IEOR 265 - Lecture 4
Stochastic DP: Perfect Information Case

1 Stochastic Dynamic Programming

We now return to our base problem, where the disturbance vector w_k is re-introduced. Namely, we have the DP problem:

$$J^*(x_0) = \min_{\pi \in \Pi} \mathbb{E}_w \big[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, \mu_k(x_k), w_k) \big]$$

$$x_{k+1} = f_k(x_k, u_k(x_k), w_k), \forall k \in \{0, 1, ..., N-1\} \tag{1}$$

where the expectation on the right-hand side is taken w.r.t. the joint probability distribution of $(w_0, ..., w_{N-1})$. Differently from the deterministic case, we need to obtain closed-loop policies, as they will usually represent a major improvement to open-loop sequences. Hence our goal is to find the optimal policy:

$$\pi^* = \{\mu_0^*(x_0), ..., \mu_{N-1}^*(x_{N-1})\} \tag{2}$$

And due to the disturbance vectors, forward DP will not work in general and we have to rely again the backward DP recursion in order to solve the stochastic DP problem:

$$J_N(x_n) = g_N(x_N) \tag{3}$$

$$J_i(x_i) = \min_{u_i \in U_i(x_i)} \big\{ \mathbb{E}_{w_i} \big[g_i(x_i, u_i, w_i) + J_{i+1}(f_i(x_i, u_i, w_i)) \big] \big\}, \forall i \in \{0, ..., N-1\} \tag{4}$$

1.1 Example: Optimal Stopping Problem

Stochastic DP's are so pervasive that a huge array of problems, such as motion planning, scheduling, inventory control, to portfolio analysis, can be formulated as an Stochastic DP. On this subsection we will study one of such problems: the Optimal Stopping Problem, which has central importance in finance and quantitative trading.

The Optimal Stopping Problem can be described as follows: Suppose you are in charge of some process (or system) and at every point in time you have to make a decision: You can "push a button" and stop the process from evolving, which will incur you some loss, or let the process continue. If you elect to let it continue, you need to specify what action to take, incurring some costs for the continuation of the process. Typically, you only have two possible actions: "to stop" and "to continue". As we shall see, the optimal closed-loop policy will

involve a set of *stopping states*, that is states on which the optimal decision is to stop.

It is remarkable how many practical problems of interest can be framed as an "Optimal Stopping". For instance, suppose you are in charge of a single asset, for example an apartment, for which you are offered an amount of money from period to period. Let's denote by $w_0, ..., w_{N-1}$ the money offers that you receive at each time period from 0 to $N-1$. Assume that $w_k \geq 0, \forall k$, and are random and independent, where if $w_k = 0$, then it means nobody offered you money for the apartment at period k. If you accept the offer, then you can invest the money you obtained at some (known) rate r; If you reject the offer, you wait for next time period to receive the next offer, but once you reject an offer, that offer is lost forever. In addition, if you rejected all offers up to period $N-1$ you must accept the last offer w_{N-1}. The goal is to find a policy for accepting/rejecting offers that maximizes your revenue at the N'th period.

As with all DP problems, we need to start by defining the systems states: We let the state $x_k \in \{\mathbb{R}, T\}$, that is the state x_k can take any value along the real line, augmented with a special *termination state* T. By writing that the system is at a state $x_k = T$ at some time $k \leq N-1$, it means that your apartment has been sold already. On the hand if $x_k \neq T$ at some time $k \leq N-1$, it means the apartment has not been sold yet and the offer under consideration is equal to x_k (which is also equal to w_{k-1}). We take the initial state to a dummy offer, so $x_0 = 0$. According to our notation, we view the offers w_k as our disturbances in the DP framework. Lastly, we let $u_k \in \{\text{"sell"}, \text{"do not sell"}\}$ be our control decision (note that in the Optimal Stopping framework, "sell" is equivalent to "stop" and "do not sell" is equivalent to "continue").

Now we are in a position to write the dynamics as follows:

$$x_{k+1} = f_k(x_k, u_k, w_k), , \forall k \in \{0, ..., N-1\} \tag{5}$$

where the functions f_k are defined via the relation:

$$x_{k+1} = \begin{cases} T, \text{ if } x_k = T, \text{ or if } x_k \neq T \text{ and } u_k = \text{"sell"} \\ w_k, \text{ otherwise} \end{cases} \tag{6}$$

We observe that the constraint requiring that we must accept the w_{N-1} in case all other offers we rejected is not being enforced by our dynamics above. This will be done in the objective function formulation, as follows:

$$\mathbb{E}_w\big[g_N(x_N) + \sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)\big] \tag{7}$$

where:

$$g_N(x_N) = \begin{cases} x_N, \text{ if } x_N \neq T \\ 0, \text{ otherwise} \end{cases} \tag{8}$$

$$g_k(x_k, u_k, w_k) = \begin{cases} (1+r)^{N-k} x_k, \text{ if } x_k \neq T \text{ and } u_k = \text{"sell"} \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

Now we see that that terminal costs captures the fact that if no other offers has been accept then we must accept the offer w_{N-1}, which by our dynamics is

equal to x_N. On the hand, the stage cost captures the fact that if we sell the apartment at some stage $k \leq N-1$, then we get to invest for the remaining $N-k$ periods, at some rate r.

Based on these formulations, we can write the associate backward DP algorithm:

$$J_N(x_N) = \begin{cases} x_N, \text{ if } x_N \neq T \\ 0, \text{ if } x_N = T \end{cases} \tag{10}$$

$$J_k(x_k) = \begin{cases} \max\{(1+r)^{N-k}x_k, \mathbb{E}_{w_k}[J_{k+1}(w_k)]\}, \text{ if } x_k \neq T \\ 0, \text{ if } x_k = T \end{cases} \tag{11}$$

We observe that Eq.4 simplifies to Eq.11 in this example, since the maximization is taken over only two decisions (to sell or not to sell), hence the maximization boils down to a comparison between the revenue from selling at stage k, that is $(1+r)^{N-k}x_k$ and the expected revenue after rejecting the offer at stage k, that is $\mathbb{E}_{w_k}[J_{k+1}(w_k)]$. Hence we can write the optimal policy as follows:

$$\mu^*(x_k) = \begin{cases} \text{“sell”, if } x_k > \alpha_k \\ \text{“do not sell”, if } x_k \leq \alpha_k \end{cases} \tag{12}$$

where α_k is a *threshold* given by:

$$\alpha_k = \frac{\mathbb{E}_{w_k}[J_{k+1}(w_k)]}{(1+r)^{N-k}} \tag{13}$$

this threshold α_k represents the expected revenue discounted to the present time. This make the optimal policy fairly intuitive: We decide to sell the apartment if the current offer is larger than the expected revenue discounted to the current time; otherwise, we reject the offer and do not sell it. This could be argued with intuitive reasoning, but we can see that it can be explicitly obtained as the optimal closed-loop policy of a stochastic DP.

Now let's add the assumption that all $w_k's$ are i.i.d. and we will drop the subscript k when we perform expectations so $\mathbb{E}_{w_k}[\cdot] = \mathbb{E}_w[\cdot], \forall k$. Now let's define the following *potential* functions:

$$P_k(x_k) = \frac{J_k(x_k)}{(1+r)^{N-k}}, \; x_k \neq T \tag{14}$$

Thus we can substitute back in Eq.10 and in Eq.11, to obtain:

$$P_N(x_N) = x_N = w_{N-1} \tag{15}$$

$$P_k(x_k) = \max\{x_k, (1+r)^{-1}\mathbb{E}_w[P_{k+1}(w)]\}, \forall k \in \{0, 1, ..., N-1\} \tag{16}$$

Now we can re-write the thresholds $\alpha_k's$ as:

$$\alpha_k = \frac{\mathbb{E}_w[V_{k+1}(w)]}{1+r} \tag{17}$$

Now, observe that as long as $x_{N-1} \neq T$, then it holds that:

$$P_{N-1}(x) \geq P_N(x), \forall x \neq T \tag{18}$$

By proceeding backwards, for $N-2$ and $N-1$, and using Eq.18, we obtain for all $x \neq T$:

$$P_{N-2} = \max\{x, (1+r)^{-1}\mathbb{E}_w[P_{N-1}(w)]\} \geq \\ \max\{x, (1+r)^{-1}\mathbb{E}_w[P_N(w)]\} = P_{N-1}(x) \tag{19}$$

Hence we can continue on the same manner backwards and see that:

$$P_k(x) \geq P_{k+1}(x), \forall x \neq T, \forall k \in \{0, ..., N-1\} \tag{20}$$

Now using the Eq.7, the inequality above translate to:

$$\alpha_k \geq \alpha_{k+1}, \forall k \in \{0, ..., N-1\} \tag{21}$$

This is an interesting property: The thresholds are non-increasing, which makes sense. An offer that is good enough to be accept today must be good enough to be accepted tomorrow or any day after, since after each day, there is less and less chance for improvement. This fact can be showcased in a figure:

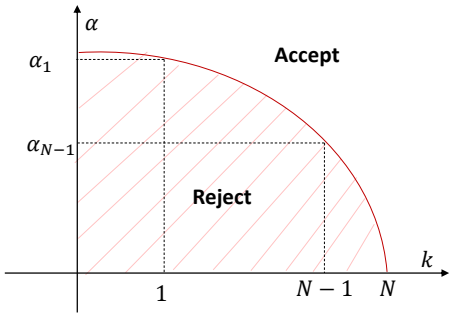

Figure 1: Schematic figure of the rejection region of offers and the

The area "outside" the curve is the *stopping set*: any offer that "falls" into that area triggers the stopping of the process: In our case that means selling the apartment and investing the money.

2 Linear-Quadratic Regulator (LQR)

We will now consider a special case of the DP problem (Eq.1), which is probably the most well-known problem in Optimal Control: the Linear-Quadratic Regulator Problem (LQR). On this section, our DP problem will have linear dynamics and quadratic costs. Namely our dynamics function f_k will be given as follows:

$$x_{k+1} = A_k x_k + B_k u_k + w_k, \forall k \in \{0, 1, ..., N-1\} \tag{22}$$

where A_k and B_k are matrices with appropriate dimensions. And the quadratic cost is given by:

$$\mathbb{E}_w\big[x_N^\top Q_N x_N + \sum_{k=0}^{N-1} (x_k^\top Q_k x_k + u_k^\top R_k u_k)\big] \tag{23}$$

where the matrices Q_k are assumed to be symmetric and positive semidefinite, while the R_k are assumed to be symmetric positive definite. (We will see why such an assumption of definiteness is made). In addition, we assume that w_k are independent of random vectors such that the probability distributions do not depend on x_k and u_k (so $w_k \sim \Pr_k(\cdot)$). Furthermore, each w_k has zero-mean and finite second moment:

$$\mathbb{E}[w_k] = 0,\ \mathbb{E}[w_k^2] < \infty,\ \forall k \in \{0, 1, ..., N-1\} \tag{24}$$

Several important problems in automation and process control can be formulated as an LQR problem. The reason it is called a "regulator" is because the problem can be seen as trying to regulate a system back to the origin $x = 0$: Note that if the system starts at the origin (so $x_0 = 0$), then the disturbances w_k will make the system "drift" away from it. Hence the controls u_k are needed to bring the system back to the origin. This problem is so important, that in many practical applications it is used as the first approach, acting as a benchmark: If you face a problem in automation, a sensible strategy is to first try the LQR formulation before going to more complex modeling and algorithms. Therefore, understanding the DP recursion on the LQR problem is essential as a start point to more sophisticated and complex approaches.

Lastly, it is worth pointing out a variation of such problem: Instead of trying to keep the system state at the origin, our goal can be to track some state reference $(\bar{x}_0, ..., \bar{x}_N)$, then the objective function becomes: (this is called, as expected, the *Tracking Problem* in control):

$$\mathbb{E}_w\big[(x_N - \bar{x}_N)^\top Q_N (x_N - \bar{x}_N) + \sum_{k=0}^{N-1} ((x_k - \bar{x}_k)^\top Q_k (x_k - \bar{x}_k) + u_k^\top R_k u_k)\big] \tag{25}$$

we we point out that if the matrices R_k and Q_k are all identity, then the objective function becomes a least-squares objective.

2.1 LQR and Dynamic Programming

Let's start by writing the Backwards DP algorithm for the LQR problem:

$$J_N(x_N) = x_N^\top Q_N x_N$$

$$J_k(x_k) = \min_{u_k} \left\{ \mathbb{E}_{w_k} \left[x_k^\top Q_k x_k + u_k^\top R_k u_k + J_{k+1}(A_k x_k + B u_k + w_k) \right] \right\} \quad (26)$$

Now we will obtain the optimal closed-loop policy by subsequently applying the DP algorithm. Starting from $k = N-1$:

$$J_{N-1}(x_{N-1}) = \min_{u_{N-1}} \Big\{ \mathbb{E}_{w_{N-1}} \big[x_{N-1}^\top Q_{N-1} x_{N-1} + u_{N-1}^\top R_{N-1} u_{N-1} + (A_{N-1} x_{N-1} + B u_{N-1} + w_{N-1})^\top Q_N (A_{N-1} x_{N-1} + B u_{N-1} + w_{N-1}) \big] \Big\} \quad (27)$$

where we used the definition of $J_N(x_N)$ to obtain the last quadratic term of Eq.8. Now by expanding this last term we can write:

$$\begin{aligned} J_{N-1}(x_{N-1}) = x_{N-1}^\top Q_{N-1} x_{N-1} + \min_{u_{N-1}} \Big\{ & u_{N-1}^\top R_{N-1} u_{N-1} + \\ & u_{N-1}^\top B_{N-1}^\top Q_N B_{N-1} u_{N-1} + 2 x_{N-1}^\top A_{N-1}^\top Q_N B_{N-1} u_{n-1} + \\ & x_{N-1}^\top A_{N-1}^\top Q_N A_{N-1} x_{N-1} + \mathbb{E}_{w_{N-1}} \left[w_{N-1}^\top Q_N w_{N-1} \right] \Big\} \end{aligned} \quad (28)$$

where we used the fact that $\mathbb{E}_{w_{N-1}}[w_{N-1}^\top Q_N (A_{N-1} x_{N-1} + B_{N-1} u_{N-1})] = 0$, since $\mathbb{E}[w_{N-1}] = 0$. In addition, we pushed the expected value all the way to only contain terms that have w_{N-1}.

Now, if we examine Eq.9, we see that the optimization problem is unconstrained and more, the objective function is strictly convex (since R_{n-1} is positive definite). Then we can find the optimal u_{N-1}^* by simply differentiating Eq.9 w.r.t. u_{N-1} and setting the derivative equal to zero. Doing so we obtain:

$$(R_{N-1} + B_{N-1}^\top Q_N B_{N-1}) u_{N-1} + B_{N-1}^\top Q_N A_{N-1} x_{N-1} = 0 \quad (29)$$

Now by moving the term containing x_{N-1} to the right-hand side of Eq.10 and using the fact that $(R_{N-1} + B_{N-1}^\top Q_N B_{N-1})$ is a positive definite matrix (therefore invertible) we can write the optimal control vector as:

$$u_{N-1}^* = -(R_{N-1} + B_{N-1}^\top Q_N B_{N-1})^{-1} B_{N-1}^\top Q_N A_{N-1} x_{N-1} \quad (30)$$

We stop here to highlight that the above expression is **linear** in x_{N-1}. Hence if we define K_{N-1} as:

$$K_{N-1} = -(R_{N-1} + B_{N-1}^\top Q_N B_{N-1})^{-1} B_{N-1}^\top Q_N A_{N-1} \quad (31)$$

Then the optimal control function $\mu_{N-1}^*(x_{N-1})$ becomes:

$$\mu_{N-1}^*(x_{N-1}) = K_{N-1} x_{N-1} \quad (32)$$

which is a linear function of the state. Now, if we substitute, the above in Eq.9, after some algebraic manipulation we can write:

$$J_{N-1}(x_{N-1}) = x_{N-1}^\top P_{N-1} x_{N-1} + \mathbb{E}_{w_{N-1}} \left[w_{N-1}^\top Q_N w_{N-1} \right] \quad (33)$$

where P_{N-1} is equal to:

$$P_{N-1} = A_{N-1}^\top (Q_N - Q_N B_{N-1} (B_{N-1}^\top Q_N B_{N-1} + R_{N-1})^{-1} B_{N-1}^\top Q_N) A_{N-1} + Q_{N-1} \quad (34)$$

and we note that the matrix P_{N-1} is symmetric and positive semidefinite. To see this note that for any x, we have:

$$x^\top P_{N-1} x = \min_u \left\{ x^\top Q_{N-1} x + u^\top R_{N-1} u + (A_{N-1}x + B_{N-1}u)^\top Q_N (A_{N-1}x + B_{N-1}u) \right\} \tag{35}$$

since Q_{N-1} and Q_N are positive semidefinite and R_{N-1} is positive definite, then the expression inside the minimization is nonnegative. Since the minimization preserves the nonnegativity, then it follows that P_{N-1} is positive semidefinite.

Now we can proceed the argument by induction to compute J_{N-2}, and we obtain a similar positive semidefinite matrix P_{N-2}. Thus, proceeding all the way back to $k = 0$ we can obtain the optimal closed-loop policy (Eq. 2):

$$\mu_k^*(x_k) = K_k x_k \tag{36}$$

where the matrix K_k (called the *control gain matrix*) is defined as:

$$K_k = -(R_k + B_k^\top P_{k+1} B_k)^{-1} B_k^\top P_{k+1} A_k \tag{37}$$

and the symmetric positive semidefinite matrices P_k are given by the backwards recursion:

$$\begin{aligned} P_N &= Q_N \\ P_k &= A_k^\top (P_{k+1} - P_{k+1} B_k (B_k^\top P_{k+1} B_k + R_k)^{-1} B_k^\top P_{k+1}) A_k + Q_k \end{aligned} \tag{38}$$

At the end of such recursion, we can write the optimal value function:

$$J_0^*(x_0) = x_0^\top P_0 x_0 + \sum_{k=0}^{N-1} \mathbb{E}_{w_k} \left[w_k^\top P_{k+1} w_k \right] \tag{39}$$

There are several key remarks to be made regarding the solution of the LQR problem:

1. The optimal closed-loop policy is **linear** in the states, which makes policy evaluation (that is computing the actual values) very simple and efficient. This policy is often called the *linear feedback control law*.

2. The backwards recursion necessary to compute the matrices P_k can be done **very** efficiently. The linear algebra involved can be exploited to yield very fast computations. And improvements, such as the "islanding" (very similar to what we saw for HMM's) can be implemented to speed up the computations even more.

3. The optimal value function is **quadratic** in the initial state x_0. The value function and the policy are simple and interpretable. This makes the LQR formulation very desirable from the computation point of view.

4. Lastly, observe that the optimal closed-loop policy **does not** depend on the disturbance vectors $w_k's$. In particular, if we replace each w_k by it's expected value, the optimal policy does not change. In addition, even if the w_k has non-zero mean, the optimal policy will only depend on the uncertainty via it's expectation (we leave the derivation as an exercise). This phenomenon is called the *Certainty Equivalence Principle*, and it appears on most (but not all) stochastic dynamic problems involving linear systems and quadratic costs.

2.2 The Riccati Equation

The backward recursion formula (Eq.19) is so famous that is has a name: *the discrete-time Riccati equation.* One interesting feature of the Riccati equation is that if the matrices are A_k, B_k, Q_k, R_k are all constant across time and equal to A, B, Q, R respectively, then we can extend the solution of the equation to the infinite-horizon setting. We will study infinite-horizon problems when we study Reinforcement Learning, but for now, just imagine a problem where the end period N is very large and "goes to infinity". Under mild assumptions, the matrices P_k will converge to a *steady-state* solution satisfying:

$$P = A^\top (P - PB(B^\top PB + R)^{-1} B^\top P)A + Q \tag{40}$$

This equation (Eq.21) is called the *algebraic Riccati equation.* As we will show, this means that for the linear system:

$$x_{k+1} = Ax_k + Bu_k + w_k, \, \forall k \in \{0, ..., N-1\} \tag{41}$$

with a large number of stages N, one can reasonably approximate the optimal closed-loop policy π^* by a *stationary* policy, that is by $\pi^* = \{\mu^*, \mu^*, ..., \mu^*\}$, where:

$$\mu^*(x) = Kx$$
$$K = -(B^\top KB + R)^{-1} B^\top PA \tag{42}$$

In order to show convergence of the Riccati equation (Eq.19) to it's algebraic form (Eq.21), we need to define two important concepts in Optimal Control: Controllability and Observability:

Definition 1 *A pair of matrices (A, B), where A is an $n \times n$ matrix and B is an $n \times m$ matrix is said to be* ***controllable*** *os the $n \times nm$ matrix:*

$$\begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \tag{43}$$

has full rank (linearly independent rows).

A pair os matrices (A, C), where C is an $m \times n$ is said to be ***observable*** *if the pair $(A^\top, C^\top)$ is controllable.*

The notion of a controllable (A, B) can be explained intuitively: It states that for any initial state x_0, there exists a sequence of control vectors $u_0, u_,..., u_{n-1}$ that force the state x_n of the linear system:

$$x_{k+1} = Ax_k + Bu_k \tag{44}$$

to be equal to zero at time n. Observe that we successfully apply the dynamics to "rollout" the dynamics, obtaining:

$$x_n = A^n x_0 + Bu_{n-1} + ABu_{n-2} + ..., A^{n-1}Bu_0 \tag{45}$$

where x_n is given explicitly as a function of the control sequence $(u_0, u_,..., u_{n-1})$ and the initial state x_0. Equivalently, in matrix notation, we have:

$$x_n - A^n x_0 = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \begin{bmatrix} u_{n-1} \\ u_{n-2} \\ \vdots \\ u_0 \end{bmatrix} \tag{46}$$

Now note that if (A, B) is controllable then the matrix on the rhs of Eq.27 is the same as in Eq.17 and it has full-rank. Then by appropriately selecting $(u_0, u_,..., u_{n-1})$ we can obtain any vector in $\mathbb{R}^n$ (which for example can be $-A^n x_0$, which would yield $x_n = 0$).

The notion of observability allows a similar interpretation, but in the context of estimation: Suppose we obtain some observations $z_0, z_1, ..., x_{n-1}$ of the form:

$$Z_k = Cx_k, \forall k \in \{0, ..., n-1\} \tag{47}$$

It is possible to infer the initial state x_0 of the linear system $x_{k+1} = Ax_k$ by using the following:

$$\begin{bmatrix} z_{n-1} \\ \vdots \\ z_1 \\ z_0 \end{bmatrix} = \begin{bmatrix} CA^{n-1} \\ \vdots \\ CA \\ C \end{bmatrix} x_0 \tag{48}$$

In addition, we can verify that the above is equivalent to the property that, in the absence of control inputs, if $Cx_k \to 0$, then $x_k \to 0$.

Next, we present the very important notion of **stability**: Suppose we utilize the stationary control policy (also called control law):

$$\mu(x_k) = Kx_k, \forall k = 0, 1, ... \tag{49}$$

Then if we replace that in the linear system (Eq.25) we obtain the *closed-loop system*:

$$x_{k+1} = (A + BK)x_k, \forall k = 0, 1, ... \tag{50}$$

We say that the above system is **stable** if x_k tends to zero as $k \to \infty$. By doing the same process as before we can write:

$$x_k = (A + BK)^k x_0 \tag{51}$$

and it follows that the system is stable if and only if $(A+BK)^k \to 0$, which is equivalent to requiring that the eigenvalues of the matrix $(A + BK)$ are strictly within the unit circle. Lastly, we say that the pair of matrices (A, B) are **stabilizable** if there exists a matrix K such that $(A + BK)^k \to 0$.

With these definitions, we can establish the convergence of the Riccati Equation, via the following proposition (taken from [1]):

Proposition 1 *Let A be an $n \times n$ matrix, B be a $n \times m$ matrix, Q be and $n \times n$ positive semidefinite symmetric matrix. Consider the discrete-time Riccati Equation:*

$$P_{k+1} = A_k^\top (P_k - P_k B_k (B_k^\top P_k B_k + R_k)^{-1} B_k^\top P_k) A_k + Q_k, \forall k = \{0, 1, ...\} \tag{52}$$

where, w.l.o.g., the indices's are reversed: we start from an initial positive semidefinite matrix P_0. Assume that the pair (A, B) is controllable. Assume also that Q may be written as $Q = C^\top C$, where the pair (A, C) is observable. Then:

(a) There exists a positive definite symmetric matrix P such that, for every positive semidefinite symmetric initial matrix P_0 we have:

$$\lim_{k\to\infty} P_k = P \tag{53}$$

Further, P is the unique solution of the algebraic matrix equation:

$$P = A^\top (P - PB(B^\top PB + R)^{-1}B^\top P)A + Q. \tag{54}$$

within the class of positive semidefinite symmetric matrices.

(b) The corresponding closed-loop system is stable; that is, the eigenvalues of the matrix:

$$D = A + BK \tag{55}$$

where:

$$K = -(B^\top KB + R)^{-1}B^\top PA \tag{56}$$

are strictly within the unit circle.

proof: The proof is fairly long, and we detail it next, for those who are are interested in the more technical details. The proof follows closely the proof provided in [1]:

We proceed in several steps: First we show that the sequence in Eq.52 converges when the initial matrix P_0 is equal to zero.

Step 1: $P_0 = 0$: Consider the optimal control problem of finding the sequence $u_0, ..., u_{k-1}$:

$$\min \sum_{i=0}^{k-1} x_i^\top Q x_i + u_i^\top R u_i$$
$$\text{s.t.: } x_{i+1} = Ax_i + Bu_i, \forall i \in \{0, ..., k-1\} \tag{57}$$

where x_0 is given. If we apply the discrete-time Riccati Equation we obtain the matrix $P_k(0)$, where it is given by Eq.52 using $P_0 = 0$, and the final optimal objective value is given by $x_0^\top P_k(0)x_0$. Then, for any control sequence $(u_0, .., u_k)$ we can write the inequality:

$$\sum_{i=0}^{k-1}(x_i^\top Q x_i + u_i^\top R u_i) \leq \sum_{i=0}^{k}(x_i^\top Q x_i + u_i^\top R u_i) \tag{58}$$

and:

$$x_0^\top P_k(0)x_0 = \min_{(u_0,...,u_{k-1})} \left\{ \sum_{i=0}^{k-1}(x_i^\top Q x_i + u_i^\top R u_i) \right\} \leq$$

$$\min_{(u_0,...,u_k)} \left\{ \sum_{i=0}^{k-1}(x_i^\top Q x_i + u_i^\top R u_i) \right\} = x_0^\top P_{k+1}(0)x_0$$

where minimizations are done w.r.t the linear system equations $x_{i+1} = Ax_i + Bu_i$. Now, for every x_0 and for every k, $x_0^\top P_k(0)x_0$ is bounded from above by the cost corresponding to a control sequence that forces x_0 to the origin (the zero-vector) in n steps and applying zero control after that (so the system stays at the origin). Such a sequence exists by the controllability assumption. Thus the sequence $\{x_0^\top P_k(0)x_0\}$ is nondecreasing with respect to k and bounded from above, and therefore converges to some real number for every $x_0 \in \mathbb{R}^n$. It follows that the sequence $\{P_k(0)\}$ converges to some matrix P in the sense that each of the sequences of the elements of $P_k(0)$ converges to the corresponding elements of P (this can be verified by using x_0 as the elementary vector). Hence, we can write:

$$\lim_{k\to\infty} P_k(0) = P \tag{59}$$

where $P_k(0)$ are generated using Eq.52 with $P_0 = 0$. Furthermore, since $P_k(0)$ is positive semidefinite and symmetric, so is the limit matrix P. Now by taking the limit in Eq.52 it follows that P satisfies:

$$P = A^\top(P - PB(B^\top PB + R)^{-1}B^\top P)A + Q \tag{60}$$

In addition, by using Eq. 55 and Eq.56:

$$D = A + BK \tag{61}$$

where:

$$K = -(B^\top KB + R)^{-1}B^\top PA \tag{62}$$

we can write the following equation by substituting them back on Eq.60:

$$P = D^\top PD + Q + K^\top RK \tag{63}$$

Step 2: Stability of the Closed-Loop System: Consider the linear system:

$$x_{k+1} = (A + BK)x_k = Dx_k \tag{64}$$

for any initial state x_0. We will show that $x_k \to 0$ as $k \to \infty$. We have for all k, using Eq. 63:

$$x_{k+1}^\top Px_{k+1} - x_k^\top Px_k = x_k^\top(D^\top PD - P)x_k = -x_k^\top(Q + R^\top RK)x_k \tag{65}$$

Hence

$$x_{k+1}^\top Px_{k+1} = x_0^\top Px_0 - \sum_{i=0}^{k} x_i^\top(Q + K^\top RK)x_i \tag{66}$$

The left-hand side of this equation is bounded below by zero, so it follows that:

$$\lim_{k\to 0} x_k^\top(Q + K^\top RK)x_k = 0 \tag{67}$$

since R is positive definite and Q may be written as $C^\top C$, we obtain:

$$\lim_{k\to\infty} Cx_k = 0, \qquad \lim_{k\to\infty} Kx_k = \lim_{k\to\infty} \mu^*(x_k) = 0 \tag{68}$$

Now we due to observability assumption, it follows that $x_k \to 0$. To see this, we can write the following equation, using the linear system (Eq.64):

$$\begin{bmatrix} C(x_{k+n-1} - \sum_{i=1}^{n-1} A^{i-1}BKx_{k+n-i-1}) \\ C(x_{k+n-2} - \sum_{i=1}^{n-2} A^{i-1}BKx_{k+n-i-2}) \\ \vdots \\ C(x_{k+1} - BKx_k) \\ Cx_k \end{bmatrix} = \begin{bmatrix} CA^{n-1} \\ CA^{n-2} \\ \vdots \\ CA \\ C \end{bmatrix} x_k \tag{69}$$

since $Kx_k \to 0$, by Eq.68, the left-hand side tends to zero and hence the right-hand side tends to zero also. By the observability assumption, by multiplying x_k on the right-hand side of Eq.69 has full-rank. So it follows that $x_k \to 0$.

Step 3: Positive-Definiteness of P: We show this by contradiction: Assume there exists a $x_0 \neq 0$ such that $x_0^\top P x_0 = 0$. Since P is positive semidefinite, from Eq.66 we obtain:

$$x_k^\top (Q + K^\top RK) x_k = 0,\ k \in \{0, 1, ...\} \tag{70}$$

Since $x_k \to 0$, we obtain $x_k^\top Q x_k = x_k^\top C^\top C x_k = 0$ and $x_k^\top K^\top RK x_k = 0$, or:

$$Cx_k = 0,\ Lx_k = 0,\ k \in \{0, 1, ...\} \tag{71}$$

Thus all the controls $\mu^*(x_k) = Kx_k$ of the closed-loop system are zero while we have $Cx_k = 0$ for all k. Now using Eq.69 for $k = 0$. By the preceding equalities, the left-hand side is zero and hence:

$$0 = \begin{bmatrix} CA^{n-1} \\ \vdots \\ CA \\ C \end{bmatrix} x_0 \tag{72}$$

Since the matrix multiplying x_0 above has full rank by the observability assumption, we obtain $x_0 = 0$, which contradicts the hypothesis $x_0 \neq 0$ and proves that P is positive definite.

Step 4: Arbitrary Initial Matrix P_0: Now we show that the sequence $\{P_k(P_0)\}$, defined by Eq.52 when the starting matrix in arbitrary positive semidefinite symmetric matrix P_0, converges to $P = \lim_{k\to\infty} P_k(0)$. Consider the optimal control below:

$$\begin{aligned} &\min x_k^\top P_0 x_k + \sum_{i=0}^{k-1} x_i^\top Q x_i + u_i^\top R u_i \\ &\text{s.t.: } x_{i+1} = Ax_i + Bu_i, \forall i \in \{0, ..., k-1\} \end{aligned} \tag{73}$$

and we let $x_k^\top P_0(P_0) x_0$ be the optimal objective value. Then, for every $x_0 \in \mathbb{R}^n$:

$$x_0^\top P_k(0) x_0 \leq x_0^\top P_k(P_0) x_0 \tag{74}$$

Consider now the cost of Eq.73 corresponding to the controller $\mu(x_k) = u_k = Kx_k$, where L is defined by Eq.56, the cost is equal to:

$$x_0^\top \left(D^{k\top} P_0 D^k + \sum_{i=0}^{k-1} D^{i\top} (Q + K^\top RK) D^i \right) x_0 \tag{75}$$

Hence for all k and all $x \in \mathbb{R}^n$, we can write the following inequality:

$$x^\top P_k(0)x \leq x^\top P_k(P_0)x \leq x^\top \left(D^{k\top} P_0 D^k + \sum_{i=0}^{k-1} D^{i\top}(Q + K^\top RK)D^i \right) x \quad (76)$$

We have already prove that:

$$\lim_{k\to\infty} P_k(0) = P \quad (77)$$

and we also have, using the fact $\lim_{k\to\infty} D^{k\top} P_0 D^k = 0$, and the relation $Q + K^\top RK = P - D^\top PD$ (Eq.63) the following:

$$\lim_{k\to\infty} \left\{ D^{k^\top} P_0 D_k + \sum -i = 0^{k-1} D^{i\top}(Q + K^\top RK)D^i \right\} =$$

$$\lim_{k\to\infty} \left\{ \sum -i = 0^{k-1} D^{i\top}(Q + K^\top RK)D^i \right\} =$$

$$\lim_{k\to\infty} \left\{ \sum -i = 0^{k-1} D^{i\top}(P - D^\top PD)D^i \right\} = P \quad (78)$$

Combining the preceding equations we obtain:

$$\lim_{k\to\infty} P_k(P_0) = P \quad (79)$$

for an arbitrary positive semidefinite symmetric initial matrix P_0.

Step 5: Uniqueness of the solution: Lastly, if $\bar{P}$ is another positive semidefinite symmetric solution of the algebraic Riccati Equation (Eq.54), then we have $P_k(\bar{P}) = \bar{P}$ for all $k \in \{0, 1, ...\}$. From the convergence result on Step 4, we then obtain:

$$\lim_{k\to\infty} P_k(\bar{P}) = P \quad (80)$$

which yields $\bar{P} = P$. QED.

References

[1] D. P. Bertsekas, D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.