Solutions to Assignment 4

Consider the following problem: Let \( X = \{1, 2\}, U = \{1, 2\} \), where \( X \) denotes whether a fading channel is in a good state \( (x = 2) \) or a bad state \( (x = 1) \). There exists an encoder who can either try to use the channel \( (u = 2) \) or not use the channel \( (u = 1) \). The goal of the encoder is send information across the channel.

Suppose that the encoder’s cost (to be minimized) is given by:

\[
c(x, u) = -1_{\{x=2,u=2\}} + \alpha u,
\]

for \( \alpha = 1/2 \).

Suppose that the transition kernel of the channel is given by (if the encoder uses a channel, the channel is likely to be bad in the next state):

\[
\begin{align*}
P(x_{t+1} = 2|x_t = 2, u_t = 2) &= 0.8, & P(x_{t+1} = 1|x_t = 2, u_t = 2) &= 0.2 \\
P(x_{t+1} = 2|x_t = 2, u_t = 1) &= 0.2, & P(x_{t+1} = 1|x_t = 2, u_t = 1) &= 0.8 \\
P(x_{t+1} = 2|x_t = 1, u_t = 2) &= 0.5, & P(x_{t+1} = 1|x_t = 1, u_t = 2) &= 0.5 \\
P(x_{t+1} = 2|x_t = 1, u_t = 1) &= 0.9, & P(x_{t+1} = 1|x_t = 1, u_t = 1) &= 0.1
\end{align*}
\]

We will consider either a discounted cost for some \( \beta \in (0, 1) \)

\[
\inf_{\Pi} E_x^\Pi \left[ \sum_{t=0}^{\infty} \beta^t c(x_t, u_t) \right]
\]

or an average cost

\[
\inf_{\Pi} \limsup_{T \to \infty} \frac{1}{T} E_x^\Pi \left[ \sum_{t=0}^{T-1} c(x_t, u_t) \right],
\]

over all admissible policies.

**Problem 1: Policy/Value Iteration or Q-Learning [20 Points]**

Using Matlab, obtain a solution to the problem given above in (1) through the following:

(i) [20 Points] Policy Iteration or Value Iteration.

(ii) [20 Points] Q-Learning. Note that a common way to pick \( \alpha \) coefficients in the algorithm is to take for every \( x, u \) pair:

\[
\alpha_t(x, u) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{x_k=x, u_k=u\}}}
\]
For Policy Iteration, see the discussion in Chapter 8, Section 8.1 of the Lecture Notes on the course web site. For Value Iteration, see Theorem 5.4.2. For Q-Learning, see Section 8.2.1 of the Lecture Notes. You may also consult resources on the Internet.

Problem 2: Scalar Kalman Filter

Consider a linear system with the following dynamics:

\[ x_{t+1} = ax_t + u_t + w_t, \]

and let the controller have access to the observations given by:

\[ y_t = x_t + v_t. \]

Here \( \{w_t, v_t, t \in \mathbb{Z}\} \) are independent, zero-mean, Gaussian random variables, with variances \( E[w^2] \) and \( E[v^2] \). The controller at time \( t \in \mathbb{Z} \) has access to \( I_t = \{y_s, u_s, \quad s \leq t - 1\} \cup \{y_t\} \).

The initial state has a Gaussian distribution, with zero mean and variance \( E[x_0^2] \), which we denote by \( \nu_0 \). We wish to find for some \( r > 0 \):

\[ \inf_{\Pi} J(x_0, \Pi) = E_{\nu_0} \left[ \sum_{t=0}^{3} x_t^2 + ru_t^2 \right], \]

Compute the optimal control policy and the optimal cost. It suffices to provide a recursive form.

Hint: Explicitly show that the optimal control has a separation structure. Compute the conditional estimate through the Kalman Filter.

Solution:

You can obtain the solution directly from the lecture notes on the web site. In the following, we take an more instructive approach, repeating a number of steps followed while deriving the Kalman filter in class. We apply dynamic programming. Let us first consider the control at time \( t = 3 \).

It follows that

\[ J_3(I_3) = \min_{u_3} E \left[ x_3^2 + ru_3^2 | I_3, u_3 \right], \]

which leads to the solution \( u_3 = 0 \).

For \( t = 2 \), we have

\[ J_2(I_2) = \min_{u_2} E \left[ x_2^2 + ru_2^2 + E[J_3(I_3)] | I_2, u_2 \right], \]

or

\[ J_2(I_2) = \min_{u_2} E \left[ x_2^2 + ru_2^2 + P_3 x_3^2 | I_2, u_2 \right], \]

where \( P_3 \) is equal to 1, but we would like to keep it as \( P_3 \) to generate a parametric recursion. Now, we have

\[ J_2(I_2) = \min_{u_2} E \left[ x_2^2 + ru_2^2 + P_3(ax_2 + u_2 + w_2)^2 | I_2, u_2 \right], \]
The quantity
\[
E\left[x_2^2 + ru_2^2 + P_3(ax_2 + u_2 + w_2)^2\right|I_2, u_2]
\]
\[
e = E\left[(1 + P_3a^2)x_2^2 + 2aP_3x_2u_2 + (r + P_3)u_2^2\right|I_2, u_2]
+ P_3E[w_2^2]
\]
\[
e = E\left[(\sqrt{r + P_3}u_2 + \frac{aP_3}{\sqrt{r + P_3}}x_2)^2 + x_2^2(1 + P_3a^2 - \left(\frac{P_3a}{r + P_3}\right)^2)\right|I_2, u_2]
+ P_3E[w_2^2],
\] is minimized by
\[
u_2 = -\frac{aP_3}{r + P_3}E[x_2|I_2],
\]
and
\[
J_2(I_2) = E[P_2x_2^2] + E[K_2(x_2 - E[x_2|I_2])^2] + P_3E[w_2^2],
\]
where
\[
P_2 = 1 + P_3a^2 - \left(\frac{P_3a}{r + P_3}\right)^2,
\]
and
\[
K_2 = \frac{a^2P_3}{r + P_3}.
\]
For time \(t = 1\) and \(t = 0\), we have, by the same recursions:
\[
J_0(I_0) = P_0E[x_0^2] + \sum_{t=0}^{2} \left(K_t(E[(x_t - E[x_t|I_t])^2]) + P_{t+1}E[w_t^2]\right),
\]
with
\[
P_t = 1 + P_{t+1}a^2 - \left(\frac{P_{t+1}a}{r + P_{t+1}}\right)^2
\]
and
\[
K_t = \frac{a^2P_{t+1}^2}{r + P_{t+1}}.
\]
We now need to compute \(E[(x_t - E[x_t|I_t])^2]\) and \(E[x_t|I_t]\) for \(t = 0, 1, 2\).
Now, note that,
\[
E[x_0|y_0] = \frac{E[x_0^2]}{E[x_0^2] + E[v_0^2]}(y_0),
\]
and
\[
\Sigma_0 := E[(x_0 - E[x_0|y_0])^2] = \frac{(E[x_0^2])(E[v_0^2])}{E[x_0^2] + E[v_0^2]}.
\]
For time \(t = 1\), we consider \(x_1 = ax_0 + u_0 + w_0\) and write this as
\[
x_1 = aE[x_0|y_0] + a(x_0 - E[x_0|y_0]) + u_0 + w_0,
\]
and
\[
y_1 = x_1 + v_1.
\]
As such, now the controller wishes to estimate \(a(x_0 - E[x_0|y_0]) + w_0\); since \(aE[x_0|y_0]\) and \(u_0\) are available at the controller.
In the next iteration, the controller updates its estimate as:

\[
E[x_1|y_0, u_0, y_1] = aE[x_0|y_0] + u_0 + E[(a(x_0 - E[x_0|y_0]) + u_0)|y_1, y_0, u_0]
\]

or

\[
E[x_1|y_0, u_0, y_1] = aE[x_0|y_0] + u_0 + \alpha_1(y_1 - aE[x_0|y_0] - u_0)
\]

where

\[
\alpha_1 = \frac{a^2\Sigma_0 + E[w_0^2]}{a^2\Sigma_0 + E[w_0^2] + E[v_1^2]}
\]

We compute:

\[
\Sigma_1 := E[(x_1 - E[x_1|I_1])^2] = \frac{(a^2\Sigma_0 + E[w_0^2])(E[v_1^2])}{(a^2\Sigma_0 + E[w_0^2]) + E[v_1^2]}
\]

We now apply the same procedure for time \(t = 2\):

\[
E[x_2|y_0, u_0, y_1, u_1, y_2] = aE[x_1|y_1] + u_1 + \alpha_2(y_2 - a(E[x_1|I_1] + u_0))
\]

where

\[
\alpha_2 = \frac{a^2\Sigma_1 + E[w_1^2]}{a^2\Sigma_1 + E[w_1^2] + E[v_2^2]}
\]

Thus,

\[
\Sigma_2 := E[(x_2 - E[x_2|I_2])^2] = \frac{(a^2\Sigma_1 + E[w_1^2])(E[v_2^2])}{(a^2\Sigma_1 + E[w_1^2]) + E[v_2^2]}
\]

And, the overall cost is:

\[
J = P_0E_{\nu_0}[x_0^2] + \sum_{t=0}^{2} K_t \Sigma_t + \sum_{t=0}^{2} P_{t+1}E[w_t^2].
\]

**Problem 3: Convex Analytic Method**

a) Consider the convex analytic method we discussed in class. Let \(\mathbb{X}, \mathbb{U}\) be countable sets and consider the occupation measure:

\[
v_T(A) = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{1}_{(x_t \in A)}, \quad \forall A \in \mathcal{B}(\mathbb{X}).
\]

While proving that the limit of such a measure process lives in a specific set, the following is used, which you are asked to prove: Let \(\mathcal{F}_t\) be the \(\sigma\)-field generated by \(\{x_s, u_s, s \leq t\}\). Define a \(\mathcal{F}_t\) measurable process

\[
F_t(A) = \left(\sum_{s=1}^{t} \mathbb{1}_{x_s \in A} - t \sum_{X \times U} P(A|x) \nu_t(x, u)\right),
\]

where \(\Pi\) is some arbitrary but admissible control policy. Show that, \(\{F_t(A), \quad t \in \{1, 2, \ldots, T\}\}\) is a martingale sequence for every \(T \in \mathbb{Z}\).
b) Let, for a Markov control problem, \( x_t \in X, u_t \in U \), where \( X \) and \( U \) are finite sets denoting the state space and the action space, respectively. Consider the optimal control problem of the minimization of 
\[
\lim_{T \to \infty} \frac{1}{T} E_x \left[ \sum_{t=0}^{T-1} c(x_t, u_t) \right],
\]
where \( c \) is a bounded function. Further assume that under any stationary control policy, the state transition kernel \( P(x_{t+1}|x_t, u_t) \) leads to an irreducible Markov Chain.

Does there exist an optimal control policy? Propose a way to find the optimal policy.

c) Consider the example stated earlier with the cost criterion given in (2). Apply your algorithm in part b) to this problem by solving the corresponding linear program. What is the optimal policy? In Matlab, the command `linprog` can be used to solve linear programming problems.

**Solution:**

a) We first observe that for all \( t \in \{1, 2, \ldots, T\} \)
\[
\left( \sum_{s=1}^{t} 1_{(x_s, u_s) \in A} - \sum_{s=0}^{t-1} \sum_{x \times U} P^\pi(A|x) v_t(x, u) \right)
= \left( \sum_{s=1}^{t} 1_{(x_s, u_s) \in A} - \sum_{s=0}^{t-1} \sum_{x \times U} P^\pi(A|x) 1_{((x_s, u_s) = (x, u))} \right)
\]

Thus, for \( t \geq 2 \),
\[
E[F_t(A)|F_{t-1}]
= E \left[ \left( \sum_{s=1}^{t} 1_{(x_s, u_s) \in A} - \sum_{s=0}^{t-1} \sum_{x \times U} P^\pi(A|x) 1_{((x_s, u_s) = (x, u))} \right) \right] |F_{t-1}]
= E \left[ \left( 1_{(x_t, u_t) \in A} - \sum_{x \times U} P^\pi(A|x) 1_{((x_{t-1}, u_{t-1}) = (x, u))} \right) \right] |F_{t-1}]
+ \left( \sum_{s=1}^{t-1} 1_{(x_s, u_s) \in A} - \sum_{s=0}^{t-2} \sum_{x \times U} P^\pi(A|x) 1_{((x_s, u_s) = (x, u))} \right)
= 0
+ \left( \sum_{s=1}^{t-1} 1_{(x_s, u_s) \in A} - \sum_{s=0}^{t-2} \sum_{x \times U} P^\pi(A|x) 1_{((x_s, u_s) = (x, u))} \right) |F_{t-1}]
= F_{t-1}(A),
\]

where (3) follows from the fact that \( E[1_{(x_t, u_t) \in A}|F_{t-1}] = P((x_t, u_t) \in A|F_{t-1}) \).

b) By the discussion in class, it follows that every converging sequence of occupational measures \( v_t(.) \) converges to a set \( \Gamma \) defined by:
\[
\Gamma = \{ \mu : \mu(y \times U) = \sum_{x \times U} P(y|x, u) \mu(x, u) \}.\]
We also observed that every cost achievable under an arbitrary (possibly non-stationary) policy is lower bounded by the value attained by a converging subsequence under such a policy.

Furthermore, since there is a unique invariant distribution $\mu$ for every stationary policy (by the irreducibility condition and the fact that the state space is finite), it follows that, the cost

$$\limsup_{T \to \infty} \frac{1}{T} E_x \left[ \sum_{t=0}^{T-1} c(x_t, u_t) \right],$$

is lower bounded by

$$\sum_{x \times U} \mu(x, u) c(x, u).$$

As such, the problem reduces to one of the minimization of the above quantity, subject to:

$$\mu \in \Gamma.$$ 

Let $\mu^*$ be the optimal occupation measure (this exists since the state space is finite, and thus $\Gamma$ is compact, and $\sum_{x \times U} \mu(x, u) c(x, u)$ is continuous in $\mu(., .)$). This induces an optimal policy $\pi(u|x)$ as:

$$\pi(u|x) = \frac{\mu^*(x, u)}{\sum_{u} \mu^*(x, u)}.$$

Thus, we can find the optimal policy conveniently, without using dynamic programming.

c) The goal is to solve

$$\sum_{x \times U} \nu(x, u) c(x, u).$$

subject to

$$\nu \in \Gamma = \{ \mu : \mu(y) = \sum_{x \times U} P((x, u); y) \mu(x, u) \}.$$ 

which can also be written as

$$\sum_{j} \mu(y, j) = \sum_{x \times U} P((x, u); y) \mu(x, u) \}$$

and

$$\nu(x, y) \geq 0, \sum_{x, y} \nu(x, y) = 1$$

All of these are linear constraints.

**Problem 4: Convex Analytic Method**

Consider a two-state, controlled Markov Chain with state space $\mathbb{X} = \{0, 1\}$, and transition kernel for $t \in \mathbb{Z}_+$:

$$P(x_{t+1} = 0|x_t = 0) = u_t^0$$
\[
P(x_{t+1} = 1|x_t = 0) = 1 - u^0_t \\
P(x_{t+1} = 1|x_t = 1) = u^1_t \\
P(x_{t+1} = 0|x_t = 1) = 1 - u^1_t.
\]

Here \(u^0_t \in [0.2, 1]\) and \(u^1_t \in [0, 0.8]\) are the control variables. Suppose, the goal is to minimize the quantity

\[
\lim_{T \to \infty} \frac{1}{T} E_0 \left[ \sum_{t=0}^{T-1} c(x_t, u_t) \right],
\]

where

\[
c(0, u^0) = 1 + u^0, \\
c(1, u^1) = 1.5, \quad \forall u^1 \in [0, 0.8],
\]

with given \(\alpha, \beta \in \mathbb{R}_+\).

Find an optimal policy and the optimal cost.

**Solution:**

An optimal policy is stationary without any loss, by the convex analytic method. Consider any deterministic policy \(u^0, u^1\). The invariant distribution corresponding to this policy would be

\[
\pi = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} = \begin{bmatrix} 1 - u^1 & 1 - u^0 \\ 2 - u^0 - u^1 & 2 - u^0 - u^1 \end{bmatrix}
\]

The expected cost is then: \(\frac{1 - u^1}{2 - u^0 - u^1} (1 + u^0) + \frac{1 - u^0}{2 - u^0 - u^1} 1.5\). For every fixed \(u^1\), we see by a derivative analysis that the expected cost is increasing in \(u^0\); hence we may take \(u^0 = 0.2\). For \(u^0 = 0.2\), the expected cost is increasing in \(u^1\) hence the optimal policy is to take \(u^1 = 0\).