Problem set 3 - AI for Engineering Systems track

Introduction to Reinforcement Learning

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

Outline

Model-based stochastic rollout, Adaptive sampling, Sequential dynamic programming

Infinite-horizon problems, Policy iteration

Q-Learning

Model-based stochastic rollout, Adaptive sampling, Sequential dynamic programming

Problem 1.1 Consider the following switched linear system subject to stochastic disturbances

$$x_{k+1} = A_{\sigma_k} x_k + w_k$$

in the interval $k \in \{0,1,2\}$ with $\sigma_k \in \{1,2\}$ and $\{w_k | k \in \{0,1,2\}\}$ are zero-mean independent and identically distributed random variables with

$$\mathbb{E}[w_k w_k^\intercal] = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

and

$$A_1 = \begin{bmatrix} 0.9 & 0.1 \\ 0 & 0.8 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.8 & 0.1 \\ 0 & 0.9 \end{bmatrix}$$

Suppose that the following cost should be minimized

$$\mathbb{E}[\sum_{k=0}^{2} x_k^\intercal x_k + x_3^\intercal x_3]$$

(i) Provide a stochastic rollout policy for σ_k with horizon $H = 1$, $k \in \{0,1,2\}$, with base policy $\sigma_k = 1$ for every $k \in \mathbb{N}_0$.

(ii) Provide the sequence resulting from the policy in (i) if the disturbances are

$$w_0 = \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} \quad w_1 = \begin{bmatrix} 0.2 \\ 0 \end{bmatrix} \quad w_2 = \begin{bmatrix} -0.05 \\ 0.1 \end{bmatrix}$$

and the initial condition is $x_0 = [0.2 \quad 1]^\intercal$.

Problem 1.2 [1] Consider the following switched linear system subject to stochastic disturbances

$$x_{k+1} = A_{\sigma_k} x_k + w_k$$

in the interval $k \in \mathbb{N}_0$ with $\sigma_k \in \{1,2\}$ and $\{w_k | k \in \mathbb{N}_0\}$ is a sequence of zero-mean independent and identically distributed random variables with

$$\mathbb{E}[w_k w_k^\intercal] = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$$

[1] This question needs results from discounted cost problems that we have not covered and it is therefore an advanced question

and

$$A_1 = \begin{bmatrix} 1.1 & 0 \\ -0.2 & 0.5 \end{bmatrix} \quad A_2 = \begin{bmatrix} -0.1 & 0 \\ -0.1 & -0.9 \end{bmatrix}$$

Suppose that the following cost should be minimized

$$\sum_{k=0}^{\infty} \mathbb{E}[\alpha^k x_k^\intercal Q x_k], \quad \alpha = 0.9, \quad Q = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$$

Provide a stochastic rollout policy for σ_k, $k \in \mathbb{N}_0$ with horizon $H = 2$, with base policy $\sigma_k = 2$ for every $k \in \mathbb{N}_0$.

Problem 1.3 Suppose that in the context of a model free stochastic rollout method for switched linear systems, one wishes to evaluate through simulations four switching sequences, labelled 1,2,3,4. Let a_n be the switching sequence evaluated at iteration n using adaptive sampling; assume that $a_1 = 1$, $a_2 = 2$, $a_3 = 3$, $a_4 = 4$, and a_n, for $n \geq 5$, is picked according to the rule

$$a_{n+1} = \arg \min_{j \in \{1,2,3,4\}} \frac{\sum_{s=1}^{n} \bar{J}^s(a_s)\, 1_{[a_s = j]}}{\sum_{s=1}^{n} 1_{[a_s=j]}} - c\sqrt{\frac{\log(n)}{\sum_{s=1}^{n} 1_{[a_s=j]}}}$$

with $c = 1$, where $\bar{J}^s(a_s)$ is the cost obtained by simulating the switching sequence a_s. Suppose that these costs and decisions up to iteration $n = 10$ are the ones summarized in the table below. Determine the label of the switching sequence $a_{11} \in \{1, 2, 3, 4\}$ which should be evaluated at iteration $n = 11$.

n	1	2	3	4	5	6	7	8	9	10
a_n	1	2	3	4	4	3	3	2	2	4
$\bar{J}^n(a_n)$	1.1	0.9	0.5	0.2	1	0.33	0.4	0.25	0.4	0.8

Problem 1.4 Repeat Problem 1.1 with $c = 0$ and then $c = 2$ and reflect on the new choices of optimal decision.

Problem 1.5 Suppose that in the context of sequential dynamic programming, the algorithm starts at the last decision stage with a set of representative states $\{x_{h-1}^s | s \in \{1, 2, \dots, q\}\}$ and by approximately (via simulations) computing

$$\beta_{h-1}^s = \min_{u_{h-1} \in U_{h-1}(x_{h-1}^s)} \mathbb{E}\left[g_{h-1}\left(x_{h-1}^s, u_{h-1}, w_{h-1}\right) + g_h\left(f_{h-1}\left(x_{h-1}^s, u_{h-1}, w_{h-1}\right)\right)\right]$$

The state belongs to $\mathbb{R}^n$ with $n = 1$ and the pairs $(x_{h-1}^s, \beta_{h-1}^s)$ are summarized in the table below.

s	1	2	3	4	5
x_{h-1}^s	-4	-2	0	1	3
β_{h-1}^s	17.49	6.87	0.73	-0.55	-1.19

Suppose the following linear parametrization architecture is used for approximating the cost-to-go at stage $h - 1$

$$\tilde{J}_{h-1}(x_{h-1}, r_{h-1}) = \sum_{\ell=1}^{m_{h-1}} r_{\ell,h-1} \phi_{\ell,h-1}(x_{h-1}) = r_{h-1}^\top \phi_{h-1}(x_{h-1})$$

and the parameters of such architecture are found as follows

$$\hat{r}_{h-1} = \arg\min_r \sum_{s=1}^{q} \left(\tilde{J}_{h-1}\left(x_{h-1}^s, r\right) - \beta_{h-1}^s\right)^2$$

Compute $\hat{r}_{h-1}$ for $m_{h-1} = 3$ assuming the feature functions are

$$\phi_{\ell,h-1}(x) = x^{\ell-1}$$

for every $\ell \in \{1, 2, 3\}$.

Problem 1.6 Repeat Problem 1.3 but consider $m_{h-1} = 2$ assuming the feature functions are

$$\phi_{1,h-1}(x) = \cos(x) \quad \phi_{2,h-1}(x) = \sin(x)$$

Infinite-horizon problems, Policy iteration

Problem 2.1 Consider the following finite horizon optimal control problem

$$\mathbb{E}[\sum_{k=0}^{2} g(x_k, u_k) + g_3(x_h)]$$

$$x_{k+1} = \min(\max(x_k + u_k + w_k, 0), 3)$$

with $g(x_k, u_k) = x_k + u_k$ and $g_3(x_3) = x_3$, and $x_k \in \{0, 1, 2, 3\}$, $u_k = \{-1, 0, 1\}$, $w_k = \{-1, 0, 1\}$, $\text{Prob}[w_k = -1] = \text{Prob}[w_k = 0] = \text{Prob}[w_k = 1] = \frac{1}{3}$.

(i) Write it as a shortest path problem.

(ii) Write the Bellman equation for the shortest path problem and show that the solution of the Bellman equation provides the optimal cost-to-go for every $k \in \{0, 1, 2\}$ for the original problem.

Problem 2.2 Consider the following discounted cost optimal control problem

$$\sum_{k=0}^{\infty} \mathbb{E}[\alpha^k g(x_k, u_k)]$$

$$x_{k+1} = \min(\max(x_k + u_k + w_k, 0), 3)$$

with $\alpha = 0.6$, $g(x_k, u_k) = x_k + u_k$ and $g_3(x_3) = x_3$, and $x_k \in \{0, 1, 2, 3\}$, $u_k = \{-1, 0, 1\}$, $w_k = \{-1, 0, 1\}$, $\text{Prob}[w_k = -1] = \text{Prob}[w_k = 0] = \text{Prob}[w_k = 1] = \frac{1}{3}$.

(i) Write it as a shortest path problem.

(ii) Write the Bellman equation for the shortest path problem and show that it coincides with the Bellman equation for the original discounted cost problem.

Problem 2.3 Consider an inventory control problem where the number of items of a given product over h 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 denotes the capacity. The objective is to minimize

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

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

$$c_2(j) = \begin{cases} 5j + 0.9 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 9$ is the selling price per item, and $g_h(i) = -4i$, $i \in \{0, \dots, N\}$, is the terminal cost. Suppose that $h = 3$, $N = 2$, $d_0 = 1$, $d_2 = 1$ and that the demand d_1 is uncertain and characterized by $\text{Prob}[d_1 = 0] = \text{Prob}[d_1 = 1] = \text{Prob}[d_1 = 2] = 1/3$.The optimal policy which minimizes the expected value of the cost (3) can be obtained via **policy iteration**. In this context, run one step of policy evaluation, i.e., compute the cost-to-go of the initial policy $\mu_k^0(i) = 0$ for every k, i, and one step of policy improvement, obtaining an improved policy $\mu_k^1(x_k)$.

Problem 2.4 Consider an inventory control problem where the number of items of a given product is described by

$$x_{k+1} = \max\{x_k + u_k - d_k, 0\}, \quad k \in \{0, 1, \dots\}, \tag{2}$$

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 denotes the capacity. Suppose that $N = 2$, and that the demand d_k is uncertain and characterized by $\text{Prob}[d_k = 0] = \text{Prob}[d_k = 1] = \text{Prob}[d_k = 2] = 1/3$ for every $k \in \{0, 1, \dots\}$. The objective is to find a policy that minimizes

$$\sum_{k=0}^{\infty} \mathbb{E}[\alpha^k \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big)].$$

for $\alpha = 0.9$ where $c_1(i) = i$, $i \in \{0, \dots, N\}$, is the storage cost,

$$c_2(j) = \begin{cases} 5j + 0.9 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 9$ is the selling price per item. The optimal policy which minimizes the expected cost can be found using **policy iteration**. In this context: (i) run one step of policy evaluation, i.e., compute the cost function of the initial policy $\mu_0(i) = 0$ for every $i \in \{0, 1, \dots, N\}$,

$$J_{\mu_0}(i) = \sum_{k=0}^{\infty} \mathbb{E}[\alpha^k \big(c_1(x_k) + c_2(\mu_0(x_k)) - p \min\{x_k + \mu_0(x_k), d_k\}\big)|x_0 = i], \quad i \in \{0, 1, \dots, N\},$$

when $u_k = \mu_0(x_k)$ in (2); and (ii) run one step of policy improvement, obtaining an improved policy $\mu_1(i)$, $i \in \{0, 1, 2\}$.

Problem 2.5 Suppose that we wish to sell a used car by finding an optimal policy to accept offers. The car is exactly two years old and the offer for cars of this model and of this age in the initial month of the sale follows a given distribution modeled by a random variable w_0 which takes values in a discrete space $\{c_1, c_2, \dots, c_C\}$. Consider $c_i = 10000 + (i - 1)1000$ euros $, C = 5)$ with $\text{Prob}[w_0 = c_i] = p_i$, $\sum_{i=1}^{C} p_i = 1 - p_0$, where$p_0$ is the probability that there is no offer. Assume that $p_0 = 0.01$, $p_1 = 0.05$, $p_2 = 0.2$, $p_3 = 0.3$, $p_4 = 0.3$, $p_5 = 0.1$. The car depreciates at a rate α every month after the initial one. That is if $\bar{x}_k$ is the offer at month k after the initial one (at which the car's age is exactly two years), then $\frac{\bar{x}_k}{\alpha^k}$ has the same probability distribution as w_0. After 10 years the residual value of the car is almost zero. Therefore we can consider the following cost

$$\sum_{k=0}^{\infty} \alpha^k g(x_k, u_k)$$

to be minimized, where $u_k \in \{0, 1\}$is the decision at time k either to sell ($u_k = 1$) or not ($u_k = 0$); the state x_k can take either a numerical value or two non-numerical values: $x_k = \text{N}$ if there is no offer and $x_k = T$ if an offer was already accepted. In the case that non of the two latter cases is met, x_k is equal to

$$x_k = \frac{\bar{x}_k}{\alpha^k} \in \{c_1, c_2, \dots, c_C\}$$

i.e., the value of the offer at month k multiplied by $1/\alpha_k$so that we can model

$$x_k = w_{k-1}, \text{ if } x_k \neq N, x_{k-1} \neq T \text{ and } u_{k-1} = 0$$

,

$$x_k = T \text{ if } u_{k-1} = 1 \text{ or } x_{k-1} = T$$

where w_k are independent and identically distributed random variables (with the same probability distribution) as w_0. The cost function is such that $g(N, u_k) = 0$ for every u_k, $g(T, u_k) = 0$ for every u_k, and

$$g(x_k, 1) = x_k, \quad g(x_k, 0) = 0.$$

Compute the optimal policy using policy iteration, assuming the initial policy $\mu_0(x_0)$ is do not sell for any state, i.e.,

$$\mu_0(x_k) = 0 \text{ if } x_k \in \{10000, 11000, 12000, 13000, 140000\}$$

Q-Learning

Problem 3.1 Consider an inventory control problem where the number of items of a given product over h stages is described by

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

where $x_k \in \{0, \dots, N\}$, $u_k \in U_k(x_{k+1}) := \{0, \dots, N - x_k\}$, d_k denote the number of items, the supply and the demand at time k, respectively, and N denotes the capacity. The objective is to minimize

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

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

$$c_2(j) = \begin{cases} 5j + 0.9 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 9$ is the selling price per item, and $g_h(i) = -4i$, $i \in \{0, \dots, N\}$, is the terminal cost. Suppose that $h = 3$, $N = 2$, $d_0 = 1$, $d_2 = 1$ and that the demand d_k is uncertain and characterized by $\text{Prob}[d_k = 0] = \text{Prob}[d_k = 1] = \text{Prob}[d_k = 2] = 1/3$ for every $k \in \{0, 1, \dots\}$.

Suppose that Q-Learning is used to obtain the optimal policy. The estimates at iteration ℓ, coinciding with each new state/ control pair visited, for a given state $i \in \{0, 1, \dots, N\}$, at stage $k \in \{0, 1, \dots, h-1\}$ using control $u \in \{0, \dots, N - i\}$ is denoted by $Q_\ell(k, i, u)$. A given random policy with random initial conditions is used to generate trajectories leading to the data/ simulations summarized in the table below. Suppose that all the Q-estimates are initialized with zero values $Q_1(k, i, u) = 0$ for every k, i, u.

ℓ	1	2	3	4	5	6	7	8	9	10
k	0	1	2	3	0	1	2	3	0	1
x_k	1	0	2	2	0	0	1	0	0	1
u_k	1	2	0	-	1	2	1	-	1	1
d_k	2	0	0	-	1	1	2	-	0	1

Provide the estimates after $L = 10$ iterations, $Q_{11}(k, i, u)$ for every k, i, u. Assume that the parameters γ^k in the update equations are constant and equal to $\gamma^k = 0.5$.

$$Q_{\ell+1}(k, x_k, u_k) = \left(1 - \gamma^\ell\right) Q_\ell(k, x_k, u_k) + \gamma^\ell \left(g(x_k, u_k, d_k) + \min_{v \in U_k(x_{k+1})} Q_\ell(k+1, x_{k+1}, v) \right), \quad k \neq h-1$$

$$Q_{\ell+1}(h-1, x_{h-1}, u_{h-1}) = \left(1 - \gamma^\ell\right) Q_\ell(h-1, x_{h-1}, u_{h-1}) + \gamma^\ell \left(g(x_h, u_{h-1}, d_{h-1}) + g_{h-1}(x_h) \right),$$

where $g(x_k, u_k, d_k) = \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big)$.

Problem 3.2 Consider an inventory control problem where the number of items of a given product is described by

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

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 denotes the capacity. Suppose that $N = 2$. The objective is to minimize

$$\sum_{k=0}^{\infty} \mathbb{E}[\alpha^k \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big)].$$

for $\alpha = 0.9$ where $c_1(i) = i$, $i \in \{0, \dots, N\}$, is the storage cost,

$$c_2(j) = \begin{cases} 5j + 0.9 \text{ if } j > 0 \\ 0 \text{ if } j = 0 \end{cases},$$

is the cost of ordering j items, $p = 9$ is the selling price per item.

Suppose that Q-Learning is used to obtain the optimal policy. The estimates at iteration ℓ, coinciding with each new state/ control pair visited, for a given state $i \in \{0, 1, \dots, N\}$, at stage $k \in \{0, 1, \dots, \}$ using control $u \in \{0, \dots, N - i\}$ is denoted by $Q_\ell(k, i, u)$. A given random policy with random initial conditions is used to generate trajectories leading to the data/ simulations summarized in the table below for the first 10 iterations. Suppose that all the Q-estimates are initialized with zero values $Q_1(k, i, u) = 0$ for every k, i, u.

ℓ	1	2	3	4	5	6	7	8	9	10
k	0	1	2	3	4	5	6	7	8	9
x_k	1	0	2	2	0	0	1	0	0	1
u_k	1	2	0	0	1	2	1	2	1	0.8
d_k	2	0	0	2	1	1	2	2	0	0.8

Provide the estimates after $L = 10$ iterations, $Q_{11}(k, i, u)$ for every k, i, u. Assume that the parameters γ^k in the update equations are constant and equal to $\gamma^k = 0.6$.

$$Q_{\ell+1}\left(k, x_k, u_k\right) = \left(1 - \gamma^\ell\right) Q_\ell\left(k, x_k, u_k\right) + \gamma^\ell \left(g\left(x_k, u_k, d_k\right) + \min_{v \in U_k(x_{k+1})} Q_\ell\left(k + 1, x_{k+1}, v\right) \right),$$

where $g(x_k, u_k, d_k) = \big(c_1(x_k) + c_2(u_k) - p \min\{x_k + u_k, d_k\}\big)$.

Problem 3.3 Suppose that we wish to sell a used car by finding an optimal policy to accept offers. The car is exactly two years old and the offer for cars of this model and of this age in the initial month of the sale follows a given distribution modeled by a random variable w_0 which takes values in a discrete space $\{c_1, c_2, \dots, c_C\}$. Consider $c_i = 10000 + (i - 1)1000$ euros ,$C = 5$) with Prob$[w_0 = c_i] = p_i$, $\sum_{i=1}^{C} p_i = 1 - p_0$, where$p_0$ is the probability that there is no offer. Assume that $p_0 = 0.01$, $p_1 = 0.05$, $p_2 = 0.2$, $p_3 = 0.3$, $p_4 = 0.3$, $p_5 = 0.1$. The car depreciates at a rate α every month after the initial one. That is if $\bar{x}_k$ is the offer at month k after the initial one (at which the car's age is exactly two years), then $\frac{\bar{x}_k}{\alpha^k}$ has the same probability distribution as w_0. After 10 years the residual value of the car is almost zero. Therefore we can consider the following cost

$$\sum_{k=0}^{\infty} \alpha^k g(x_k, u_k)$$

to be minimized, where $u_k \in \{0, 1\}$is the decision at time k either to sell ($u_k = 1$) or not ($u_k = 0$); the state x_k can take either a numerical value or two non-numerical values: $x_k = \text{N}$ if there is no offer and $x_k = T$ if an offer was already accepted. In the case that non of the two latter cases is met, x_k is equal to

$$x_k = \frac{\bar{x}_k}{\alpha^k} \in \{c_1, c_2, \dots, c_C\}$$

i.e., the value of the offer at month k multiplied by $1/\alpha_k$so that we can model

$$x_k = w_{k-1}, \text{ if } x_k \neq N, x_{k-1} \neq T \text{ and } u_{k-1} = 0$$

,

$$x_k = T \text{ if } u_{k-1} = 1 \text{ or } x_{k-1} = T$$

where w_k are independent and identically distributed random variables (with the same probability distribution) as w_0. The cost function is such that $g(N, u_k) = 0$ for every u_k, $g(T, u_k) = 0$ for every u_k, and

$$g(x_k, 1) = x_k, \quad g(x_k, 0) = x_k.$$

Suppose that the model parameters are not available (namely the probabilities c_k of the probabiltiy distribution of w_k) and we want to use a learning technique to compute the optimal policy.

Show that the general Q-Learning algorithm (where $\gamma^\ell = \bar{\alpha}$ is constant).

$$Q^\ell(x_k, 0) = (1 - \bar{\alpha}) Q^\ell(x_k, 0) + x \bar{\alpha} \alpha \max\{Q^\ell(x_{k+1}, 0), Q^\ell(x_{k+1}, 1)\}$$

greatly simplifies and all the Q functions can be computed from the variables β_ℓ that satisfy

$$\beta_{\ell+1} = (1 - \bar{\alpha}) \beta_\ell + \bar{\alpha} \alpha \max\{\beta_\ell, x_{\ell+1}\}$$