Linear programming solution methods

A slightly less frequently described method for MDPs: solution via linear programming

Basic idea: we can capture the constraint

$$V(s) \geq R(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s, |s, a) V(s')$$

via the set of $|\mathcal{A}|$ linear constraints

$$V(s) \geq R(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) V(s'), \ \ \forall a \in \mathcal{A}$$

Now consider the linear program

$$\begin{array}{ll}\underset{V}{\text{minimize}} & \displaystyle\sum_{s} V(s) \\ \text{subject to} & \displaystyle V(s) \geq R(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V(s'), \;\; \forall a \in \mathcal{A}, s \in \mathcal{S}\end{array}$$

Theorem: the optimal value of this linear program will be $V^\star$

Proof: Suppose there exists some $s \in \mathcal{S}$ with

$$V(s) > R(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V(s')$$

Then we can construct a solution with only $V(s)$ changed to make this an equality: this will have a lower objective value, but be feasible, since it can only decrease right hand side for other constraints

Comments on LP formulation

In objective, we can optimize any positive linear function of $V(s)$ and the result above still holds

If we optimize

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & \sum_{s} d(s) V(s) \\ \text{subject to} \quad & V(s) \geq R(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V(s'), \quad \forall a \in \mathcal{A}, s \in \mathcal{S} \end{aligned}$$

where $d(s)$ is a distribution over states, then objective is equal to total expected accumluted reward when beginning at a state drawn from this distribution

Adding dual variables $\mu(s, a)$ for each constraint, dual problem is (after some simplification)

$$\begin{aligned} \underset{\mu(s,a)}{\text{maximize}} \quad & \sum_{s \in \mathcal{S}} R(s) \sum_{a \in \mathcal{A}} \mu(s, a) \\ \text{subject to} \quad & \sum_{a \in \mathcal{A}} \mu(s', a) = d(s') + \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P(s'|s, a)\mu(s, a) \;\; \forall s' \in \mathcal{S} \\ & \mu(s, a) \geq 0 \end{aligned}$$

These have the interpretation that

$$\mu(s, a) = \sum_{t=0}^{\infty} \gamma^t P(S_t = s, A_t = a)$$

i.e., they are discounted state-action counts, which directly encode the optimal policy

$$\pi^\star(s) = \max_{a \in \mathcal{A}} \mu(s, a)$$

LP versus value/policy iteration

Some surprising connections between LP formulation and standard value and policy iteration algorithms: e.g. a certain form of dual simplex is equivalent to policy iteration

Typically, best specialized MDP algorithms (e.g. modified policy iteration) are faster than general LP algorithms, but the LP formulation provides a number of connections to other methods, and has also been the basis for much work in approximate large-scale MDP solutions (e.g., de Farias and Van Roy, 2003)