)}\}$, of m initial states according to the initial distribution of the original (PO)MDP.
- Choose a truncation horizon equal to the ϵ-horizon time $H = \log_\gamma(\epsilon(1-\gamma)/2R_{max}$.
- For each initial state $s_0^{(i)}$, $i = 1, m$, generate a *scenario* of Hd uniform random numbers. These will be reused in policy search.

Policy search

- Choose an initial policy $\pi(s|\theta)$, where θ parametrizes the policy in a smooth way ($d\pi/d\theta$ exists).
- For each scenario i, generate a sequence $(s_0^{(i)}, s_0^{(i)}, \ldots, s_H^{(i)})$ of H states by means of the deterministic simulative model and the current policy π.
- For each sequence of states, compute the corresponding rewards.
- The value $V(\pi)$ is the average of all mH rewards – it is a deterministic function of the policy for a fixed set of scenarios.
- Compute $dV(\pi)/d\theta$ numerically and use it to maximize $V(\pi)$.

Experimental results

- A simple problem: a gridworld POMDP with 25 states.
- A total of 65536 possible policies, which were searched exhaustively.
- Discounting factor $\gamma = 0.99$, horizon $H = 100$.
- The algoritm finds an almost optimal policy with as few as $m = 30$ scenarios.
- A very hard problem: a bycicle simulator with six-dimensional continuous state space.
- Actions are parametrized as sigmoids on 15 manually chosen features of state space.
- A total of $m = 30$ scenarios.
- Discounting factor $\gamma = 0.998$, horizon $H = 500$.
- The distance to the goal is $1km$ – the algorithm finds solutions ranging from 995 meters to $1.07km$.