Module 8
Linear Programming

CS 886 Sequential Decision Making and Reinforcement Learning

University of Waterloo

Policy Optimization

- Value and policy iteration
  - Iterative algorithms that implicitly solve an optimization problem
- Can we explicitly write down this optimization problem?
  - Yes, it can be formulated as a linear program

Primal Linear Program

primalLP(MDP)

$\min_V \sum_s w(s)V(s)$

subject to $V(s) \geq R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s') \quad \forall s, a$

return V

- Variables: $V(s)\ \forall s$
- Objective: $\min \sum_s w(s)V(s)$
  where $w(s)$ is a weight assigned to state s
- Constraints:
  $V(s) \geq R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s') \quad \forall s, a$

Objective

- Why do we minimize a weighted combination of the values? Shouldn't we maximize value?

- Value functions V that satisfy the constraints are upper bounds on the optimal value function V^*

$$V(s) \geq V^*(s) \;\; \forall s$$

- Minimizing value ensures that we choose the lowest upper bound

$$\min_{\mathrm{V}} V(s) = V^*(s) \;\; \forall s$$

Upper bound

- Theorem: Value functions V that satisfy $V(s) \geq R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s') \;\; \forall s, a$ are upper bounds on the optimal value function V^*

$$V(s) \geq V^*(s) \; \forall s$$

- Proof:
  - Since $V(s) \geq R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s') \;\; \forall s, a$
  - Then $V(s) \geq \max_a R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a)\, V(s') \; \forall s$

$$= H^*(V)(s) \; \forall s$$

  - Furthermore

$$V \geq H^*(V) \geq H^*(H^*((V)) \geq \cdots \geq (H^*)^\infty(V) = V^*$$

Weight function (initial state)

- How do we choose the weight function?

- If the policy always starts in the same initial state s_0, then set

$$w(s) = \begin{cases} 1 & s = s_0 \\ 0 & \text{otherwise} \end{cases}$$

- This ensures that $\sum_s w(s)V(s) = V^*(s_0)$

Weight function (any state)

- If the policy may start in any state, then assign a positive weight to each state, i.e. $w(s) > 0 \ \ \forall s$
- This ensures that V is minimized at each s and therefore $V(s) = V^*(s) \ \forall s$
- The magnitude of the weight doesn't matter when the LP is solved exactly. We will revisit the choice of $w(s)$ when we discuss approximate linear programming.

Optimal Policy

- Linear program finds V^*
- We can extract π^* from V^* as usual:

$$\pi^*(s) \leftarrow argmax_a \; R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) V^*(s')$$

- Or check the active constraints
  - For each s, check which a^* leads to equality

$$V(s) = R(s,a^*) + \gamma \sum_{s'} \Pr(s'|s,a^*) V(s')$$

$$V(s) \geq R(s,a) + \gamma \sum_{s'} \Pr(s'|s,a) V(s') \;\; \forall a$$

  - Set $\pi^*(s) \leftarrow a^*$

Direct Policy Optimization

- The optimal solution to the primal linear program is V^*, but we still have to extract π^*

- Could we directly optimize π?
  - Yes, by considering the dual linear program

Dual Linear Program

dualLP(MDP)

$\max_y \sum_{s,a} y(s,a)R(s,a)$

subject to $\sum_{a'} y(s',a') = b(s') + \gamma \sum_{s,a} \Pr(s'|s,a)y(s,a) \; \forall s$

$y(s,a) \geq 0 \;\; \forall s,a$

Let $\pi(a|s) = \Pr(a|s) = y(s,a) / \sum_a y(s,a)$

return π

- Variables: $y(s,a) \; \forall s,a$
  - frequency of each $\langle s,a \rangle$-pair (proportional to π)
- Objective: $\max_y \sum_{s,a} y(s,a)R(s,a)$
- Constraints: $\sum_{a'} y(s',a') = b(s') + \gamma \sum_{s,a} \Pr(s'|s,a)y(s,a)$

Duality

- For every primal linear program in the form

$$\min_x c^T x$$
$$\text{s.t. } Ax \geq b$$

- There is an equivalent dual linear program in the form

$$\max_y b^T y$$
$$\text{s.t. } A^T y = c \text{ and } y \geq 0$$

- Where $\min_x c^T x = \max_y b^T y$

Interpretation:

$c = w$

$x = V$

$y \propto \pi$

$A = [I - \gamma T^a] \forall a$

$b = [R^a] \forall a$

State Frequency

- Let $f(s)$ be the frequency of s under policy π.

0 step: $f_0(s) = w(s)$

1 step: $f_1(s') = w(s') + \gamma \sum_s \Pr(s'|s, \pi(s)) w(s)$

2 steps: $f_2(s'') = w(s'') + \gamma \sum_{s'} \Pr(s''|s', \pi(s')) w(s')$

$$+\gamma^2 \sum_{s,s'} \Pr(s''|s', \pi(s')) \Pr(s'|s, \pi(s)) w(s)$$

…

n steps:

$$f_n(s^{(n)}) = w(s^{(n)}) + \gamma \sum_{s^{(n-1)}} \Pr(s^{(n)}|s^{(n-1)}, \pi(s^{(n-1)})) f_{n-1}(s^{(n-1)})$$

∞ steps: $f(s') = w(s') + \gamma \sum_s \Pr(s'|s, \pi(s)) f(s)$

State-Action Frequency

- Let $y(s,a)$ be the state-action frequency

$$y(s,a) = \pi(a|s) f(s)$$

where $\pi(a|s) = \Pr(a|s)$ is a stochastic policy

- Then the following equations are equivalent

$$f(s') = w(s') + \gamma \sum_s \Pr(s'|s,\pi(s))\, f(s)$$

$$\Leftrightarrow \sum_{a'} \pi(a'|s')\, f^{\pi}(s') = w(s') + \sum_s \Pr(s'|s,a)\, \pi(a|s) f^{\pi}(s)$$

$$\Leftrightarrow \underbrace{\sum_{a'} y(s',a') = w(s') + \sum_s \Pr(s'|s,a)\, y(s,a)}$$

Constraint of dual LP

Policy

- We can recover π from y.

$$y(s,a) = \pi(a|s)\, f(s) \qquad \text{(by definition)}$$

$$\pi(a|s) = \frac{y(s,a)}{f(s)} \qquad \text{(isolate } \pi\text{)}$$

$$\pi(a|s) = \frac{y(s,a)}{\sum_a y(s,a)} \qquad \text{(by definition)}$$

- π may be stochastic
- Actions with non-zero probability are necessarily optimal

Objective

- Duality theory guarantees that the objectives of the primal and dual LPs are equal

$$\max_{y} \sum_{s,a} y(s,a)R(s,a) = \min_{V} \sum_{s} w(s)\,V(s)$$

- This means that $\sum_{s,a} y(s,a)R(s,a)$ implicitly measures the value of the optimal policy.

Solution Algorithms

- Two broad classes of algorithms:
  - Simplex (corner search)
  - Interior point methods (interior iterative methods)
    - Polynomial complexity (MDP is in P, not NP)
- Many packages for linear programming
  - CPLEX (robust, efficient and free for academia)