IE 3186: Approximate Dynamic Programming **Fall 2018**

Lecture 13: Benchmarking & Information Relaxation

Lecturer: Daniel Jiang *Scribes: Ibrahim El Shar, Boyuan Lai*

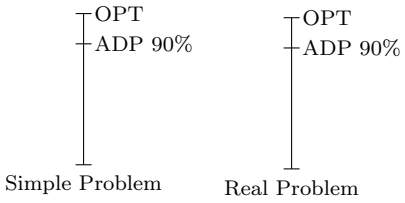
References:

D. B. Brown, J. E. Smith, P. Sun. *Information relaxations and duality in stochastic dynamic programs.* Operations Research, 58(4-part-1), 785-801, 2010.

13.1 Benchmarking

How do we *test the optimality* of an ADP algorithm on a specific problem instance?

1. Benchmark on a "simple" problem for which you can solve the MDP. The logic is that: if on a simple problem we can use ADP to compute a solution (which may in itself take many hours or days to compute) that is, e.g., 90% of the optimal objective then hopefully with normalized computation we can also achieve 90% of the optimal objective on the real problem. Repeating this experiment across a number of distinct problem domains can be convincing.



2. Or, test directly on the real problem instance (one that you can't solve) and compare against baselines/heuristics. Often the default route; self-explanatory.

3. Alternatively: produce an upper-bound on the optimal solution of a problem that you cannot solve.

Here, one would make a statement such as: if $\frac{\text{ADP}}{\text{UB}} = 80\%$, then $\frac{\text{ADP}}{\text{OPT}} \geq 80\%$. There are various ways to do this. In this lecture, we discuss the information relaxation approach to producing this upperbound.

13.2 Information Relaxation and Duality in MDPs

We will be in the setting of maximizing rewards. Any feasible policy is a lower bound. The main idea of Brown et. al. is that to get an upper bound, we can relax nonanticipativity constraints (i.e., constraints requiring decisions to depend only on the information available at the time a decision is made).

13.2.1 Framework

Slightly more general DP framework.

- $\mathbb{F} = (\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_T)$ (standard filtration),
- DM (decision's maker's) state of information,
- $\mathcal{A}_t$ is the set of actions at time t,
- $\mathbb{A} = \mathcal{A}_0 \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_T$,
- Standard DP: the policy π_t depends only on information in $\mathcal{F}_t$, and this information is summarized by the state variable.
- For each $\omega \in \Omega$, π will select a sequence of actions $a_0, a_1, \ldots, a_T$. $(\pi : \Omega \to \mathbb{A})$.
- Let $\Pi_{\mathbb{F}}$ be all nonanticipative policies.
- Let $r_t(a) = r_t(a_0, a_1, \ldots, a_t)$ be a reward function at time t. Let $r(\pi)$ be the total reward generated by following π.
- We can write the recursion:

$$V_t(a_0, a_1, \ldots, a_{t-1}) = \sup_{a_t \in \mathcal{A}_t} \{r_t(a_0, \ldots, a_t) + \mathbf{E}[V_{t+1}(a_0, \ldots, a_t) \,|\, \mathcal{F}_t]\}.$$

13.2.2 The Dual approach

Let $\mathbb{G}$ be a relaxed filtration, i.e., for all t, $\mathcal{F}_t \subseteq \mathcal{G}_t \iff \mathbb{F} \subseteq \mathbb{G}$. Under $\mathbb{G}$, DM knows more information and since $\mathcal{F}_t \subseteq \mathcal{G}_t$, we know that $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. Hence, we also have

$$\sup_{\pi \in \Pi_{\mathbb{F}}} \mathbf{E}[r(\pi)] \leq \sup_{\pi \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi)].$$

Example 13.1 (Perfect Information)**.** *At every time period t, you know information all of the information, as if you were at T. To be precise, we define $\mathbb{I} = \{\mathcal{F}, \mathcal{F}, \ldots, \mathcal{F}\}$, i.e., all randomness is known at every period. Thus, we have $\Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}} \subseteq \Pi_{\mathbb{I}}$ and*

$$\sup_{\pi \in \Pi_{\mathbb{F}}} \mathbf{E}[r(\pi)] \leq \mathbf{E}\left[\sup_{a \in \mathcal{A}} r(a)\right] = \sup_{\pi \in \Pi_{\mathbb{I}}} \mathbf{E}[r(\pi)].$$

Main issue: this upper bound can (and is expected to) be weak. How can we improve it? Solution: we could somehow *penalize* the use of future information.

Dual Feasible Penalty

A *dual feasible penalty* z satisfies

$$\mathbb{Z}_{\mathbb{F}} = \{z : \mathbf{E}[z(\pi)] \leq 0, \text{ for all } \pi \in \Pi_{\mathbb{F}}\}.$$

Here z is defined similar to the way we defined r. Note that z assigns no penalty (negative penalty) to nonanticipative π.

Lemma 13.2 (Weak Duality)**.** *For any $\pi_F \in \Pi_{\mathbb{F}}$,*

$$\mathbf{E}[r(\pi_F)] \leq \sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z(\pi_G)].$$

Proof. We have $\mathbf{E}[r(\pi_F)] \leq \mathbf{E}[r(\pi_F) - z(\pi_F)] \leq \sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z(\pi_G)]$.

The first inequality holds because $z \in \mathbb{Z}_{\mathbb{F}}$ (thus $\mathbf{E}[z(\pi_F)] \leq 0$) and the second because $\pi_F \in \Pi_{\mathbb{F}} \subseteq \Pi_{\mathbb{G}}$. □

Corollary 13.3 ($\mathbb{G} = \mathbb{I}$)**.** *Perfect information case: for any $\pi_F \in \mathcal{A}_{\mathbb{F}}$ and $z \in \mathbb{Z}_{\mathbb{F}}$ we have*

$$\mathbf{E}[r(\pi_F) - z(\pi_F)] \leq \mathbf{E}\Big[\sup_{a \in \mathcal{A}}\{r(a) - z(a)\}\Big]. \tag{13.1}$$

Remark 13.4. *The upper bound in* (13.1) *is good for simulation. We can estimate the expected value of the right hand side of* (13.1) *by (1) generating samples (2) optimizing deterministically (3) taking the average.*

If $\mathbb{G} \neq \mathbb{I}$, then we have "partial future knowledge" e.g. maybe we know demand, but not price in an inventory problem. In this case, the simulation of the bound involves (1) generating a sample of future demands, (2) solve simpler stochastic DP, where the demand is deterministic, (3) average.

Theorem 13.5 (Strong Duality)**.** *Let $\mathbb{G}$ be a relaxation of $\mathbb{F}$. Then,*

$$V^* = \sup_{\pi_F \in \Pi_{\mathbb{F}}} \mathbf{E}[r(\pi_F)] = \inf_{z \in Z_{\mathbb{F}}} \Big\{ \sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z(\pi_G)] \Big\}. \tag{13.2}$$

Proof. By weak duality, we have

$$\sup_{\pi_F \in \Pi_{\mathbb{F}}} \mathbf{E}[r(\pi_F)] \leq \inf_{z \in Z_{\mathbb{F}}} \{ \sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z(\pi_G)] \}.$$

Let $z^*(a) = r(a) - V^*$ and note that $\mathbf{E}[z^*(\pi)] = \mathbf{E}[r(\pi)] - V^*$ but since $\mathbf{E}[r(\pi)] \leq V^*$ for all $\pi \in \Pi_F$ (by defn), z^* is dual feasible. If $V^* < \infty$ then so is the RHS of (13.2). So RHS $= V^*$ (plugin z^* in (13.2)). □

Theorem 13.6 (Complementary Slackness)**.** *Let π_F^* and z^* be feasible to primal, dual problems with respect to information relaxation $\mathbb{G}$. They are optimal to their respective problems if and only if (1) $\mathbf{E}[z^*(\pi_F^*)] = 0$ and (2)*

$$\mathbf{E}[r(\pi_F^*) - z^*(\pi_F^*)] = \sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z^*(\pi_G)].$$

Proof. ($\Leftarrow$ direction). By (1) and (2) we have

$$\sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z^*(\pi_G)] = \mathbf{E}[r(\pi_F^*)]$$

By weak duality, π_F^* and z^* must be optimal.

($\Rightarrow$ direction). For any $\pi_F^* \in \Pi_{\mathbb{F}}$, $z^* \in Z_{\mathbb{F}}$

$$\begin{aligned}
\sup_{\pi_G \in \Pi_{\mathbb{G}}} \mathbf{E}[r(\pi_G) - z^*(\pi_G)] &\geq \sup_{\pi_F \in \Pi_{\mathbb{F}}} \mathbf{E}[r(\pi_F) - z^*(\pi_F)] \\
&\geq \mathbf{E}[r(\pi_F^*) - z^*(\pi_F^*)] \quad (\pi_F^* \in \Pi_{\mathbb{F}}) \\
&\geq \mathbf{E}[r(\pi_F^*)] \quad (z^* \in Z_{\mathbb{F}})
\end{aligned}$$

By strong duality, the first and last terms are equal. Thus: $\mathbf{E}[z^*(\pi_F^*)] = 0$ and so

$$\mathbf{E}[r(\pi_F^*) - z^*(\pi_F)] = \sup_{\pi_F \in \Pi_{\mathbb{F}}} E[r(\pi_F) - z^*(\pi_F)]$$

as desired. □

Interpretation:

- With optimal penalty, DM chooses nonanticipative policy, even though it has the option to choose otherwise.
- Optimal penalty assigns zero to optimal primal policy.

Proposition 13.7 (Good Penalty)**.** *Let filtration $\mathbb{G}$ be an information relaxation of $\mathbb{F}$. Let $(w_0(a,\omega), w_1(a,\omega), \ldots, w_T(a,\omega))$ be a sequence of generating functions on $\mathcal{A} \times \Omega$, where w_t depends only on the first $t+1$ actions of a: $(a_0, a_1, \ldots, a_t)$. Let $z_t(a) = \mathbf{E}[w_t(a) \,|\, \mathcal{G}_t] - \mathbf{E}[w_t(a) \,|\, \mathcal{F}_t]$ and $z(a) = \sum_{t=0}^{T} z_t(a)$. Then:*

1. *For all $\pi_F \in \Pi_{\mathbb{F}}$, $\mathbf{E}[z_t(a) \,|\, \mathcal{F}_t] = 0$ and $\mathbf{E}[z(\pi_F)] = 0$.*
2. *$(z_0(a), \ldots, z_T(a))$ is adapted to $\mathbb{G}$ and depends on the first $t+1$ actions of a.*

Part 1 says that this choice of penalty assigns all nonanticipative policies zero penalty. Part 2 says that the penalized objective, $r - z$, can be decomposed and solved as another DP.

$$V_t^{\mathcal{G}}(a_0, \ldots, a_{t-1}) = \sup_{a_t} \{ r_t(a_0, \ldots, a_t) - z_t(a_0, \ldots, a_t) + \mathbf{E}[\, V_{t+1}^{\mathcal{G}}(a_0, \ldots, a_t) \,|\, \mathcal{G}_t \,] \}$$

Theorem 13.8 (Ideal Penalty)**.** *Consider letting $w_t(a) = V_{t+1}(a_0, \ldots, a_t)$ and let $\mathbb{G} = \mathbb{I}$. Then,*

$$z_t(a) = V_{t+1}(a_0, \ldots, a_t) - \mathbf{E}[\, V_{t+1}(a_0, \ldots, a_t) \,|\, \mathcal{F}_t \,]$$

is dual feasible and optimal.

Therefore, the optimal penalty can be constructed using the optimal value function. This suggests than value function approximations from ADP could be a good choice.