Problem Set 1

Discrete optimization problems

Optimal control and reinforcement learning, TU/e, 2022-2023

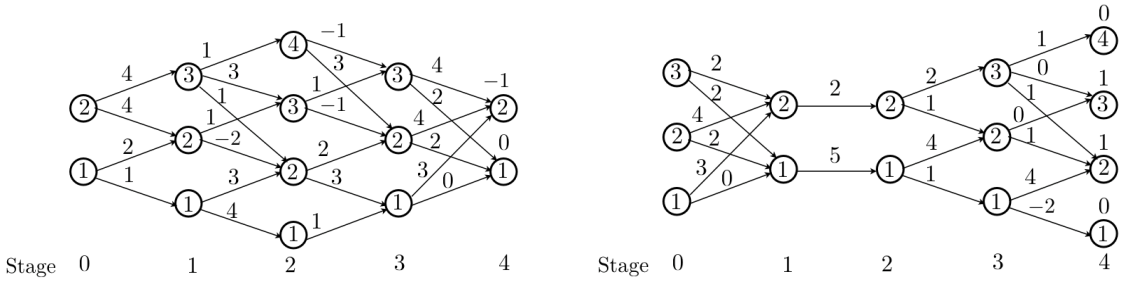
Outline

1. Modeling and the dynamic programming algorithm
2. Stochastic dynamic programming
3. Search methods in graphs
4. Bayes' filter and POMDP

---

1. Modeling and the dynamic programming algorithm

Problem 1.1 Answer the following questions for both transition diagrams (a) and (b).



(a) (b)

(i) For each initial state at stage 0 compute one optimal path minimizing the cost incurred from stage 0 to stage 4, using the dynamic programming algorithm.

(ii) Indicate one optimal policy and the costs-to-go in the transition diagram.

(iii) Are the optimal polices and the optimal paths for each initial state unique? If not, compute the number of optimal paths for each initial state and the number of optimal polices. Are the costs-to-go different for different optimal policies?

Problem 1.2 Consider an inventory control problem where the number of items of a given product along $h = 3$ stages is described by

$$x_{k+1} = \max\{x_k + u_k - d_k, 0\}, \quad k \in \{0, \ldots, h-1\},$$

where $x_k \in \{0, \dots, N\}$, $u_k \in \{0, \dots, N - x_k\}$, d_k denote the number of items, the supply and the demand at time k, respectively, and $N = 3$ is the capacity. The objective is to minimize the cost

$$\sum_{k=0}^{h-1} \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big) + g_h(x_h), \tag{1}$$

where $c_1(i) = 0.3i$, $i \in \{0, \dots, N\}$, is the storage cost,

$$c_2(j) = \begin{cases} 2.5j + 0.5 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 4$ is the selling price per item, and $g_h(i) = -2i$, $i \in \{0, \dots, N\}$, is the terminal cost. Suppose that $d_0 = 1$, $d_1 = 1$, $d_2 = 1$. Sketch the transition diagram for the problem and compute the possible optimal policies and the costs-to-go.

Problem 1.3 The following equation describes the evolution of the number of items of a given product in a shop over h stages

$$x_{k+1} = x_k + u_k - d_k, \quad k \in \{0, \dots, h-1\},$$

where x_k, u_k, d_k denote the number of items, the supply and the demand at time k. Negative stocks indicate backlogged excess demands which are filled in as soon as additional inventory becomes available. A maximum number of item N can be stored, i.e., $x_k \leq N$ which imposes the following condition on the supplies $u_k \in \{0, 1, \dots, N - x_k\}$. At each time k a cost $c_1(x_k)$ is incurred for either positive stock or negative stock and ordering u_k items has a cost $c_2(u_k) = cu_k + c_{\text{tr}}\|u_k\|_0$, where c_{tr} is the transportation price, $\|u_k\|_0$ equals zero if u_k is zero and equals one otherwise, and c is the cost per item. There is a terminal cost at stage h taking the form

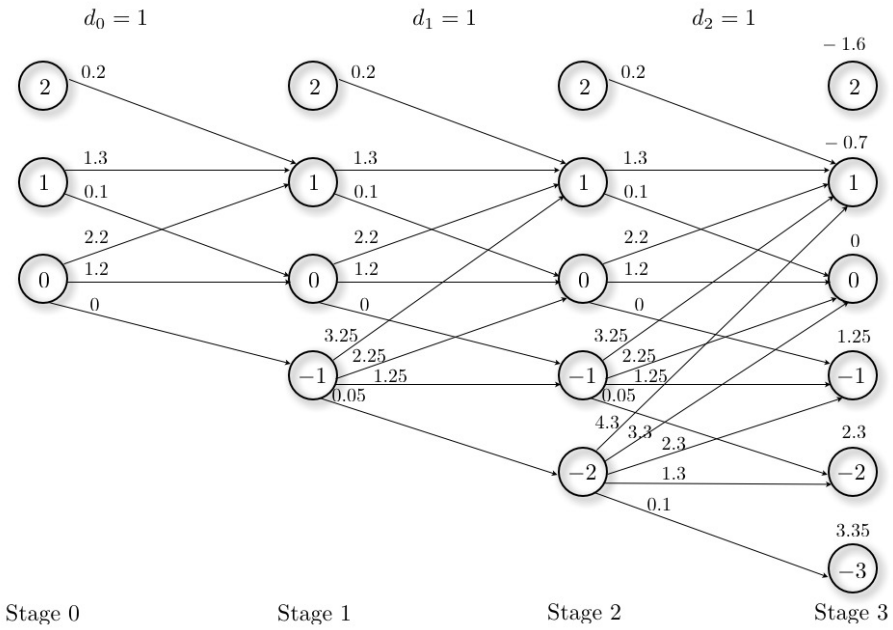
$$g_h(x_h) = \begin{cases} -cx_h + c_1(x_h) + c_{tr}, \text{ if } x_h \neq 0, \\ 0, \text{ if } x_h = 0, \end{cases}$$

which results from assuming that items in excess $x_h > 0$ are returned whereas unmet demand $x_k < 0$ is exactly compensated for by a final supply ($u_h = -x_h$ with cost cu_h). In both cases transportation and storage prices must be paid. The total cost to be minimized is then given by [1]

$$\sum_{k=0}^{h-1} \big(c_1(x_k) + c_2(u_k)\big) + g_h(x_h). \tag{2}$$

(i) Suppose that $h = 3$, $d_0 = 1$, $d_1 = 1$, $d_2 = 1$, $N = 2$, $c = 1$, $c_{\text{tr}} = 0.2$, and $c_1(i) = 0.1i$ if $i \geq 0$ and $c_1(i) = -0.05i$ if $i < 0$. Assuming that the initial inventory is positive, one can obtain the transition diagram depicted in the figure below, where each discrete state denotes the possible value for the number of items at each stage and arrows denote possible supply options ($u_k \in \{0, \dots, N - x_k\}$, associated with cost $(c_1(x_k) + c_2(u_k))$.The terminal costs are also indicated. Compute the optimal policy for this inventory control problem. Provide the optimal supplies u_0, u_1, u_2 for each of the possible initial states $x_0 \in \{0, 1, 2\}$. For each initial state indicate if such optimal supplies are unique.

[1] Note that this model is different from the one discussed in class. For example here negative stocks are allowed. Moreover, since all the demand is met, possibly by ordering more items at the final stage which is accounted for by the final cost, the cost function does not need to take into account the profit made by selling items.

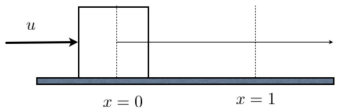


(ii) Suppose now that $c_{\text{tr}} = 0.01$ instead of $c_{\text{tr}} = 0.2$. Recompute the costs for each transition and the final cost and compute also the optimal policy. Is it the same as the one obtained considering $c_{\text{tr}} = 0.2$?

(iii) Consider again that $c_{\text{tr}} = 0.2$ and suppose now that d_1 is uncertain and that $\text{Prob}[d_1 = 1] = 0.6$, $\text{Prob}[d_1 = 0] = 0.4$. For each of the optimal paths obtained in (i) for the initial state $x_0 = 0$ (possibly more than one), with supplies $\bar{u}_0$, $\bar{u}_1$, $\bar{u}_2$ answer the following questions. If we apply an open loop policy, i.e., apply these supplies $\bar{u}_0$, $\bar{u}_1$, $\bar{u}_2$, what is the expected cost? What is the expected cost if we use the policy obtained in (i) to readjust the supply at stage 2 (if needed)? Compare the costs of open loop and closed loop strategies.

Problem 1.4 Formulate each of the following problems as either a discrete optimization problem, a stage-decision problem or a continuous-time optimal control problem:

a. We wish to move a mass on a plane surface from $x = 0$ to $x = 1$ in minimum time, using a force $F = u$. A model can be derived using Newton's law. Assume that the friction force is proportional to the velocity.



b. Redo a. assuming that the mass is controlled digitally.

c. Consider the problem of planing a project consisting of several activities, as depicted in the diagram of Figure 1. In the diagram, circles represent stages of the project and arrows represent activities. The duration of each activity in weeks is also indicated. Note that some activities can only start after other activities are concluded. The goal is to find the overall duration of the project as well as the critical activities, i.e., those for which the project will be delayed if the activity is delayed.

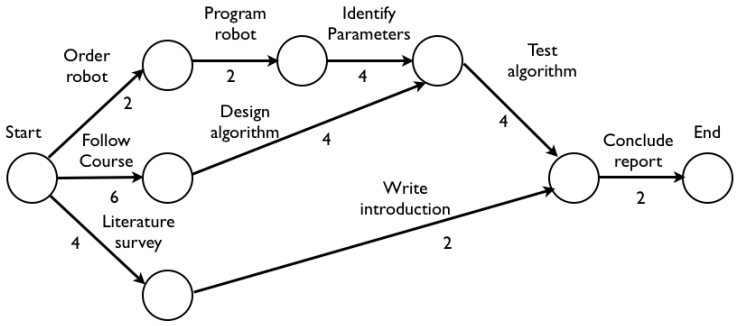
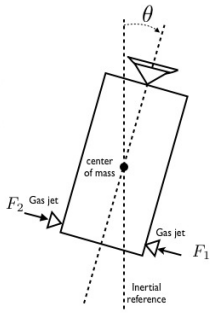


Figure 1: Example of a graph of an activity network

d. The attitude of a satellite must be stabilized and controlled so that the satellite's antenna is pointed to a particular point on Earth for communication. Consider the on board control of one of the rotational degrees of freedom by thrusters as indicated in Figure 2. The goal is to choose the torque T to move the satellite angle θ from an initial angle at time t_0, $\theta(t_0)$, to a desired angle θ_{des} at time $t_0 + h$, while minimizing the fuel consumption $c(T)$, which is a function of the applied thrust T.

(a) Source: wikipedia.



(b) Satellite control

Figure 2: Control of a satellite

2. Stochastic dynamic programming

Problem 2.1 Consider the following two-stage stochastic dynamic programming problems

$$\text{Problem 1:} \quad \begin{cases} \min \sum_{k=0}^{2} |x_k| \\ x_{k+1} = x_k + u_k + d_k \end{cases} \qquad \text{Problem 2:} \quad \begin{cases} \min \sum_{k=0}^{2} -x_k \\ x_{k+1} = x_k + u_k + d_k \end{cases}$$

where $u_k \in \{-1, 0, 1\}$, $\text{Prob}[d_k = -1] = \text{Prob}[d_k = 0] = \text{Prob}[d_k = 1] = \frac{1}{3}$, for $k \in \{0, 1\}$.

(i) Given that $x_0 \in X_0$, $X_0 = \{0\}$, find the sets X_1, X_2 of possible values for x_1 and x_2, respectively, for both Problems 1 and 2.

(ii) Find the optimal policies for Problems 1 and 2, using the (stochastic) dynamic programming algorithm.

(iii) What is the expected optimal cost for Problems 1 and 2?

Problem 2.2 Consider Problem 1.3 and assume as in question (iii) that $c_{\text{tr}} = 0.2$ and $\text{Prob}[d_1 = 1] = 0.6$, $\text{Prob}[d_1 = 0] = 0.4$.

(i) Use the (stochastic) dynamic programming algorithm to compute an optimal policy which minimizes the expected cost.

(ii) What is the expected cost obtained with the stochastic dynamic programming policy for an initial condition $x_0 = 0$? Is it better than the open-loop and closed loop options computed in Problem 1.3?

Problem 2.3 Consider an inventory control problem where the number of items of a given product along $h = 3$ stages is described by

$$x_{k+1} = \max\{x_k + u_k - d_k, 0\}, \quad k \in \{0, \dots, h-1\},$$

where $x_k \in \{0, \dots, N\}$, $u_k \in \{0, \dots, N - x_k\}$, d_k denote the number of items, the supply and the demand at time k, respectively, and $N = 3$ is the capacity. The objective is to minimize the cost

$$\sum_{k=0}^{h-1} \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big) + g_h(x_h), \tag{3}$$

where $c_1(i) = 0.5i$, $i \in \{0, \dots, N\}$, is the storage cost,

$$c_2(j) = \begin{cases} 2j + 0.8 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 5$ is the selling price per item, and $g_h(i) = -1.8i$, $i \in \{0, \dots, N\}$, is the terminal cost.

(i) Suppose that $d_0 = 1$, $d_1 = 1$, $d_2 = 1$. Sketch the transition diagram for the problem and compute the possible optimal policies. Confirm that for $x_0 = 2$, $u_0 = 0$, $u_1 = 0$, $u_2 = 1$ are optimal supplies.

(ii) Suppose now that the demand d_1 is uncertain and characterized by $\text{Prob}[d_1 = 0] = 0.4$, $\text{Prob}[d_1 = 1] = 0.2$, $\text{Prob}[d_1 = 2] = 0.4$, while, again, $d_0 = 1$, $d_2 = 1$. Consider that $x_0 = 2$.

(a) Suppose that the supplies $u_0 = 0$, $u_1 = 0$, $u_2 = 1$ corresponding to the optimal supplies in (i) are still applied at stages $k = 0$, $k = 1$, $k = 2$, respectively, independently of the value of the disturbance. What is the expected value of the cost?

(b) Suppose that we use the *policy* computed in (i) to compute the decision u_2 which therefore may now depend on d_1. What is now the expected value of the cost?

(c) Using the (stochastic) dynamic programming algorithm, compute the optimal policy minimizing the expected value of the cost and provide the corresponding expected value of the cost.

(d) Discuss and compare the expected value of the costs obtained in (ii).a, (ii).b, and (ii).c.

Problem 2.4 A lady wants to buy a car, which is only available at three car dealers (1, 2, and 3). She goes first to dealer 1 and receives an offer. She must then decide either to accept the offer and buy the car or refuse the offer and go to the next dealer 2, never returning to dealer 1. In the latter case, she must decide again either to accept the offer of dealer 2 or go to dealer 3, never returning to dealer 2. If she goes to dealer 3, she must accept the offer of dealer 3 (although dealer 3 does not know this). Based on a study of previous deals she estimates that the dealers $k \in \{1, 2, 3\}$ make their offers independently and their prices a_k follow the same probability distribution described by

$$\text{Prob}[a_k = 8] = \frac{1}{3}, \quad \text{Prob}[a_k = 9] = \frac{1}{12}, \quad \text{Prob}[a_k = 10] = \frac{1}{6}, \quad \text{Prob}[a_k = 11] = \frac{1}{12}, \quad \text{Prob}[a_k = 12] = \frac{1}{3}$$

(i) Determine the optimal decisions, to buy or not to buy, after receiving the offer of dealer 1 and after receiving the offer of dealer 2. [Suggestion: formulate the problem as a discrete optimization problem and use the dynamic programming algorithm].

(ii) Compute the expected cost of the car before going to dealer 1. Compare this with the expected cost of simply accepting whichever offer dealer 1 makes.

Problem 2.5 [1, Ex. 4.19, p. 211] A driver is looking for parking on the way to his destination. Each parking place if free with probability p independently of whether other parking places are free or not. The driver cannot observe whether a parking place is free until he reaches it. If he parks k places from his destination, he incurs a cost k. If he reaches the destination without having parked the cost is C.

(a) Let F_k be the minimal expected cost if he is k parking places from his destination, where $F_0 = C$. Show that

$$F_k = p \min(k, F_{k-1}) + qF_{k-1}, \quad k = 1, 2, \ldots,$$

where $q = 1 - p$.

(b) Show that an optimal policy is of the form: never park if $k \geq k^*$, but take the first free place if $k < k^*$, where k is the number of parking places from the destination and k^* is the smallest integer i satisfying $q^{i-1} < (pC + q)^{-1}$.

Problem 2.6 A gentleman wants to sell his house in the next three months, June, July or August, labeled by 1, 2, 3, respectively. After a market study he estimates that the best offers he will get for each month, denoted by w_1 (June), w_2 (July), w_3 (August) are distributed according to the values shown in the table

Prob$[w_j = P]$	$P = 100$	$P = 125$	$P = 150$	$P = 175$	$P = 200$
$j = 1$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
$j = 2$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{3}{10}$
$j = 3$	0	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{3}{10}$	$\frac{3}{10}$

The offers are independent and are only available for that particular month. If the gentleman has not sold the house by August, he has to take whichever best offer he receives in August. Let $u_1 \in \{0, 1\}$ and $u_2 \in \{0, 1\}$ indicate the decision to sell ($u_j = 1$) or not ($u_j = 0$) the house on months $j = 1$ (June) and $j = 2$ (July) given the offers w_1 and w_2, respectively. Provide the optimal policy u_1 and u_2 (i.e., the one that maximizes the expected selling price for the house).

Problem 2.7 [1, Adapted from Ex. 1.4, p. 52] This assignment considers a two-player game inspired by blackjack and is a variation of the game described in [1, Ex. 1.4, p. 52]. Both players A and B start by throwing a dice. Based on both outcomes, player A may decide to stop or may decide to throw the dice again and add the result to the result of his/her previous throw. In the latter case, player A may again decide to stop or throw the dice again and add the result of the new throw to the sum of his previous throws, and so on. However, if at a given stage this sum exceeds seven, player A loses the game (busts). If player A stops before exceeding seven, then player B takes over and throws the dice successively until either his/her sum equals or exceeds the score of player A, in which case player B wins, or the sum of player B is over seven, in which case player B loses the game. Using (stochastic) dynamic programming, determine a stopping strategy for player A that maximizes the probability of winning, assuming that the initial throw of player B is two. Provide the probabilities of player A winning as a function on the initial throw. [Suggestion: follow a similar hint to the one provided in [1, Ex. 1.4, p. 52]. Note, however, that in the context of [1, Ex. 1.4, p. 52], Player B would stop when it would achieve a sum of 4 or higher.]

3. Search methods in graphs

Problem 3.1 Consider a weighted graph characterized by the following matrix

$$W = \begin{bmatrix} 0 & 0 & \infty & \infty & 1 \\ 3 & 0 & 5 & 3 & 1 \\ \infty & 5 & 0 & 1 & 3 \\ \infty & \infty & 1 & 0 & 1 \\ 1 & 1 & 3 & 1 & 0 \end{bmatrix}$$

where each component W_{ij} of row i and column j denotes the cost of the link from node i to node j ($W_{ij} = \infty$ indicates that there is no link from node i to node j).

i) Compute a shortest path from node 1 to node 3 using the dynamic programming algorithm.

i) Compute a shortest path from node 1 to node 3 using the Dijkstra's algorithm.

iii) For each node $i \in \{1, \ldots, 5\}$, indicate the first node (immediately after i) of the optimal path from node i to node 3 with only 3 hops.

Problem 3.2 Consider the transition diagram depicted in Figure 3. For each initial state at stage 0 compute one optimal path minimizing the cost incurred from stage 0 to stage 4, using the Dijkstra's algorithm.

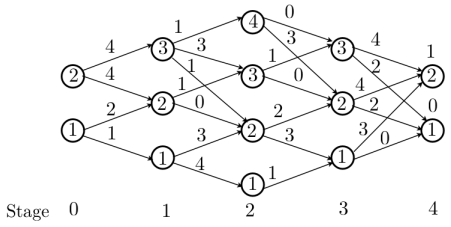


Figure 3: Transition diagram

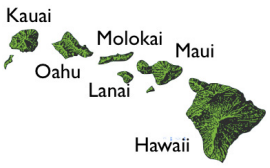
Problem 3.3 A cost-minded traveler visiting Hawaii wants to fly from Kauai to Maui and checks the prices of the flights for the next day, summarized in the table below. A direct flight costs 300 money units, and the traveler believes that there should be a cheaper non-direct flight. He estimates that he can do trips with at most 4 hops and he is interested in knowing what is the best price with at most 2, 3 and 4 hops to compare the options.

(i) Provide the answers that the traveler is looking for using the dynamic programming algorithm.

(ii) Would there be a better option if 5 hops were allowed?

(iii) Solve (i) with the Dijkstra's algorithm.

Arrival \Departure	Oahu	Hawaii	Kauai	Lanai	Maui	Molokai
Oahu	-	90	70	105	130	95
Hawaii	95	-	80	55	100	115
Kauai	75	105	-	135	250	85
Lanai	105	35	120	-	40	100
Maui	130	95	300	45	-	140
Molokai	90	120	80	95	95	-



Problem 3.4 Consider the weighted graph depicted in Figure 4. Compute the shortest path from node 1 to node 11 using:

i) the Dijkstra's algorithm.

ii) the A* algorithm with the following heuristic h

node i	1	2	3	4	5	6	7	8	9	10	11
$h(i)$	0	25	50	15	16	50	9	7	50	0	0

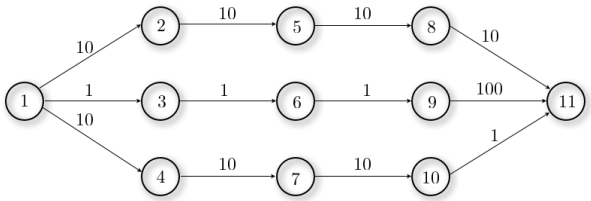
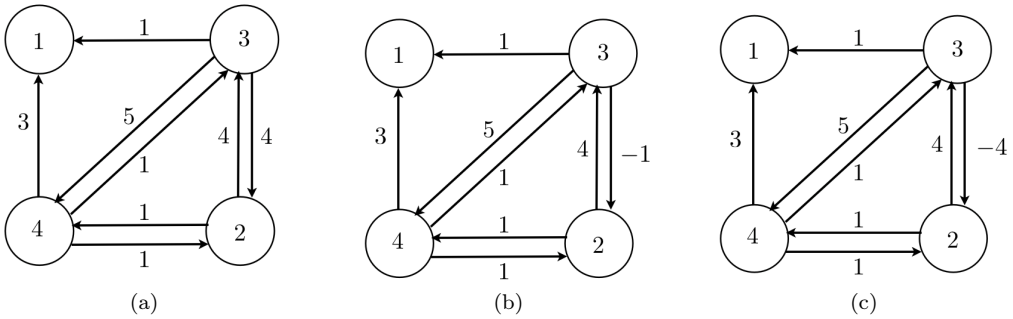


Figure 4: Directed graph

Problem 3.5 Although we have only considered so far graphs with non-negative weights, it is possible to apply the dynamic programming algorithm to find the shortest path in graphs with negative weights, under a certain condition. The aim of this exercise is to find such a condition. Consider the following three graphs.



(a) (b) (c)

(i) Apply the dynamic programming algorithm to obtain the shortest path from the initial node 2 to the final node 1 for graphs (a), (b) and (c).

(ii) Compare the shortest path that you have found for graph (c) with the cost of the following path $2 \to 4 \to 3 \to 2 \to 4 \to 3 \to 1$? Is the path found in (i) optimal?

(iii) Can you find a path with a smaller cost for graph (b) than the one found in (i)?

(iii) Can you find a condition under which the dynamic programming algorithm finds the shortest path for graphs with negative weights?

(iv) Would it be possible to apply the Dijkstra's algorithm?

4. Bayes' filter and POMDP

Problem 4.1 Suppose that we wish to estimate the angle of a spinning wheel with a camera. Different colors with continuously varying values of hue according to the HSV color representation are placed on the wheel. Due to space limitations, the camera is placed close to the wheel and can only detect a limited range of the wheel. The angle is quantized into N values and this induces a quantization also for the value of hue of the camera measurements. Due to changes in brightness and other exterior conditions, the camera processing algorithm can provide wrong estimates of the hue, but these are typically close to the correct value. The setting is depicted in Figure 5. In this exercise, for simplicity, we assume $N = 4$, considering four angles and four associated colors : (i) green, $\theta = 0$; (ii) orange, $\theta = \pi/2$; (iii) red, $\theta = \pi$; (iv) blue $\theta = 3\pi/2$. The wheel is assumed to be spinning anti clock-wise according to the following model for the angle θ_k at time $k \in \mathbb{N} \cup \{0\}$,

$$\theta_{k+1} = \theta_k + w_k$$

where

$$\text{Prob}[\theta_0 = \beta] = \begin{cases} 0.25 \text{ if } \beta = 0 \\ 0.25 \text{ if } \beta = \dfrac{\pi}{2} \\ 0.25 \text{ if } \beta = \pi \\ 0.25 \text{ if } \beta = \dfrac{3\pi}{2} \end{cases} \qquad \text{Prob}[w_k = \alpha] = \begin{cases} 0.2 \text{ if } \alpha = 0 \\ 0.6 \text{ if } \alpha = \dfrac{\pi}{2} \\ 0.2 \text{ if } \alpha = \pi \end{cases} \quad \forall k$$

The measurement at time k obtained from the algorithm processing camera data is denoted by $y_k \in \{\text{green}, \text{orange}, \text{red}, \text{blue}\}$ and follows the model

$$\text{Prob}[y_k = \text{c}|\theta_k = 0] = \begin{cases} 0.1 \text{ if c} = \text{blue} \\ 0.8 \text{ if c} = \text{green} \\ 0.1 \text{ if c} = \text{orange} \end{cases} \quad \text{Prob}[y_k = \text{c}|\theta_k = \frac{\pi}{2}] = \begin{cases} 0.1 \text{ if c} = \text{green} \\ 0.8 \text{ if c} = \text{orange} \\ 0.1 \text{ if c} = \text{red} \end{cases}$$

$$\text{Prob}[y_k = \text{c}|\theta_k = \pi] = \begin{cases} 0.1 \text{ if c} = \text{orange} \\ 0.8 \text{ if c} = \text{red} \\ 0.1 \text{ if c} = \text{blue} \end{cases} \quad \text{Prob}[y_k = \text{c}|\theta_k = \frac{3\pi}{2}] = \begin{cases} 0.1 \text{ if c} = \text{red} \\ 0.8 \text{ if c} = \text{blue} \\ 0.1 \text{ if c} = \text{green} \end{cases} \quad \forall k$$

Compute, for $k \in \{0, 1, 2, 3, 4\}$, and for $c \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$,

$$\text{Prob}[\theta_k = c|I_k]$$

where $I_k = \{y_0, \ldots, y_k\}$ and the measurements are $y_0 = \text{green}$, $y_1 = \text{orange}$, $y_2 = \text{orange}$ $y_3 = \text{blue}$, $y_4 = \text{blue}$.

Problem 4.2 The goal of this exercise is to repeat Problem 4.1 but assuming that the sensor can only detect one of the angles without noise, as suggested in Figure 6. The sensor model is then

$$\text{Prob}[y_k = 1|\theta_k = 0] = 1 \quad \text{Prob}[y_k = 0|\theta_k = 0] = 0$$

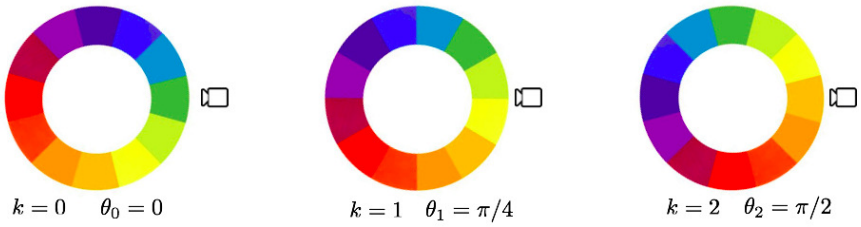


Figure 5: Problem setting 4.1

and, for $\beta \in \{\frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$,

$$\text{Prob}[y_k = 1|\theta_k = \beta] = 0 \quad \text{Prob}[y_k = 0|\theta_k = \beta] = 1.$$

Compute, for $k \in \{0, 1, 2, 3, 4\}$, and for $c \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$,

$$\text{Prob}[\theta_k = c|I_k]$$

where $I_k = \{y_0, \ldots, y_k\}$ and the measurements are $y_0 = 1$, $y_1 = 0$, $y_2 = 0$ $y_3 = 1$, $y_4 = 0$.

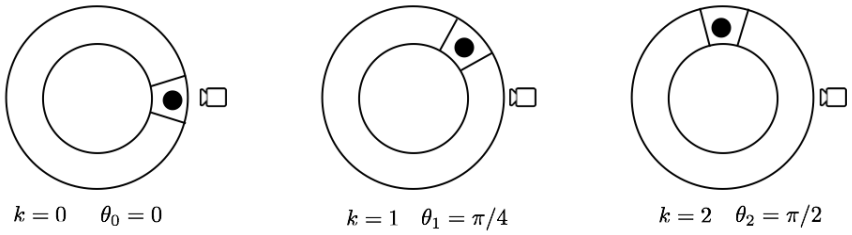


Figure 6: Problem setting 4.2

Problem 4.3 [1, Ex. 5.6] Consider a machine that can be in one of two states, good or bad. Suppose that the machine produces an item at the end of each period. The time produced is either good or bad depending on whether the machine is in a good or bad state at the beginning of the corresponding period, respectively. We suppose that once the machine is in a bad state it remains in that state until it is replaced. If the machine is in a good state at the beginning of a certain period, then with probability t it will be in the bad state at the end of the period. Once an item is produced, we may inspect the item at a cost I or not inspect. If an inspected item is found to be bad, the machine is replaced with a machine in good state at cost R. The cost of producing a bad item is $C > 0$. Write a DP algorithm for obtaining an optimal inspection policy assuming a machine initially in good state and a orizon of N periods. Solve the problem for $t = 0.2$, $I = 1$, $R = 3$, $C = 2$ and $N = 8$. (the optimal policy is to inspect at the end of the thir period and not inspect in any other period).

Problem 4.4 [1, Ex. 5.10] A person is offered 2 to 1 odds in a coin-tossing game where he wins whenever a tail occurs. However, he suspects that the coin is biased and has an a priori probability

distribution $F(p)$ for the probability p that a head occurs at each toss. The problem is to find an optimal policy of deciding whether to continue or stop participating in the game given the outcomes of the game so far. A maximum of N tossings is allowed. Indicate how such a policy an be found by means of DP.

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control.* Athena Scientific, 2005.