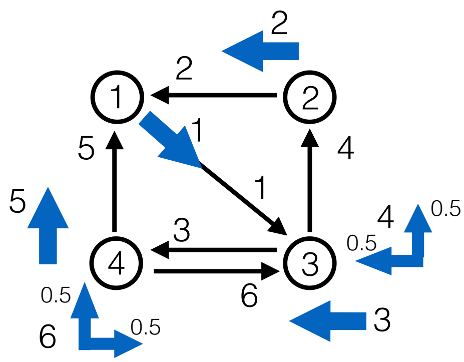
EE578B - Convex Optimization - Winter 2021

Homework 7

Due Date: Wednesday, Mar 3^{rd}, 2021 at 11:59 pm

Consider the Markov Decision Process with the following graph and action structure.



with states $\mathcal{S}$, edges $\mathcal{E}$, and actions $\mathcal{A}$. The actions are given in blue with the associated transition probabilities labeled (when not obvious).

1. **Transition Kernel Constraints**

- **(PTS:0-2)** Write down the incidence matrices for the graph.

$$E_i \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{E}|}, \quad E_o \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{E}|}, \quad P \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \quad A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \quad W \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{A}|}$$

- **(PTS:0-2)** For the incidence matrices given above show the following identities

$$\mathbf{1}^T E_i = \mathbf{1}^T E_o = \mathbf{1}^T$$
$$\mathbf{1}^T A = \mathbf{1}^T P =$$
$$\mathbf{1}^T W = \mathbf{1}^T$$
$$E_i W = P, \quad E_o W = A$$

where the dimension of each $\mathbf{1}$ is determined by context.

- **(PTS:0-2)** Consider two policies with the following actions chosen from each state

Policy 1: State 1: Action 1, State 2: Action 2,
State 3: Action 4, State 4: Action 6

Policy 2: State 1: Action 1, State 2: Action 2,
State 3: $\begin{matrix}\text{50\% Action 3}\\ \text{50\% Action 4}\end{matrix}$, State 4: $\begin{matrix}\text{50\% Action 5}\\ \text{50\% Action 6}\end{matrix}$

Write each policy in matrix form $\Pi \in \mathbb{R}^{6\times 4}$. Compute the corresponding Markov matrix $M = P\Pi$. Also show that $A\Pi = I$ for each policy.

- **(PTS:0-4)** The stationary (state) distribution associated with each Markov chain is the solution to the equation $\rho = M\rho$. Compute this stationary distribution by finding the eigenvector with eigenvalue 1. (You can use the function `eig` in Matlab or `numpy.linalg.eig` in Python.). Make sure to scale the eigenvector so that it is an appropriate probability distribution that sums to 1 and has all positive values. Compute the corresponding action distribution y as $y = \Pi\rho$.
- **(PTS:0-2)** Show that each y from the previous part satisfies $Py = Ay$ and $\mathbf{1}^T y = 1$. Compute the corresponding edge mass vector for each $x = Wy$. Show that x is in the nullspace of $E = E_i - E_o$.

2. **Infinite Horizon, Average Reward**

Consider the following optimization problem for finding the optimal steady-state action distribution $y \in \mathbb{R}^{|\mathcal{A}|}$

$$\begin{aligned} \max_{y} \quad & r^T y \\ \text{s.t.} \quad & Py = Ay,\ \mathbf{1}^T y = 1,\ y \geq 0 \end{aligned} \tag{1}$$

for reward vector $r \in \mathbb{R}^{|\mathcal{A}|}$.

- **(PTS:0-2)** Write the dual optimization problem with dual variables $\lambda \in \mathbb{R}$ associated with the constraint $\mathbf{1}^T y = 1$, $v \in \mathbb{R}^{|\mathcal{S}|}$ associated with constraint $Py = Ay$, $\mu \in \mathbb{R}_+^{|\mathcal{A}|}$ associated with the constraint $y \geq 0$.
- **(PTS:0-2)** The KKT conditions at optimum (for either the primal or dual problem) are given by

$$\begin{aligned} r^T - \lambda \mathbf{1}^T + v^T(P - A) + \mu^T = 0, \quad & \mu \geq 0 \\ Py - Ay = 0, \quad \mathbf{1}^T y = 1, \quad & y \geq 0 \\ \mu^T y = 0 & \end{aligned}$$

  Use these conditions to show that λ is an upper bound on the primal objective $r^T y$ for any feasible y. What does $\mu^T y$ represent for a specific y? What does the condition $\mu^T y = 0$ imply about the optimal y?
- **(PTS:0-4)** Use `cvx` or `cvxpy` to solve the above optimization problem for the transition kernel given initially and each reward vector

$$r^T = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}$$
$$r^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

  What is the optimal joint distribution y in each case? What is the expected average reward $r^T y$ in each case?
- **(PTS:0-2)** What is the steady-state state distribution associated with each solution $\rho = Ay$? What is the optimal policy associated with y? Use the formula

$$(\pi_s)_a = \frac{y_a}{\rho_s} = \frac{y_a}{\sum_{a \in \mathcal{A}_s} y_a}$$

  You could also put the policy in matrix form using the formula

$$\Pi = \text{diag}(y) A^T \text{diag}(\rho)^{-1}$$

- **(PTS:0-2)** Now suppose you apply the policy

$$\Pi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}$$

What reward do you achieve in each case? (Hint: compute ρ such that $\rho = P\Pi\rho$ and then y using $y = \Pi\rho$.) How much does this reward differ from the optimal average reward? How does this difference relate to the quantity $\mu^T y$ where μ is the optimal dual variable?

3. **Finite Horizon, Total Reward**

Consider the following optimization problem for finding the optimal finite horizon policy.

$$\begin{aligned} \max_{y(t),\ t \in \mathcal{T}} \quad & \sum_{t=0}^{T-1} r(t)^T y(t) + g^T A y(T) \\ \text{s.t.} \quad & Ay(0) = \rho(0), \quad y(0) \geq 0 \\ & Ay(t+1) = Py(t), \quad y(t+1) \geq 0, \quad t \in \mathcal{T} \end{aligned} \tag{2}$$

where $\mathcal{T} = \{0, \ldots, T-1\}$, $\rho(0) \in \mathbb{R}^{|\mathcal{S}|}$ is a given initial state distribution, and $g \in \mathbb{R}^{|\mathcal{S}|}$ is a final cost on each of the states.

- **(PTS:0-4)** Write the dual optimization problem with dual variables $v(0) \in \mathbb{R}^{|\mathcal{S}|}$ associated with the constraint $Ay(0) = \rho(0)$, $v(t+1) \in \mathbb{R}^{|\mathcal{S}|}$ associated with constraint $Py(t) = Ay(t+1)$, and $\mu(t) \in \mathbb{R}_+^{|\mathcal{A}|}$ associated with the constraint $y(t) \geq 0$.
- **(PTS:0-4)** The KKT optimality conditions for the primal and dual optimization problems are given by

$$\begin{aligned} g^T A - v(T)A + \mu(T)^T &= 0, \quad \mu(T) \geq 0 \\ r(t)^T + v(t+1)^T P - v(t)^T A + \mu(t)^T &= 0, \quad \mu(t) \geq 0, \quad t \in \mathcal{T} \\ Ay(0) &= \rho(0), \quad y(0) \geq 0 \\ Ay(t+1) &= Py(t), \quad y(t+1) \geq 0, \quad t \in \mathcal{T} \\ \mu(t)^T y(t) &= 0, \quad t \in \mathcal{T},\ t = T \end{aligned}$$

Use these conditions to show that $v(0)^T \rho(0)$ is an upper bound on the primal objective $\sum_t r(t)^T y(t) + g^T A y(T)$ for any feasible $y(t)$ that satisfies the mass flow equations. What does $\sum_t \mu(t)^T y(t)$ represent for a specific mass flow $y(t), t \in \mathcal{T}$.
- **(PTS:0-4)** Use `cvx` or `cvxpy` to solve the above optimization problem for the MDP given initially with the following rewards

$$r(t)^T = \begin{bmatrix} 2 & 1 & 2 & 1 & 2 & 1 \end{bmatrix} \text{ for } t \in \mathcal{T}, \quad g^T = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

for ten time steps $T = 10$ and initial distribution $\rho(0) = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}^T$

What is the optimal action distribution $y(t)$ at each time step? What is the expected total reward $\sum_t r(t)^T y(t)$?

- **(PTS:0-4)** What is the policy $\Pi(t)$ chosen at each time step? Use the formula

$$(\pi_s)_a(t) = \frac{y_a(t)}{\rho_s(t)} = \frac{y_a(t}{\sum_{a\in\mathcal{A}_s} y_a(t)}$$

  where $\rho(t) = Ay(t)$.

- **(PTS:0-4)** Now suppose you apply the policy

$$\Pi(t) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}$$

  at each time step. Start by computing $y(0) = \Pi(0)\rho(0)$. $\rho(t)$ is then given by $Py(0) = \rho(1)$. Use $\rho(1)$ to compute $y(1) = \Pi(1)\rho(1)$, etc. What total reward do you achieve? What is the quantity $\sum_t \mu(t)^T y(t)$? How does this relate the total reward to the optimal total reward?