Assignment 3. Due on 8 Dec 11:59 PM.

Exercise 1. An energetic salesman works every day of the week. He can work in only one of two towns A and B on each day. For each day he works in town A (or B) his expected reward is r_A (or r_B, respectively). The cost of changing towns is c. Assume that $c > r_A > r_B$, and that there is a discount factor $\alpha < 1$.

(a) Show that for α sufficiently small, the optimal policy is to stay in the town he starts in, and that for α sufficiently close to 1, the optimal policy is to move to town A (if not starting there) and stay in A for all subsequent times.

(b) Solve the problem for $c = 3, r_A = 2, r_B = 1$ and $\alpha = 0.9$ using policy iteration.

(c) Use a computer to solve the problem of part (b) by value iteration.

Exercise 2. Consider a DM making decisions over a finite horizon, with time periods denoted by $t \in \{1, \ldots, T\}$. In each period, the DM makes decisions for N subproblems indexed by $i = 1, \ldots, N$. A state variable $\boldsymbol{s}_t = (s_{t,i})$ summarizes the DM's information about each subproblem in period t. We denote the state space for subproblem i by $S_{t,i}$ and the state space for all subproblems by $\boldsymbol{S}_t = \otimes_{i=1}^{N} S_{t,i}$, where $\otimes$ denotes the Cartesian product. We assume that $S_{t,i}$ are finite and the initial state $\boldsymbol{s}_1$ is given.

In each period t, after observing $\boldsymbol{s}_t$, the DM selects an action $a_{t,i}$ for each subproblem i from the set $A_{t,i}(s_{t,i})$, which we assume to be a finite set. We denote the actions selected for all subproblems in period t by $\boldsymbol{a}_t = (a_{t,1}, \ldots, a_{t,N})$ and denote the set of actions in period t and state $\boldsymbol{s}_t$ by $\boldsymbol{A}_t(\boldsymbol{s}_t) = \otimes_{i=1}^{N} A_{t,i}(s_{t,i})$. We interpret $\ell_{t,i}(s_{t,i}, a_{t,i})$ as the amount of resources consumed by subproblems i in state $s_{t,i}$ when action $a_{t,i}$ selected. In period t, the DM has limited amount of resources shared across all subproblems, which is of the form $\sum_{i=1}^{N} \ell_{t,i}(s_{t,i}, a_{t,i}) \leq b_t$, where $\ell_{t,i}(s_{t,i}, a_{t,i}) \in \mathbb{R}$ and $b_t \in \mathbb{R}$. We assume for all t, i, and states $s_{t,i}$, there exists an action $\hat{a}_{t,i}(s_{t,i}) \in A_{t,i}(s_{t,i})$ – which we denote simply by $\hat{a}_{t,i}$ – such that

$$\ell_{t,i}(s_{t,i}, \hat{a}_{t,i}) \leq \ell_{t,i}(s_{t,i}, a_{t,i}), \ \forall a_{t,i} \in A_{t,i}(s_{t,i}),$$

where the inequality holds component-wise across all L_t constraints.

We denote the DM's set of feasible actions in a given period and state by

$$\mathcal{A}_t(\boldsymbol{s}_t) = \left\{ \boldsymbol{a}_t \in \boldsymbol{A}_t(\boldsymbol{s}_t) : \sum_{i=1}^{N} \ell_{t,i}(s_{t,i}, a_{t,i}) \leq b_t \right\},$$

and assume this set is nonempty in every period and state, i.e., $\sum_{i=1}^{N} \ell_{t,i}(s_{t,i}, \hat{a}_{t,i}) \leq b_t$ for any time t and state $\boldsymbol{s}_t$. Without loss of generality, we assume $\ell_{t,i}(s_{t,i}, a_{t,i}) \geq 0$ for any $(t, i, s_{t,i}, a_{t,i})$.

In each period t, subproblem i generates a reward $r_{t,i}(s_{t,i}, a_{t,i})$ and we denote the total reward in period t by $\boldsymbol{r}_t(\boldsymbol{s}_t, \boldsymbol{a}_t) = \sum_{i=1}^{N} r_{t,i}(s_{t,i}, a_{t,i})$. The states for each subproblem i evolve randomly according to a transition function and a random variable so that $s_{t+1,i} = f_{t,i}(s_{t,i}, a_{t,i}, \tilde{\varepsilon}_{t,i})$, and we assume $(\tilde{\varepsilon}_{t,i})$ are independent across subproblems and time.

The DM's goal is to maximize the total expected rewards.

(a) (DP Formulation) Write down the optimal Bellman equation for the DP problem described above. Let $V_1^*(\boldsymbol{s}_1)$ denote the optimal value.

(b) (Perfect Information Relaxation) Write down the perfect information relaxation bound, i.e., suppose there is a prophet who knows all realization of the randomness $(\varepsilon_{t,i})$ at the beginning of the time horizon. Let $\mathbb{E}_{\tilde{\boldsymbol{\varepsilon}}}\left[V_1^{\mathrm{P}}(\boldsymbol{s}_1; \tilde{\boldsymbol{\varepsilon}})\right]$ denote the corresponding bound.

(c) (Improving Bounds) Improve the perfect information relaxation obtained in (b) using information relaxation and prove that the new bound is indeed better than $\mathbb{E}_{\tilde{\boldsymbol{\varepsilon}}}\left[V_1^{\mathrm{P}}(\boldsymbol{s}_1; \tilde{\boldsymbol{\varepsilon}})\right]$.