Examples of Stochastic Optimization Problems

In this chapter, we will give examples of three types of stochastic optimization problems, that is, optimal stopping, total expected (discounted) cost problem, and long-run average cost problem. The setup and solution of these problem will require the familiarity with probability theory. For the sake of convenience, some basic concepts and theorems are listed below. Even though they are necessary for mathematical completeness, they are not pre-requisite for understanding the idea of DP.

1 Basics of probability theory

1.1 Three theorems

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a sequence of random variables $\{X_n : n \in \mathbb{N}\}$ that are defined on this space.

Fatou Lemma: If $X_n \geq 0$ for each n, then

$$E\left[\liminf_n X_n\right] \leq \liminf_n E[X_n].$$

Monotone Convergence Theorem (MCT): Suppose $X_n \geq 0$ for each n, and $X_1 \leq X_2 \leq \cdots \leq X_n \leq \cdots$. Let $X \doteq \lim_n X_n$. Then

$$E[X] = \lim_n E[X_n].$$

Dominated Convergence Theorem (DCT): Suppose $X_n \to X$ (almost surely or in probability). If there exists a random variable $Y \geq 0$ such that $E[Y] < \infty$ and $|X_n| \leq Y$ for each n, then

$$E[X] = \lim_n E[X_n].$$

Tower property of conditional expectation: Suppose X is a random variable, and $(Y_1, \ldots, Y_k)$ is an arbitrary collection of random variables. Then

$$E\left[E[X\,|Y_1, \ldots, Y_k]\right] = E[X].$$

1.2 Martingales

Consider a probability space $(\Omega, \mathcal{F}, P)$ carrying a sequence of random variables $\{X_n : n = 0, 1, \ldots\}$.

The sequence $\{X_n\}$ is said to be a *martingale* (resp. submartinagle, supermartingale) if for every n,

$$E[X_{n+1} \,|X_n, \ldots, X_1, X_0] \equiv (\geq, \ \ \leq) X_n.$$

Roughly speaking, a martingale (resp. submartingale, supermartingale) *on average* stay the same (resp. non-decreasing, non-increasing).

An important concept in martingale theory is the so-called *stopping time*, say τ, which is a mapping $\Omega \to \{0, 1, \ldots, +\infty\}$ such that, for each $n \geq 0$, the event

$$\{\tau = n\} = \{(X_0, \ldots, X_n) \in A_n\}$$

for some Borel set $A_n \in \mathbb{R}^{n+1}$. A stopping time τ is said to be *finite* if $P(\tau < \infty) = 1$, and *bounded* if $P(\tau \leq N) = 1$ for some fixed number N.

Intuitively, a stopping time is a random time such that whether it takes value n (i.e., stops at time n) is totally determined by the historical information $(X_0, X_1, \ldots, X_n)$ up to time n. It **cannot** depend on future information $(X_{n+1}, X_{n+2}, \ldots)$ at all. For example, for any fixed number x, the first passage time (to level x)

$$\tau \doteq \inf\{n \geq 0 : X_n \geq x\}$$

with convention $\inf\{\emptyset\} = +\infty$, is a stopping time, since

$$\{\tau = n\} = \{X_0 < x, X_1 < x, \ldots, X_{n-1} < x, X_n \geq x\}.$$

However, the random time

$$\tau \doteq \inf\{n \geq 0 : X_{n+1} \geq x\}$$

is not a stopping time, since one needs to look one more time step in the future to determine the value of τ, that is

$$\{\tau = n\} = \{X_0 < x, X_1 < x, \ldots, X_n < x, X_{n+1} \geq x\}.$$

Exercise **1** Suppose τ and σ are both stopping times. Show that $\tau \wedge \sigma$, $\tau \vee \sigma$, and $\tau + \sigma$ are also stopping times.

Optional sampling theorem: Suppose $\{X_0, X_1, \ldots, \}$ is a martingale (resp. submartingale, supermartingale), and τ an arbitrary stopping time. Then for any n,

$$E\left[X_{n\wedge\tau}\right] = (\geq, \leq) E[X_0].$$

In particular, if τ is bounded, then

$$E[X_\tau] = (\geq, \leq) E[X_0].$$

Remark **1** Martingales are usually defined in a more general and flexible framework as follows. Suppose $\{X_n\}$ is an integrable sequence of random variables, and $\{\mathcal{F}_n\}$ is a *filtration*; i.e.

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$$

is a sequence of σ-algebras. If X_n is $\mathcal{F}_n$-measurable, and

$$E[X_{n+1}\,|\mathcal{F}_n] = X_n$$

for each n, then $\{X_n\}$ is an $\{\mathcal{F}_n\}$-adapted martingale. The stopping time in general will be defined as mappings from Ω to $\{0, 1, \ldots, +\infty\}$ such that $\{\tau = n\} \in \mathcal{F}_n$ for each n. The optional sampling theorem still holds for in this general setting.

Example (First passage times): Suppose $(Z_1, Z_2, \ldots)$ is a sequence of iid (independent identically distributed) random variables with $P(Z_j = -1) = P(Z_j = +1) = 1/2$. Fix two positive integers $0 < x < b$. Define $X_0 = x$ and for every $n \geq 1$,

$$X_n \doteq x + \sum_{j=1}^{n} Z_j.$$

Define stopping times

$$\tau_0 \doteq \inf\{n \geq 0 : X_n = 0\}, \quad \tau_b \doteq \inf\{n \geq 0 : X_n = b\}.$$

Compute the following quantities:

$$P(\tau_0 < \tau_b), \quad P(\tau_b < \tau_0), \quad E\left[\tau_0 \wedge \tau_b\right].$$

Solution: Clearly $\{X_n : n \geq 0\}$ is a martingale. For every n, the optional sampling theorem yields

$$E\left[X_{n\wedge\tau_0\wedge\tau_b}\right] = E[X_0] = x.$$

Moreover, for every n,

$$0 \leq X_{n\wedge\tau_0\wedge\tau_b} \leq b.$$

Thus, it follows from DCT that

$$E\left[X_{\tau_0\wedge\tau_b}\right] = \lim_n E\left[X_{n\wedge\tau_0\wedge\tau_b}\right] = x.$$

But

$$E\left[X_{\tau_0\wedge\tau_b}\right] = 0 \cdot P(\tau_0 < \tau_b) + b \cdot P(\tau_b < \tau_0).$$

Therefore,

$$P(\tau_b < \tau_0) = x/b, \quad P(\tau_0 < \tau_b) = 1 - P(\tau_b < \tau_0) = (b-x)/b.$$

In order to compute $E[\tau_0 \wedge \tau_b]$, we consider the following process

$$Y_n \doteq X_n^2 - n, \quad n = 0, 1, \ldots$$

It is easy to check that $\{Y_n\}$ is also a martingale. Again, optional sampling theorem gives

$$E\left[Y_{n\wedge\tau_0\wedge\tau_b}\right] = E[Y_0]$$

for every $n \geq 0$. Or equivalently

$$E\left[X^2_{n\wedge\tau_0\wedge\tau_b}\right] = E[n \wedge \tau_0 \wedge \tau_b] + x^2.$$

Letting $n \to \infty$, applying DCT to the LHS and MCT to the RHS, we have

$$E\left[X^2_{\tau_0\wedge\tau_b}\right] = E[\tau_0 \wedge \tau_b] + x^2.$$

But

$$E\left[X^2_{\tau_0\wedge\tau_b}\right] = 0^2 \cdot P(\tau_0 < \tau_b) + b^2 \cdot P(\tau_b < \tau_0) = xb,$$

which yields

$$E[\tau_0 \wedge \tau_b] = x(b-x).$$

This complete the example. ■.

2 Optimal stopping: application to search theory

Economic search theory was pioneered by Stigler and McCall [2, 5]. Consider the model proposed by [5], searching for the lowest prices: At a constant cost of c per draw, an agent can solicit additional offers. Each offer gives a price. The agent's problem is to specify a number of offers to solicit, *in*

advance of searching, so as to minimize the expected minimum price plus the total cost of search. Suppose the offered prices are $\{X_0, X_1, \ldots\}$ that form an independent identically distributed (iid) non-negative sequence with known distribution function F. Then the problem is translated to finding a positive integer n such that

$$E\left[X_0 \wedge X_1 \wedge \ldots \wedge X_{n-1} + nc\right] \doteq a_n$$

is minimized.

This is a static optimization problem, and can be easily solved. Indeed, observe that

$$E[X_0 \wedge X_1 \wedge \ldots \wedge X_{n-1}] = \int_0^\infty [1 - F(x)]^n \, dx,$$

is a non-increasing sequence, and that

$$a_{n+1} - a_n = c - \int_0^\infty [1 - F(x)]^n F(x) \, dx$$

is a non-decreasing sequence. The optimal number of offers is

$$n^* \doteq \inf\left\{n \geq 1 : \int_0^\infty [1 - F(x)]^n F(x) \, dx < c\right\}.$$

This model has been criticized because the optimal number of offers to solicit is fixed *a priori*. Presumably, a more sensible decision rule should allow the agent to *sequentially* determine whether the search should be stopped according to the historical information. For example, if the agent is lucky enough to draw the minimum price in the very first draw, it is pointless to search further, yet the optimal policy provided by [5] would ask the agent to keep searching until n^* offers are solicited. These insights lead to a searching model involving *optimal stopping*, in which the agent has two and only two choices at each time step: either stop or continue.

2.1 Finite-horizon: finitely many offers

In the sequential search model, the agent will be asked to draw the first offer, say $X_0 = x_0$. The agent then has to decide to either draw an additional offer or stop search. This decision has to be remade at each time step. Assume that the total number of offers are finite, and these additional offers $\{X_1, X_2, \ldots, X_N\}$ form an iid non-negative sequence with distribution function F. The goal is to find a stopping time τ taking value in $\{0, 1, 2, \ldots, N\}$ so as to minimize the total expected cost

$$E\left[X_0 \wedge X_1 \wedge X_2 \wedge \ldots X_\tau + \tau c\right]$$

given that $X_0 = x_0$. In other words, we are concerned with the minimal value

$$v(x_0) \doteq \inf_{\{0 \le \tau \le N\}} E\left[X_0 \wedge X_1 \wedge X_2 \wedge \ldots X_\tau + \tau c \,|\, X_0 = x_0\right],$$

and the optimal stopping time τ^*.

The DP algorithm is very similar to that of the deterministic models. One first formally find a candidate solution, and then verify the candidate solution is indeed the optimal solution.

Step 1: Suppose the agent is at time $n = N - 1$ and the minimal price up to time n is, say x. Two choices are present: either stop searching or continue for one more (the final) offer. The first strategy will yield a minimal price x, while the second strategy will yield an expected minimal price $E[x \wedge X_N]$ plus the cost c. Thus the minimal value the agent can achieve, denoted by $V_{N-1}(x)$, should be

$$V_{N-1}(x) = \min\left\{x, c + \int_0^\infty (x \wedge y) dF(y)\right\}.$$

Note this should be true for any x.

Suppose now the agent is at time $n = N - 2$ with minimal price up to the time is, say x. Again two choices are present: either stop searching or continue for one more offer. The first strategy will yield a minimal price x, while the second strategy will yield an expected minimal total cost $c + E\left[V_{N-1}(x \wedge X_{N-1})\right]$. Thus the minimal value the agent can achieve, denoted by $V_{N-2}(x)$, should be

$$V_{N-2}(x) = \min\left\{x, c + \int_0^\infty V_{N-1}(x \wedge y) dF(y)\right\}$$

This should be true for all x also.

In general, suppose at time n the agent observes a minimal price x up to the time n, then the minimal value the agent can attain, denoted by $V_n(x)$, is

$$V_n(x) = \min\left\{x, c + \int_0^\infty V_{n+1}(x \wedge y) dF(y)\right\}.$$

If $V_n(x) = x$, then the agent should stop, and continue for at least one more offer otherwise.

In conclusion, we have the following conjecture: Let

$$V_N(x) \doteq x \tag{1}$$

$$V_n(x) \doteq \min\left\{x, c + \int_0^\infty V_{n+1}(x \wedge y) dF(y)\right\} \tag{2}$$

for $n = 0, 1, \ldots, N-1$. Then $v(x_0) = V_0(x_0)$ and the optimal stopping policy is

$$\tau^* \doteq \inf\{n \geq 0 : V_n(X_0 \wedge X_1 \wedge \ldots \wedge X_n) = X_0 \wedge X_1 \wedge \ldots \wedge X_n\}$$

Step 2: We will show that the conjectured in Step 1 indeed gives the optimal solution. For notational simplicity, let

$$Y_n \doteq X_0 \wedge X_1 \wedge \ldots \wedge X_n.$$

Observe $Y_0 \equiv X_0 = x_0$.

First we show that $V_0(x_0) \leq v(x_0)$. Consider the process

$$Z_n \doteq V_n(Y_n) + nc; \quad 0 \leq n \leq N.$$

We claim that $\{Z_n\}$ is a submartingale. Actually, since $\{X_n\}$ are iid,

$$\begin{aligned} E\left[Z_{n+1} \,|X_n, \ldots, X_0\right] &= (n+1)c + E\left[V_{n+1}(Y_n \wedge X_{n+1}) \,|X_n, \ldots, X_0\right] \\ &= (n+1)c + \int_0^\infty V_{n+1}(Y_n \wedge y)\, dF(y) \\ &\geq nc + V_n(Y_n) \\ &= Z_n, \end{aligned}$$

where the last inequality follows from equation (2).

By optional sampling theorem, we have

$$E[Z_\tau] \geq E[Z_0] = V_0(x_0)$$

for all stopping times taking values in $\{0, 1, \ldots, N\}$. But since $V_n(x) \leq x$ thanks to equation (1)-(2), we have

$$V_0(x_0) \leq E\left[Y_\tau + \tau c\right] = E\left[X_0 \wedge X_1 \wedge \ldots \wedge X_\tau + \tau c\right].$$

Taking infimum over all stopping times on the RHS, we have

$$V_0(x_0) \leq v(x_0).$$

It remains to show that $V_0(X_0) = E[Z_{\tau^*}]$. Because if this true, then

$$\begin{aligned} V_0(X_0) &= E\left[V_{\tau^*}(Y_{\tau^*}) + \tau^* c\right] \\ &= E\left[Y_{\tau^*} + \tau^* c\right] \\ &= E\left[X_0 \wedge X_1 \wedge \ldots \wedge X_{\tau^*} + \tau^* c\right], \end{aligned}$$

which in turns implies that $V_0(X_0) = v(X_0)$ and τ^* is optimal. To this end, we only need to show that, for every $n = 0, \ldots, N-1$,

$$E\left[Z_{\tau^* \wedge (n+1)}\right] = E\left[Z_{\tau^* \wedge n}\right].$$

But

$$\begin{aligned}
&E\left[Z_{\tau^* \wedge (n+1)} \,|X_0, \ldots, X_n\right] \\
&= E\left[1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} Z_{n+1} \,|X_0, \ldots, X_n\right] \\
&= 1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} E\left[Z_{n+1} \,|X_0, \ldots, X_n\right] \\
&= 1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} (n+1)c + 1_{\{\tau^* \geq n+1\}} E\left[V_{n+1}(Y_{n+1}) \,|X_0, \ldots, X_n\right] \\
&= 1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} (n+1)c + 1_{\{\tau^* \geq n+1\}} \int_0^\infty V_{n+1}(Y_n \wedge y)\, dF(y) \\
&= 1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} nc + 1_{\{\tau^* \geq n+1\}} V_n(Y_n) \\
&= 1_{\{\tau^* \leq n\}} Z_{\tau^*} + 1_{\{\tau^* \geq n+1\}} Z_n \\
&= Z_{\tau^* \wedge n}.
\end{aligned}$$

Recalling the tower property of conditional expectations, we complete the proof. ■

2.2 Infinite-horizon: infinitely many offers

In this section, we consider the same problem but with $N = \infty$. In other words, the value function is

$$u(x_0) \doteq \inf_{\tau} E\left[X_0 \wedge X_1 \wedge X_2 \wedge \ldots X_\tau + \tau c \,|\, X_0 = x_0\right],$$

where the infimum is taken over all stopping times taking values in $\{0, 1, \ldots, \infty\}$.

The DPE associated with this infinite-horizon problem is simpler (in some sense) than that of the finite-horizon problem. Even though sometimes an infinite-horizon model is not very practical, it is more likely to yield closed-form solutions and to provide valuable insights.

Again, the DP algorithm will divide into two similar steps.

Step 1 (a): In this step we will formally derive the DPE that the value function u should satisfy. At time $n = 0$, two choices are present: either stop or continue. The first strategy yield a minimal price x_0. If the second strategy is adopted, then the agent will pay cost c and solicit a new offer with price X_1. From then on, the minimal cost the agent can achieve is $u(x_0 \wedge X_1)$. It follows that the minimal total expected cost of the second

strategy is $c + E[u(x_0 \wedge X_1)]$. Thus the minimal value the agent can get from $n = 0$ has to be

$$\min\left\{x_0, c + \int_0^\infty u(x_0 \wedge y)\, dF(y)\right\}.$$

But by definition, this has to $u(x_0)$. We arrive at

$$u(x_0) = \min\left\{x_0, c + \int_0^\infty u(x_0 \wedge y)\, dF(y)\right\}.$$

This argument indeed is independent of the specific values of x_0. Therefore, one will suspect that u is a solution to the following functional equation

$$U(x) = \min\left\{x, c + \int_0^\infty U(x \wedge y)\, dF(y)\right\}. \tag{3}$$

This is the DPE associated with this infinite-horizon problem. The candidate optimal stopping policy is naturally

$$\tau^* = \inf\{n \geq 0 : U(X_0 \wedge X_1 \wedge \ldots \wedge X_n) = X_0 \wedge X_1 \wedge \ldots X_n\}.$$

Step 1 (b): For this model, the DPE also leads to an explicit solution. But let us first do a bit more guess-work. It is not hard to believe that the optimal policy is to stop searching when the minimum price you have so far is low enough. In other words, there should exist a threshold x^* such that the optimal policy is to stop searching whenever the minimum price falls below x^*. Therefore, one should expect $U(x) = x$ if $x \leq x^*$. How about $x > x^*$? Due to the structure of the problem, the specific of x is not important as long as it is bigger than x^*. Thus one should expect that $U(x)$ is a constant for $x > x^*$. If one believes that U is a continuous function, then $U(x) = U(x^*) = x^*$ for all $x > x^*$. We arrive at the following conjecture.

$$U(x) = \begin{cases} x & ; \quad \text{if } x \leq x^* \\ x^* & ; \quad \text{if } x > x^* \end{cases} \tag{4}$$

The threshold x^* is usually called the *free-boundary*.

To determine x^*, substitute (4) into (3) with $x > x^*$. We have

$$x^* = c + \int_0^{x^*} y\, dF(y) + \int_{x^*}^\infty x^*\, dF(y)$$

or,

$$c = \int_0^{x^*} (x^* - y)\, dF(y).$$

Using integration by parts, we see that x^* is determined as the (unique) solution to the following equation

$$c = \int_0^{x^*} F(y)\, dy. \tag{5}$$

We have not completely solve the equation yet. What we have obtained is a to guess a solution U defined by (4) and (5). We still need to verify that this function U is indeed a solution to the DPE (3). But this is just algebra: For $x \le x^*$, the LHS of (3) is x, and the RHS is

$$\begin{aligned} & \min\left\{x,\ c + \int_0^x y\, dF(y) + \int_x^\infty x\, dF(y)\right\} \\ = \ & \min\left\{x,\ x + \int_0^x (x-y)\, dF(y) - c\right\} \\ = \ & \min\left\{x,\ x + \int_0^x dF(y)\, dy - c\right\} \\ = \ & x. \end{aligned}$$

For $x > x^*$, the LHS is x^*, and the RHS is $\min\{x, x^*\} = x^*$. We conclude that the function U given by (4) and (5) is a solution to the DPE (3). The conjectured optimal stopping policy turns out to be

$$\tau^* = \inf\{n \ge 0 :\ X_0 \wedge X_1 \wedge \ldots \wedge X_n \le x^*\} = \inf\{n \ge 0 :\ X_n \le x^*\}. \tag{6}$$

Note this stopping time τ^* is finite, that is $P(\tau^* < \infty) = 1$.

Step 2: We will verify the explicit solution obtained in Step 1 (b) is indeed the optimal solution; namely, $u(x_0) = U(x_0)$ where U is given by (4) and (5), and τ^* by (6) is an optimal stopping policy.

The proof is similar to that of the finite-horizon problem. We first show that $u(x_0) \le U(x_0)$. Again, let $Y_n \doteq X_0 \wedge X_1 \wedge \ldots \wedge X_n$, and $Z_n = U(Y_n) + nc$. The process $\{Z_n\}$ is again a submartingale. Now the optional sampling theorem yields that, for any stopping time τ,

$$E\left[U(Y_{\tau\wedge n}) + (\tau \wedge n)c\right] = E\left[Z_{\tau\wedge n}\right] \ge E[Z_0] = U(x_0).$$

Assume τ is finite for a moment, we can take the limit as $n \to \infty$ on the LHS to obtain

$$E\left[U(Y_\tau) + \tau c\right] \ge U(x_0)$$

thanks to the DCT and MCT. This leads to, since $U(x) \le x$ for all x,

$$E\left[X_0 \wedge X_1 \wedge \ldots \wedge X_\tau + \tau c\right] = E\left[Y_\tau + \tau c\right] \ge U(x_0).$$

This inequality is automatic if τ is not finite with probability one. Now taking infimum over all stopping times, we have $u(x_0) \geq U(x_0)$.

It remains to show

$$E\left[Y_\tau^* + \tau^* c\right] = U(x_0).$$

Since τ^* is finite, $Y_{\tau^*} = U(Y_{\tau^*})$. We only need to show

$$E\left[U(Y_\tau^*) + \tau^* c\right] = U(x_0).$$

But similar to the proof of finite-horizon case, we have

$$E[Z_{\tau^* \wedge (n+1)}] = E[Z_{\tau^* \wedge n}]$$

for all n. In particular,

$$E\left[U(Y_{\tau^* \wedge n}) + (\tau^* \wedge n)c\right] = E[Z_{\tau^* \wedge n}] = E[Z_0] = U(x_0), \quad \forall\ n \geq 0.$$

Letting $n \to \infty$, observing τ^* is finite, we have

$$E[U(Y_{\tau^*}) + \tau^* c] = U(x_0),$$

thanks to DCT and MCT. This completes the verification. ■

Remark **2** The verification argument in Step 2 does not use the specific form of the solution U. It only used two facts in a crucial way: (1) the function U is a solution to the DPE; (2) the stopping time τ^* is finite almost surely. Having the explicit form of U is just a welcome bonus.

Remark **3** There is another way to obtain the DPE (3) using the finite-horizon DPE. Note that in the finite-horizon problem,

$$V_0(x) = \min\left\{x, c + \int_0^\infty V_1(x \wedge y) dF(y)\right\}.$$

Now let the horizon N goes to infinity. It is not hard to believe

$$V_0 \downarrow u, \quad V_1 \downarrow u,$$

and thus

$$u(x) \doteq \min\left\{x, c + \int_0^\infty u(x \wedge y) dF(y)\right\}.$$

This yields the same DPE. More importantly, it also hints a constructive way to obtain the solution to the infinite-horizon DPE, when a closed-from expression is not available.

3 Total expected (discounted) cost problem: application to linear regulator

In this section we study the famous linear regulator (i.e., linear quadratic, LQ) problem. This formulation is widely used in engineering and macroeconomics [4, 1, 3]. Unlike the optimal stopping model where the control (taking two values: either stop or continue) does not affect the dynamics of the system, the controls in the total expected (discounted) cost problem usually affect the system dynamics and taking values in more complicated spaces. For example, in the LQ problem we are going to consider, the control can take values in the whole Euclidean space.

3.1 Finite-horizon: LQ problem

Consider a linear system with state dynamics

$$X_{n+1} = AX_n + u_n + W_{n+1}; \quad n = 0, 1, \ldots \tag{7}$$

given the initial condition $X_0 \equiv x_0$. Here A is a $d \times d$ fixed matrix, $\{X_n \in \mathbb{R}^d\}$ is the state, $\{u_n \in \mathbb{R}^d\}$ is the control, and $\{W_{n+1} \in \mathbb{R}^d\}$ represents the noise of the system. The class of *admissible* controls u_n is defined as

$$\{u_n = F_n(X_0, X_1, \ldots, X_n) : \ F_n \text{ is measurable}\}.$$

The main point of this definition is that the control only depends on the history, and it cannot look into the "future". The objective is to find a control policy so as to attain the minimal total quadratic cost

$$v(x_0) \doteq \inf_{\{u_n\}} E\left[\sum_{j=0}^{N-1} \alpha^j \left(X_j^t Q X_j + u_j^t R u_j\right) \middle| \ X_0 = x_0\right];$$

here the constant $\alpha \in (0, 1]$. If $0 < \alpha < 1$, then the above problem is called finite-horizon discounted cost problem.

We will assume the following conditions throughout.

Condition **1** *1. The noise $\{W_1, W_2, \ldots\}$ is an iid sequence, independent of $\{X_n, u_n\}$, and has mean 0 and finite second moments.*

2. The $d \times d$ matrices Q, R are symmetric and strictly positive definite.

Step 1 (a): We will formally derive the associated DPE. Again, suppose at time $n = N - 1$, the state is $x \in \mathbb{R}^d$. Apparently, the optimal thing to do is to find a control u so as to minimize

$$\alpha^{N-1}(x^t Q x + u^t R u).$$

Denote the minimal value as $\alpha^{N-1}V_{N-1}(x)$. Then

$$V_{N-1}(x) = \inf_{u\in\mathbb{R}^d}\left[x^tQx + u^tRu\right].$$

This equality holds for an arbitrary $x \in \mathbb{R}^d$.

Suppose now at time $n = N-2$, the state is, say $x \in \mathbb{R}^d$. The optimal policy is to choose u so as to minimize the total cost

$$\alpha^{N-2}(x^tQx + u^tRu) + \alpha^{N-1}E\left[V_{N-1}(Ax + u + W_{N-1})\right].$$

Denote the minimal value by $\alpha^{N-2}V_{N-2}(x)$, and replace W_{N-1} by a generic random variable with the same distribution, then

$$V_{N-2}(x) = \inf_{u\in\mathbb{R}^d}\left\{(x^tQx + u^tRu) + \alpha E\left[V_{N-1}(Ax + u + W)\right]\right\}.$$

In general, one would expect that the minimal value one can attain from $t = n$, given the state being $x \in \mathbb{R}^d$, is $\alpha^n V_n(x)$, where

$$V_n(x) = \inf_{u\in\mathbb{R}^d}\left\{(x^tQx + u^tRu) + \alpha E\left[V_{n+1}(Ax + u + W)\right]\right\}.$$

Therefore, the candidate optimal solution is as follows. Let

$$V_N(x) \doteq 0 \tag{8}$$

$$V_n(x) \doteq \inf_{u\in\mathbb{R}^d}\left\{(x^tQx + u^tRu) + \alpha E\left[V_{n+1}(Ax + u + W)\right]\right\}, \tag{9}$$

for every $n = 0, 1, \ldots, N-1$. The equations (8)-(9) are the DPE associated with the LQ problem.

We conjecture that $v(x_0) = V_0(x_0)$ and the optimal control at time $t = n$ is determined by $u_n^*(x)$, which is the minimizer of the RHS of (9)

Step 1 (b). The DPE (8)-(9) can be explicitly solved.

Let us recall first some basics of quadratic optimization. Suppose B is a $d \times d$, symmetric, strictly positive matrix, and C is a $d \times 1$ column vector. Consider the minimization problem

$$\min_{u\in\mathbb{R}^d}\left[u^tBu + 2C^tu\right].$$

Then the minimum is attained at $u^* = -B^{-1}C$, and the minimal value is $-C^tB^{-1}C$.

It turns out that the function V_n is quadratic by straightforward computation. Indeed, suppose that

$$V_{n+1}(x) = x^t K_{n+1} x + c_{n+1},$$

with K_{n+1} a symmetric, positive definite matrix. Thanks to $E[W] = 0$, it is not difficult to check that

$$\begin{aligned} V_n(x) &= \inf_{u \in \mathbb{R}^d} \Big[x^t (Q + \alpha A^t K_{n+1} A) x + u^t (R + \alpha K_{n+1}) u \\ &\qquad + 2\alpha x^t A^t K_{n+1} u + \alpha c_{n+1} + \alpha E[W^t K_{n+1} W] \Big]. \end{aligned}$$

Since R is strictly positive definite, and K_{n+1} is positive definite, $R + K_{n+1}$ is clearly strictly positive definite. It follows that the minimizer

$$u_n^* = -\alpha (R + \alpha K_{n+1})^{-1} K_{n+1} A x$$

and

$$V_n(x) = x^t K_n x + c_n,$$

with

$$K_n = Q + \alpha A^t \left[K_{n+1} - \alpha K_{n+1} (R + \alpha K_{n+1})^{-1} K_{n+1} \right] A,$$

and

$$c_n = \alpha c_{n+1} + \alpha E[W^t K_{n+1} W].$$

It is not difficult to check that K_n preserves the symmetry and positive definiteness.

To conclude, if we define

$$K_N = 0 \tag{10}$$

$$K_n = Q + \alpha A^t \left[K_{n+1} - \alpha K_{n+1} (R + \alpha K_{n+1})^{-1} K_{n+1} \right] A \tag{11}$$

for every $n = 0, 1, \ldots, N-1$, then for every n, $\{K_n\}$ is symmetric and positive definite. The functions

$$V_n(x) \doteq x^t K_n x + \sum_{j=n+1}^{N} \alpha^{j-n} E[W^t K_j W], \qquad n = 0, 1, \ldots, N \tag{12}$$

give the solution to the DPE (8)-(9), and the minimizer

$$u_n^*(x) = -\alpha (R + \alpha K_{n+1})^{-1} K_{n+1} A x. \tag{13}$$

Step 2: We verify that $v(x_0) = V_0(x_0)$ and $\{u_n^*\}$ defined by (13) gives an optimal control policy.

Let $\{u_n\}$ be an arbitrary control sequence. Consider the process

$$Z_n \doteq \sum_{j=0}^{n-1} \alpha^j \left(X_j^t Q X_j + u_j^t R u_j \right) + \alpha^n V_n(X_n), \quad n = 0, 1, \ldots, N.$$

It is easy to show that $\{Z_n\}$ is indeed a submartingale. Actually,

$$Z_{n+1} - Z_n = \alpha^n \left[X_n^t Q X_n + u_n^t R u_n + \alpha V_{n+1}(X_{n+1}) - V_n(X_n) \right].$$

But

$$\begin{aligned} E\left[V_{n+1}(X_{n+1}) \,|X_n, \ldots, X_0\right] &= E[V_{n+1}(AX_n + u_n + W_{n+1}) \,|X_n, \ldots, X_0] \\ &= E[V_{n+1}(Ax + u + W)]|_{x=X_n, u=u_n}. \end{aligned}$$

It follows readily from equation (9) that

$$E\left[Z_{n+1} - Z_n \,|X_n, \ldots, X_0\right] = E[Z_{n+1} \,|X_n, \ldots, X_0] - Z_n \geq 0,$$

or $\{Z_n\}$ is a submartingale. In particular, since $V_N \equiv 0$, we have

$$E\left[\sum_{j=0}^{N-1} \alpha^j \left(X_j^t Q X_j + u_j^t R u_j \right) \right] = E[Z_N] \geq E[Z_0] = V_0(X_0).$$

Taking infimum on the LHS over all control $\{u_n\}$, we have $v(x_0) \geq V_0(x_0)$.

Now consider the process controlled by $\{u_n^*\}$; that is,

$$\begin{aligned} X_0^* &= x_0 \\ X_1^* &= AX_0^* + u_0^*(X_0^*) + W_1 \\ &\vdots \\ X_N^* &= AX_{N-1}^* + u_{N-1}^*(X_{N-1}^*) + W_N. \end{aligned}$$

Then the process $\{Z_n^*\}$ with

$$Z_n^* \doteq \sum_{j=0}^{n-1} \alpha^j \left((X_j^*)^t Q(X_j^*) + (u_j^*)^t R(u_j^*) \right) + \alpha^n V_n(X_n^*)$$

is easily shown to be a martingale (all the inequalities above will become equalities). In particular,

$$E\left[\sum_{j=0}^{N-1} \alpha^j \left((X_j^*)^t Q(X_j^*) + (u_j^*)^t R(u_j^*) \right) \right] = V_0(X_0).$$

This yields that $v(X_0) = V_0(x_0)$ and that $\{u_n^*\}$ defined by (13) is an optimal control sequence. ■

Remark **4** *(The forward form of the DPE)* In the DPE (8)-(9), if we let $\bar{v}_0(x) \doteq V_{N-n}(x)$, then the DPE will take the following "forward" form:

$$\bar{v}_0(x) \doteq 0 \tag{14}$$

$$\bar{v}_{n+1}(x) \doteq \inf_{u\in\mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + \alpha E\left[\bar{v}_n(Ax + u + W)\right] \right\}, \tag{15}$$

for every $n = 0, 1, \ldots, N$. Then the value function

$$v(x_0) = \bar{v}_N(x_0).$$

More precisely, for every $n \in \mathbb{N}$, the function $\bar{v}_n(x)$ is the value function for the LQ problem with time-horizon n and initial state $x \in \mathbb{R}^d$.

Remark **5** The DPE of forward form leads to the following (forward) recursive equations for $\{K_n\}$. Let $P_n \doteq K_{N-n}$. Then

$$\begin{aligned} P_0 &= 0 \\ P_{n+1} &= Q + \alpha A^t \left[P_n - \alpha P_n (R + \alpha P_n)^{-1} P_n \right] A, \end{aligned} \tag{16}$$

and for any $n \in \mathbb{N}$,

$$\bar{v}_n(x) = x^t P_n x + \sum_{j=0}^{n-1} \alpha^{n-j} E[W^t P_j W].$$

Remark **6** *(Certainty equivalence principle)* The optimal control law (13) is independent of the distribution of the noise $\{W_n\}$ (as long as the expected value is 0). It is the same optimal control law that would be obtained from the corresponding deterministic control problem, which can be considered as the special case where $\{W_n\}$ is not random but rather is known and equal to zero (its expected value). This property is called the *certainty equivalence principle*, which does not hold in general but is observed in many stochastic control problems involving linear systems and quadratic criteria.

3.2 Infinite-horizon: discounted LQ problem

The infinite-horizon problem is similar to the finite-horizon, except now the discount factor $\alpha \in (0, 1)$ and the objective is to solve the minimization problem

$$I(x_0) \doteq \inf_{\{u_n\}} E\left[\sum_{j=0}^{\infty} \alpha^j \left(X_j^t Q X_j + u_j^t R u_j \right) \middle| \, X_0 = x_0 \right].$$

Again, let us divide the solution into two steps.

Step 1 (a): Suppose at $n = 0$, an control $u_0 = u$ is adopted. Then at time $n = 1$, the state becomes $X_1 = Ax_0 + u + W_1$. If optimal policy is adopted afterwards, the minimal total expected cost incurred from time $n = 1$ on will be $\alpha I(Ax_0 + u + W_1)$. Therefore, the total expected cost will be

$$\left(x_0^t Q x_0 + u^t R u\right) + \alpha E\left[I(Ax_0 + u + W_1)\right].$$

The optimal thing to do at $n = 0$, obviously, should be to choose a u so as to minimize the above quantity. But the minimal value, by definition, is $I(x_0)$. Thus we have

$$I(x_0) = \inf_{u \in \mathbb{R}^d} \left[\left(x_0^t Q x_0 + u^t R u\right) + \alpha E\left[I(Ax_0 + u + W_1)\right]\right].$$

Since the above argument does not depend on the specific value of x_0, therefore, one should expect that I is a solution to the following equation

$$U(x) = \inf_{u \in \mathbb{R}^d} \left[\left(x^t Q x + u^t R u\right) + \alpha E\left[U(Ax + u + W)\right]\right], \tag{17}$$

where W is a generic random variable with the same distribution as W_1. This is the DPE of the infinite-horizon LQ problem. The optimal policy is expected to be the minimize for the RHS of the DPE.

Step 1 (b): One question is immediate: Does the DPE admit a solution? The answer is positive. Indeed, the DPE (17) admits a solution of quadratic form. Assume that the solution is of form

$$U(x) = x^t P x + c$$

for some symmetric, positive definite matrix P and some constant c. The (17) yields

$$\begin{aligned} x^t P x + c &= \inf_{u \in \mathbb{R}^d} \Big[x^t (Q + \alpha A^t P A) x + u^t (R + \alpha P) u \\ &\qquad + 2\alpha x^t A^t P u + \alpha E[W^t P W] + \alpha c\Big] \\ &= x^t \left[Q + \alpha A^t (P - P\alpha (R + \alpha P)^{-1} P) A\right] x + \alpha c + \alpha E[W^t P W], \end{aligned}$$

Thus, P has to be a solution of the *Riccati equation*

$$P = Q + \alpha A^t (P - \alpha P (R + \alpha P)^{-1} P) A \tag{18}$$

and

$$c = \frac{\alpha}{1 - \alpha} E[W^t P W].$$

In other words, a candidate solution is of form

$$U(x) = x^t P x + \frac{\alpha}{1-\alpha} E[W^t P W], \tag{19}$$

where P is a symmetric, positive definite matrix solving the Riccati equation (18), and the optimal control is given by

$$u^*(x) = -\alpha(R + \alpha P)^{-1} P A x. \tag{20}$$

But how should one solve the Riccati equation (18)? Worse yet, we do not even know if there exists a solution P which is symmetric and positive definite. Fortunately, the Riccati equation does admit such a solution. Even though the solution cannot be expressed in closed form, it can easily be solved numerically. We have the following lemma.

Lemma 1 *Let $\{P_n\}$ be the sequence of symmetric, positive definite matrices introduced by the forward DP algorithm; see Remark 5. Then $\{P_n\}$ converges, and the limit $P \doteq \lim P_n$ is a symmetric, positive definite matrix that satisfies the Riccati equation (18). Moreover, $U(x)$ defined by (19) is the limit of $\bar{v}_n(x)$ (see Remark 5); i.e.*

$$U(x) = \lim_n \bar{v}_n(x).$$

Note $\bar{v}_n(x)$ is the value function for the LQ problem with horizon n and initial state x.

The proof of the lemma is deferred to the Appendix.

Step 2: We verify here that the solution U equals the value function I of the infinite-horizon LQ problem, and the control policy u^* defined by (20) is optimal.

Since $\bar{v}_n$ is the value function for the LQ problem with horizon n, by definition, $I(x_0) \geq \bar{v}_n(x_0)$ for every n. Letting $n \to \infty$, Lemma 1 implies that $I(x_0) \geq U(x_0)$. Now let $\{X_n^*\}$ be the state process corresponding to the control u^*; i.e.,

$$\begin{aligned} X_0^* &= x_0 \\ X_1^* &= AX_0^* + u^*(X_0^*) + W_1 \\ &\vdots \\ X_{n+1} &= AX_n^* + u^*(X_n^*) + W_{n+1} \\ &\vdots \end{aligned}$$

Consider the process

$$Z_n^* \doteq \sum_{j=0}^{n-1} \alpha^j \left((X_j^*)^t Q(X_j^*) + (u^*(X_j^*))^t R(u^*(X_j^*)) \right) + \alpha^n U(X_n^*).$$

It is not difficult to check that $\{Z_n^*\}$ is a martingale. In particular, for every n,

$$U(x_0) = E[Z_n^*] \geq E\left[\sum_{j=0}^{n-1} \alpha^j \left((X_j^*)^t Q(X_j^*) + (u^*(X_j^*))^t R(u^*(X_j^*)) \right) \right].$$

Letting $n \to \infty$ on the RHS, it follows from MCT that

$$U(x_0) \geq E\left[\sum_{j=0}^{\infty} \alpha^j \left((X_j^*)^t Q(X_j^*) + (u^*(X_j^*))^t R(u^*(X_j^*)) \right) \right].$$

In particular, $U(x_0) \geq I(x_0)$, which in turn implies that $U(x_0) = I(x_0)$ and that u^* is optimal. This completes the solution. ■.

Remark **7** If we consider the undiscounted ($\alpha = 1$) total expected cost in the infinite-horizon problem, usually the value function will be infinite. Very strong conditions are needed if one wish the value for a infinite-horizon total expected undiscounted control problem to be finite.

4 Long-run average cost problem: application to linear regulator

The analysis of long-run average cost problems are in general more difficult than for other criteria. Actually there is no complete and powerful theory about this type of optimization problems.

Consider again the linear system that was introduced in Section 3.1, with state dynamics given by equation (7). Let $u \doteq \{u_n : n = 0, 1, \ldots\}$ be an arbitrary control process. The cost associated with this control is defined as

$$\limsup_n \frac{1}{n} E\left[\sum_{j=0}^{n-1} \left(X_j^t Q X_j + u_j^t R u_j \right) \middle| X_0 = x_0 \right] \doteq J(x_0; u).$$

The objective is to select a control policy over the infinite horizon as to minimize this average cost. The value function is thus

$$\rho(x_0) \doteq \inf_{\{u_n\}} J(x_0; u).$$

The definition of the criterion implies immediately that the control policy on an arbitrary bounded time interval will not affect the cost. In our example (and in most applications), this observation leads to the consequence that the value function is independent of the state; i.e., $\rho(x_0) \equiv \rho^*$ independent of the initial state $x_0 \in \mathbb{R}^d$.

Step 1 (a): The derivation of the DPE is not as "clean" as that for other criteria. Consider the corresponding finite horizon problem (with $\alpha = 1$). Using the DP algorithm in the forward form (Remark 4), we have

$$\begin{aligned} \bar{v}_0(x) &= 0 \\ \bar{v}_{n+1}(x) &= \inf_{u \in \mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + E\left[\bar{v}_n(Ax + u + W)\right] \right\}, \end{aligned}$$

where $\bar{v}_n(x)$ is the value function of the LQ problem with horizon n and initial state x. By definition, we have

$$\rho^* \geq \limsup_n \frac{\bar{v}_n(x)}{n}.$$

But assume for a moment that for any $x \in \mathbb{R}^d$, we indeed have

$$\rho^* = \lim_n \frac{\bar{v}_n(x)}{n}, \qquad \forall\ x \in \mathbb{R}^d,$$

and assume further that the value function

$$\bar{v}_n(x) \approx n\rho^* + h(x)$$

for some function h. Then the forward DPE yields

$$(n+1)\rho^* + h(x) = \inf_{u \in \mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + n\rho^* + E\left[h(Ax + u + W)\right] \right\},$$

or

$$\rho^* + h(x) = \inf_{u \in \mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + E\left[h(Ax + u + W)\right] \right\}.$$

This is the DPE for the average-cost LQ problem. In other words, we wish to find a couple (ρ, h) such that

$$\rho + h(x) = \inf_{u \in \mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + E\left[h(Ax + u + W)\right] \right\}. \tag{21}$$

Step 1 (b): The DPE (21) can be explicitly solved. Consider a solution of form $h(x) = x^t \bar{P} x$. Then the DPE becomes

$$\rho + x^t \bar{P} x = \inf_{u \in \mathbb{R}^d} \left\{ (x^t Q x + u^t R u) + E\left[(Ax + u + W)^t \bar{P} (Ax + u + W) \right] \right\}.$$

It follows easily that

$$\rho = E[W^t \bar{P} W] \tag{22}$$

$$\bar{P} = Q + A^t \left[\bar{P} - \bar{P}(R + \bar{P})^{-1} \bar{P} \right] A. \tag{23}$$

In other words, $\bar{P}$ satisfies the Riccati equation. The minimizer is

$$u^*(x) = -(R + \bar{P})^{-1} \bar{P} A x. \tag{24}$$

Remark **8** From Remark 5, we know

$$\bar{v}_n(x) = x^t \bar{P}_n x + \sum_{j=0}^{n-1} E[W^t \bar{P}_j W],$$

where $\bar{P}_0 = 0$, and

$$\bar{P}_{n+1} = Q + A^t \left[\bar{P}_n - \bar{P}_n(R + \bar{P}_n)^{-1} \bar{P}_n \right] A.$$

The proof of Lemma 1 implies that $\lim_n \bar{P}_n = \bar{P}$. Therefore,

$$\rho = E[W^t \bar{P}_j W] = \lim_n \frac{\bar{v}_n(x)}{n}.$$

Step 2: We verify that the solution ρ given by (22) is the value function (i.e., $\rho^* = \rho$), and u^* is an optimal control policy.

Thanks to Remark 8, we know $\lim_n \bar{v}_n(x_0)/n = \rho$. Thus by definition, $\rho^* \geq \rho$. Let $\{X_n^*\}$ be the state process with control policy $u_n^* \equiv u^*(X_n^*)$. Consider the process

$$Z_n \doteq h(X_n^*) + \sum_{j=0}^{n-1} \left[(X_j^*)^t Q X_j^* + (u^*(X_j^*))^t R (u^*(X_j^*)) \right] - n\rho$$

for $n = 0, 1, \ldots$. It is not difficult to verify that $\{Z_n\}$ is indeed a martingale. In particular,

$$h(x_0) = E[h(X_n^*)] + E\left[\sum_{j=0}^{n-1} \left[(X_j^*)^t Q X_j^* + (u^*(X_j^*))^t R (u^*(X_j^*)) \right] \right] - n\rho.$$

Since $\bar{P}$ is positive definite, we have

$$\rho + \frac{1}{n} h(x_0) \geq \frac{1}{n} E\left[\sum_{j=0}^{n-1} \left[(X_j^*)^t Q X_j^* + (u^*(X_j^*))^t R (u^*(X_j^*)) \right] \right].$$

Letting $n \to \infty$, we have $\rho \geq J(x_0; u^*)$. Moreover, by definition,

$$\rho \geq J(x_0; u^*) \geq \rho^* \geq \rho.$$

This implies that $\rho^* = \rho$ and that $\{u^*\}$ is optimal. ■

A Riccati equation

Proof of Lemma 1: We assume for now that $\{P_n\}$ converges to a matrix P. Then P is clearly symmetric and positive definite. Taking limits on both sides of equation (16), we have

$$P = Q + \alpha A^t \left[P - \alpha P(R + \alpha P)^{-1} P\right] A,$$

which is exactly the Riccati equation (18). Moreover, it is not difficult to show by the standard ε-δ language that

$$\sum_{j=0}^{n-1} \alpha^{n-j} E[W^t P_j W] \to \frac{\alpha}{1-\alpha} E[W^t P W].$$

Thus $\bar{v}_n(x) \to U(x)$.

It remains to show that $\{P_n\}$ converges. Note that

$$\bar{v}_n(x) = x^t P_n x + \sum_{j=0}^{n-1} \alpha^{n-j} E[W^t P_j W]$$

is the value function of the LQ problem with horizon n and initial state x. Also note that $\{P_n\}$ is independent of the distribution of $\{W_n\}$. It is not difficult to check that

$$x^t P_n x$$

is the value function of the (deterministic) LQ problem with horizon n, intimal state x, and noise $\{W_j\} \equiv \{0\}$. Thus we must have

$$x^t P_n x \le x^t P_{n+1} x, \quad \forall\ x \in \mathbb{R}^d.$$

Furthermore, for the deterministic control problem, one can let $u_0 = -Ax$, $u_j \equiv 0$ for $j \ge 1$. The corresponding state process is $X_0 = x$, $X_1 = X_2 = \cdots = 0$. The associated cost is

$$x^t Q x + (Ax)^t R (Ax) = x^t (Q + A^t R A) x.$$

Therefore, for any $x \in \mathbb{R}^d$ and any n,

$$x^t P_n x \le x^t (Q + A^t R A) x.$$

It follows that, for any $x \in \mathbb{R}^d$, the sequence $\{x^t P_n x\}$ is a non-decreasing, bounded sequence. So it must converge. This easily implies that $\{P_n\}$ converges. ∎

References

[1] Special issues on linear quadratic guassian problems. *IEEE. Trans. Automatic Control,* AC-16, 1971.

[2] J.J. McCall. Economics of information and job search. *Quarterly Journal of Economics*, 84:113-126, 1970.

[3] T. Sargent. *Dynamic Macroeconomic Theory*. Harvard University Press, 1987.

[4] H.A. Simon. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24:74-81, 1956.

[5] J. Stigler. The economics of informations. *Journal of Political Economy*, 69:213-225, 1961.