B9120 Dynamic Programming Lecture number - 02/10/2020

Partial Observability and Infinite Horizon Problems

Lecturer: Daniel Russo *Scribe: Jimmy Qin, Baizhi Song*

Lecture outline:

- Problems with partial observability and the separation principle in LQ control
- Intro to infinite horizon discounted objectives

1 Problems with imperfect state information

Consider a dynamic system that evolves according to $x_{k+1} = f(x_k, u_k, w_k)$ where the disturbances $\{w_k\}$ are independent. At time k, instead of observing the state x_k, we observe $y_k = O_k(x_k, u_{k-1}, \xi_k)$ where $\{\xi_k\}$ is an independent sequence. Effectively, there are known probability distributions $q_{x,u}$ depending on the state and control such that

$$y_k \,|\, x_k, u_{k-1}, \cdots, x_0, u_0 \sim q_{x_k, u_k}(\cdot).$$

As before, the objective is to minimize the cumulative expected cost

$$\inf_{\pi} \; \mathbb{E}^{\pi}_{\{w_k\},\{\epsilon_k\}} \left[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k) + g_N(x_N) \right].$$

But now the infimum is over policies $\pi = (\mu_0, \ldots, \mu_{N-1})$ where $\mu_k(\cdot)$ is a map $H_k \triangleq (y_0, \mu_0, \cdots, y_{k-1}, \mu_{k-1}, y_k) \mapsto \mu_k(H_k) \in U_k$. In other words, μ_k maps from the information available to the controller at time k, i.e. H_k, to some action in the feasible control set U_k.

1.1 Motivation: diabetes management

A lot of problems in healthcare can in principle be formulated using the dynamic programming framework. However, a common challenge there is that we only have access to partial observations. For example, if the state of the system represents the patient's health then, realistically, we only have occasional and noisy measurements of that. Let's consider a problem in diabetes management. Assume that the patient's blood glucose level evolves each day as the following dynamic system

$$x_{k+1} = f(x_k, u_k, w_k)$$

The control set may include things like physical activity activity, measuring glucose, taking insulin etc. We never see the true blood glucose level x_k but instead a noisy measurement of it in case the patient does measure their level at time k.

$$y_k = \begin{cases} x_k + \sigma(x_k)\xi_k & \text{if } \{\text{measure}\} \subset u_k \\ \emptyset & \text{o.w.} \end{cases}$$

1.2 Reduction to the perfect information case

We now show how to reduce the problem from imperfect to perfect state information. Intuitively, a reasonable strategy is to define a new system where the state at time k is the set of all observed variables which can benefit the controller in making the decision at time k. One such candidate is the information vector H_k.

1.2.1 History as state

Equations below can be viewed as describing the evolution of a system of the same nature as the one considered in the basic problem. The state of the system is H_k, the control is u_k, and y_{k+1} can be viewed as a random disturbance. The state at time $k+1$ is H_k augmented with u_k and y_{k+1},

$$H_{k+1} = (H_k, u_k, y_{k+1})$$

Furthermore,

$$\mathbb{P}\left[H_{k+1} = (H_k, u_k, y) \,|\, u_k = u, H_k\right] = \mathbb{P}(y_{k+1} = y | u_k = u, H_k) = \sum_x p_k(x) q_{x,u}(y)$$

where $p_k(x) = \mathbb{P}(x_k = x|H_k)$. Since $(y_0, \ldots, y_k)$ are part of the information vector, the probability distribution of y_{k+1} depends explicitly only on the state H_k and control u_k. We can similarly reformulate the cost function as follows.

$$\tilde{g}_k(H_k, u) = \mathbb{E}[g_k(x_k, u, w_k)|H_k] = \sum_x p_k(x)\mathbb{E}[g_k(x, u, w_k)]$$

where again note that $p_k(x) = \mathbb{P}(x_k = x|H_k)$. For simplicity, assume $g_N = 0$. Then we can write the problem objective as follows.

$$\inf_\pi \; \mathbb{E}^\pi\left[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)\right] = \inf_\pi \; \mathbb{E}^\pi\left[\sum_{k=0}^{N-1} \mathbb{E}[g_k(x_k, u_k, w_k)|H_k]\right] = \inf_\pi \; \mathbb{E}^\pi\left[\sum_{k=0}^{N-1} \tilde{g}_k(H_k, u_k)\right]$$

over policies $\pi = (\mu_0, \ldots, \mu_{N-1})$ where $u_k = \mu_k(H_k)$. However, it is critical to note that H_k grows exponentially in k. As a result this approach is typically computationally infeasible to apply in practice.

1.2.2 Posterior as state

Instead of defining $\tilde{g}(\cdot)$ as a function of history, we define it as a function of our belief about the state x_k, denoted as the posterior distribution p_k.

$$\tilde{g}_k(p_k, u) = \sum_x p_k(x)\mathbb{E}[g_k(x, u, w)]$$

with the corresponding objective being

$$\inf_\pi \; \mathbb{E}^\pi\left[\sum_{k=0}^{N-1} g_k(x_k, u_k, w_k)\right] = \inf_\pi \; \mathbb{E}^\pi\left[\sum_{k=0}^{N-1} \tilde{g}_k(p_k, u_k)\right]$$

which is optimized over policies $\pi = (u_0, \ldots, u_{N-1})$ where $u_k = \mu_k(p_k)$. Here the posterior p_k can be viewed as a sufficient summary of the history H_k and it evolves according to sequential Bayesian updating. Essentially,

$$P_{k+1}(x') = \mathbb{P}(x_{k+1} = x'|y_{k+1}, u_k, H_k) = \sum_x P_k(x)\mathbb{P}(x_{k+1} = x'|y_{k+1}, u_k, x_k = x)$$

The issue with this formulation is that the vector of beliefs can generally take on any value in the probability simplex $\{p \mid p \geq 0, \sum_x p(x) = 1\}$. In general, compting the optimal policy for problems with continuous state vectors of moderate dimension is intractable.

To summarize, we have shown a couple of approaches which in principle can reduce the problem of partial state information to the full state information case, but both these approaches remain computationally impractical. Next, we look at the case LQ control problem which has some special structure allowing us to make this reduction efficiently.

2 Linear quadratic control and the separation principle

Consider the LQ control problem where like before the system state evolves as

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

In this section, we assume that the controller does not have access to the current state x_k. Instead at the beginning of each period k, we observe a noisy measurement of it, $y_k = Cx_k + \xi_k$ where we assume $\{w_k\}, \{\xi_k\}$ to be independent sequences (and also independent of x_0). As before, the objective is to minimize the total cost

$$\inf_{\pi} \ \mathbb{E}^{\pi} \left[\sum_{k=1}^{N-1} (x_k^T Q x_k + u_k^T R u_k + x_N^T Q x_N) \right]$$

over policies $\pi = (\mu_0, \dots, \mu_{N-1})$ where $u_k = \mu_k(H_k)$. Recall that when $y_k = x_k$ (perfect state information case), the optimal policy sets

$$\mu_k^\star(H_k) = L_k x_k$$

where

$$L_i = -(B^T K_{i+1} B + R)^{-1} B^T K_{i+1} A$$

and

$$K_i = A^T (K_{i+1} - K_{i+1} B (B^T K_{i+1} B + R)^{-1} B^T K_{i+1}) A Q$$

for all $i = \{0, \dots, N-1\}$

Proposition 1 (Separation principle)**.** *The optimal policy of the LQ control with imperfect state information is* $\pi^\star = (\mu_0^\star, \dots, \mu_{N-1}^\star)$ *where*

$$\mu_k^\star(H_k) = L_k \cdot \mathbb{E}[x_k | H_k].$$

The matrices L and K can be computed recursively using the same formulae as above.

The optimal policy for LQ control with imperfect state information is very similar to that of the perfect state case. The only difference being that instead of acting on the state x_k, we now plug in our best estimate of the state $\mathbb{E}[x_k|H_k]$. Due to this remarkable fact, one can separate the problem of designing an optimal feedback controller (designing L_k) and the optimal state estimation procedure.

In the important special case where the disturbances $\{w_k\}$, $\{\xi_k\}$ and the initial state x_0 are independent Gaussian vectors, a convenient implementation of computing the conditional mean is possible by means of the Kalman filtering algorithm, which is developed in Appendix E of the textbook.

Warmup. We can get some intuition as to why using the conditional mean might be good in the LQ control case. Consider the optimization problem below with a quadratic estimation loss and a quadratic penalty.

$$\min_{u} \ \mathop{\mathbb{E}}_{x} \left[(x-u)^T Q (X-u) + u^T R u \right]$$

where $Q, R \succ 0$. It is easy to show that the minimizer to this is $u^\star = R^{-1} Q \, \mathbb{E}[x]$, which is a linear function of the mean. When $R = 0$, the optimal objective value is $\mathbb{E}[(x - \mathbb{E}[x])^T Q (x - \mathbb{E}[x])]$, which penalizes the variance of estimation error. Otherwise, the objective value separates into the sum of two terms: one of which depends on the variance of x and one which depends on the mean, which influences the energy cost $u^T R u$. The proof of the separation principle relies on a similar decomposition of the cost-to-go functions.

To prove Proposition 1, we first show the following lemma which states that the *state estimation error*, $x_k - \mathbb{E}[x_k|H_k]$ is independent of the control choice. This is due to the linearity of both the system and the measurement equation. In particular, x_k and $\mathbb{E}[x_k|H_k]$ contain the same linear terms in $(u_0, \dots, u_{k-1})$, which cancel each other out.

Lemma 2. *For every k, the estimation error, $x_k - \mathbb{E}[x_k|H_k]$, does not depend on $u_1, \dots, u_{k-1}$.*

Proof. Since there is no control when $k = 0$, the claim is obviously true. For $k > 0$, we can write x_k recursively as follows.

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_{k-1} + w_{k-1} \\ &= A(Ax_{k-2} + Bu_{k-2} + w_{k-2}) + Bu_{k-1} + w_{k-1} \\ &= \cdots \\ &= A^k x_0 + \sum_{i=0}^{k-1} A^i Bu_i + \sum_{i=0}^{k-1} A^{k-1-i} w_i \end{aligned}$$

Then,

$$x_k - \mathbb{E}[x_k|H_k] = A^k \left(x_0 - \mathbb{E}[x_0|H_k]\right) - \sum_{i=0}^{k-1} A^{k-1-i} \left(w_i - \mathbb{E}[w_i|H_k]\right),$$

which is independent of the control sequence, $\{u_1, \ldots, u_{k-1}\}$. □

We use Lemma 2 to prove the separation principle.

Proof of the Separation principle. For $k_N = Q$ and $P_N = 0$, we write the cost-to-go function as the mean cost plus the estimation variance (which does not depend on the controls)

$$J_N(H_N) = \mathbb{E}[x_N^T Q x_N|H_N] + \mathbb{E}[e_N^T P_N e_N|H_N]$$

where $e_N = x_N - \mathbb{E}[x_N|H_N]$. For stage $N-1$,

$$J_{N-1}(H_{N-1}) = \min_u \ l(H_{N-1}, u)$$

where

$$\begin{aligned} l(H_{N-1}, u) = u^T R u &+ \mathbb{E}[x_{N-1}^T Q x_{N-1}|H_{N-1}] + \mathbb{E}[e_N^T P_N e_N|H_{N-1}, u_{N-1} = u] \\ &+ \mathbb{E}[(Ax_{N-1} + Bu_{N-1} + w_{N-1})^T K_N (Ax_{N-1} + Bu_{N-1} + w_{N-1})|H_{N-1}, u_{N-1} = u] \end{aligned}$$

i.e. the expected accumulated cost-to-go at stage $N-1$ conditional on history H_{N-1} and action u is the instantaneous cost plus the cost-to-go. The cost-to-go at next stage is the sum of expectation of measurement error (not affected by action) and expectation of quadratic cost of next state, given by the linear dynamics, conditional on the history. Differentiating with respect to u we get

$$\mu^\star(H_{N-1}) = L_{N-1}\mathbb{E}[x_{N-1}|H_{N-1}]$$

where

$$L_{N-1} = -(R + B^T k_N B)^{-1} B^T k_N A.$$

Plugging the linear policy back into the quadratic function leads to

$$\begin{aligned} l(H_{N-1}, L_{N-1}\mathbb{E}[x_{N-1}|H_{N-1}]) = \mathbb{E}[w_{N-1}^T Q w_{N-1}] &+ \mathbb{E}[e_N^T P_N e_N|H_{N-1}] + \mathbb{E}[x_{N-1}^T (Q + A^T k_N A) x_{N-1}^T|H_{N-1}] \\ &- \mathbb{E}[x_{N-1}|H_{N-1}]^T P_{N-1} \mathbb{E}[x_{N-1}|H_{N-1}] \end{aligned}$$

where $P_{N-1} = A^T K_N B(R + B^T K_N B)^{-1} B K_N A$. Notice that we can write the last term as[1]

$$\mathbb{E}[x_{N-1}|H_{N-1}]^T P_{N-1} \mathbb{E}[x_{N-1}|H_{N-1}] = \mathbb{E}[x_{N-1}^T P_{N-1} x_{N-1}|H_{N-1}] - \mathbb{E}[e_{N-1}^T P_{N-1} e_{N-1}|H_{N-1}]$$

Plugging this back, we have

$$J_{N-1}(H_{N-1}) = \mathbb{E}[X_{N-1}^T K_{N-1} X_{N-1}|H_{N-1}] + \mathbb{E}[e_{N-1}^T P_{N-1} e_{N-1}|H_{N-1}] + \mathbb{E}[e_N^T P_N e_N|H_{N-1}] + C_{N-1}$$

Thus the cost-to-go function is equal to a quadratic function of state taking expectation over state, plus a bunch of terms that is not affected by the control decision. □

[1] This is a generalization of the familiar expression Variance$(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, presented here in the rearranged form $(\mathbb{E}[X])^2 = \mathbb{E}[X^2] - $ Variance(X).

3 Infinite horizon discounted objectives

We formulate a special case of our problem assume that the cost function only depends on the time through a discounted factor that encodes time preference and the state transition dynamics are independent of time. That is, for a given $\alpha \in (0, 1)$,

$$g_k(x_k, u_k, w_k) = \alpha^k g(x_k, u_k, w_k) \quad .k = 0, 1, \cdots$$

and

$$x_{k+1} = f(x_k, u_k, w_k) \qquad k = 0, 1, \cdots$$

where $\{w_k\}$ are i.i.d. random variables. Note that we had previously assumed only that the disturbances were independent. The additional assumption that they are identically distributed ensures that the dynamics are stationary across time. This section considers the finite horizon objective

$$\inf_{\pi} \mathbb{E}^{\pi} \left[\sum_{k=0}^{N-1} \alpha^k g(x_k, u_k, w_k) + \alpha^N J(X_N) \right] \tag{1}$$

as well as the infinite horizon objective

$$\inf_{\pi} \limsup_{N \to \infty} \mathbb{E}^{\pi} \left[\sum_{k=0}^{N-1} \alpha^k g(x_k, u_k, w_k) \right]. \tag{2}$$

When cost functions are uniformly bounded, we expect these two objectives to behave similarly when $N \gg \frac{1}{1-\alpha}$ as the contribution of times beyond that point becomes negligible.

The principal benefit of taking the horizon to infinity is that it makes temporal tradeoffs stationary, leading to much more elegant solutions. In the final periods of the finite horizon problem, the decision-maker is only worried about the costs in the few remaining periods. They are much more focused on near-term performance than in early periods. For infinite horizon objectives, the discount factor encodes a preference for near-term performance, but this preference does not strengthen as the problem continues. As a result, we will see optimal policies for infinite horizon discounted problems are stationary, in the sense that the decision in period k is a function of the state but not of k itself.

3.1 Technical assumptions

We make the following assumptions. Of these, the third is the most important,

- We assume the state space to be finite or countable.
  - *This assumption sidesteps subtle measure theoretic issues that arise in dynamic programming problems with general state space. For many specific models, like linear quadratic control, these issues clearly do not arise. But when developing our generic theory, our state spaces are defined over the rationals rather than the reals.*
- We assume the control space to be finite.
  - *This is used only to ensure that all minima are attained. Even when minima are not attained, most of these arguments carry through in terms of infima (i.e. sequences of policies whose performances converges to the infimum in* (2)*).*
- We assume the cost functions to be uniformly bounded, i.e. $\sup_{(x,u,w)} |g(x, u, w)| \le M < \infty$.
  - *Typically, cost functions are written as some function $g(x_k, u_k, x_{k+1})$ of the state, action and next state. In this case, the above assumption is satisfied when the state space is finite or when the state space is compact and g is continuous. For problems where the assumption is violated, the proofs given below will not work because they are based on the max-norm of cost-to-go functions, which would be infinite. Arguments are then based on other weighted max-norms.*

3.2 Rewriting the DP algorithm for finite N

As discussed in Homework 1, rather than write the DP algorithm in the form:

$$\begin{aligned} J_N(x) &= \alpha^N J(x) \\ J_{N-k}(x) &= \min_u \ \mathbb{E}\left[\alpha^{N-k} g(x,u,w) + J_{N-k+1}(f(x,u,w))\right] \end{aligned}$$

We can re-normalize to write $\tilde{J_n}^\star(x) = \frac{J^\star_{N-n}(x)}{\alpha^{N-n}}$. The DP algorithm then becomes,

$$\begin{aligned} \tilde{J}_0(x) &= J(x) \\ \tilde{J}_k(x) &= \min_{u \in U(x)} \ \mathbb{E}\left[g(x,u,w) + \alpha \tilde{J}_{k-1}(f(x,u,w))\right] \end{aligned}$$

where k in the subscript denotes the number of periods to go. For the rest of this course, we will use this re-normalized notation.

3.3 Bellman operators

To formalize the operation on function J_{k-1} that produces a new function J_k, we introduce the bellman operators. For bounded $J : S \to \mathbb{R}$, we define $(TJ) : S \to \mathbb{R}$ by:

$$TJ(x) = \min_{u \in U(x)} \ \mathbb{E}\left[g(x,u,w) + \alpha J(f(x,u,w))\right] \qquad \forall x$$

For the N period problem in (1), we can write DP algorithm for policy evaluation concisely in terms of the Bellman operator:

$$\begin{aligned} &J_0^\star = J \\ &J_1^\star = TJ \\ &\cdots \\ &J_N^\star = TJ^\star_{N-1} = \cdots = T^N J \end{aligned}$$

Bellman operator for a stationary policy

For a fixed policy $\mu : S \to \cup_x U(x)$ where $\mu(x) \in U(x)$, we define:

$$(T_\mu J)(x) = \mathbb{E}\left[g(x,\mu(x),w) + \alpha J(f(x,\mu(x),w))\right]$$

Suppose our goal is to evaluate the expected cost-to-go under the policy μ:

$$J_n^\mu(x) = \mathbb{E}^\mu\left[\sum_{k=n}^{N-1} \alpha^k g(x_k, \mu(x_k), w_k) + \alpha^N J(x_N) | x_n = x\right]$$

For a problem with N periods, we can write DP algorithm for policy evaluation concisely:

$$\begin{aligned} &J_0^\mu = J \\ &J_1^\mu = T_\mu J \\ &\cdots \\ &J_N^\mu = T_\mu J^\mu_{N-1} = \cdots = T_\mu^N J \end{aligned}$$

Greedy policies

When we apply bellman operator on cost-to-go function J, we take the minimum. We now define the set of policies that attain those minimum. We call them the greedy policies of J: $G(J) = \{\mu | T_\mu J = TJ\}$, i.e.

$$\mu(x) \in \underset{u \in U(x)}{\operatorname{argmin}} \ \mathbb{E}[g(x,u,w) + \alpha J(f(x,u,w))] \ \forall x$$

.

3.4 Main Theorem for Discounted Dynamic Programming

1. For every stationary policy μ, there exists associated cost function $J_\mu : S \to \mathbb{R}$ that solves the fixed point equation :

   $$J_\mu = T_\mu J_\mu$$

   and

   $$||T_\mu^N J - J_\mu||_\infty \leq \alpha^N ||J - J_\mu||_\infty \quad \forall J.$$

2. There exists optimal cost function $J^\star$ that satisfies:

   $$J^\star = TJ^\star$$

   and

   $$||T^N J - J^\star||_\infty \leq \alpha^N ||J - J^\star||_\infty \quad \forall J.$$

3. If $\mu \in G(J^\star)$, then $J_\mu(x) = J^\star(x)$ for all x. Moreover, under any non-stationary policy Π,

   $$\lim_{N \to \infty} \ \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g_k(x_k, u_k, w_k) | x_0 = x\right] \geq J^\star(x)$$

The proof of this theorem will be presented next class.

Interpretations:

1. The first part of the theorem states that for a given policy, its cost-to-go function is the unique fixed point of the Bellman operator T_μ. Recalling the definition of T_μ, this can be interpreted as a temporal consistency condition: the expected cost to go must equal the expected instantaneous cost plus the expected cost-to-go from the next state. Here the $T_\mu^N J$ is the cost-to-go function for an N period problem and by the second part of the theorem identifies J_μ as its infinite horizon limit:

   $$T_\mu^N J = \mathbb{E}^\mu\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k), w_k) + \alpha^N J(x_N)\right]$$

   $$J_\mu = \liminf_{N \to \infty} T_\mu^N J = \lim_{N \to \infty} \mathbb{E}^\mu\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k), w_k) + \alpha^N J(x_N)\right]$$

2. The second part of the algorithm shows that the optimal cost function $T^N J$ for a N period problem converges at a geometric rate to an infinite horizon limit:

   $$J^\star = \lim_{N \to \infty} T^N J$$

   Moreover, the optimal cost-to-go function J^* is the unique solution to the Bellman fixed point equation $J^* = TJ^*$.

3. The third part shows that a stationary policy μ is optimal when it attains the minimum in the Bellman equation.