ME 7247: Advanced Control Systems **Supplementary notes**

The principle of optimality

v1.0 (10.26.2022)

1 Problem setting and statement

For these notes, we will denote random vectors using boldface (e.g., $\boldsymbol{x}$), and particular values of the corresponding random vector using non-boldface (e.g., x).

Suppose we have a sequence of random variables $\{\boldsymbol{x}_t\}$, $t = 0, \ldots, N$. At each timestep t, we obtain information $\boldsymbol{i}_t$, and we must take an action $\boldsymbol{u}_t$ based on this information. In general, we will allow for *random* actions, that is, instead of picking a fixed u_t for each possible i_t, we will specify a probability density over actions, which we call a *policy*. This is a function of the form

$$K_t(u_t, i_t) = \mathbf{Prob}(\boldsymbol{u}_t = u_t \mid \boldsymbol{i}_t = i_t)$$

This general formulation includes *deterministic* policies as a special case, which would just be functions of the form $u_t = k_t(i_t)$. Our only assumption is that the information is *nested*, which we write informally as $\boldsymbol{i}_0 \subseteq \boldsymbol{i}_1 \subseteq \cdots \subseteq \boldsymbol{i}_N$. Therefore, as time goes on, we do not forget past information. Our task is to minimize a cost of the form:

$$J^\star(i_0) = \underset{K_0,\ldots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[\sum_{k=0}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_0 = i_0\right]$$

Where the g_t functions are known. Define the *value function* or *optimal cost-to-go* as follows.

$$V_t(i_t) := \underset{K_t,\ldots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[\sum_{k=t}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_t = i_t\right] \tag{1}$$

The optimal cost of interest is $J^\star(i_0) = V_0(i_0)$. The *principle of optimality* states that we can compute the V_t functions recursively, optimizing over one action at a time.

Theorem 1 (principle of optimality)**.** *The value function* (1) *satisfies the following recursion, which iterates backward in time starting from $t = N$:*

$$\begin{aligned} V_N(i_N) &= \mathbf{E}\left[g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_N = i_N\right] \\ V_t(i_t) &= \min_u \; \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + V_{t+1}(\boldsymbol{i}_{t+1}) \;\middle|\; \boldsymbol{i}_t = i_t, \boldsymbol{u}_t = u\right] \quad \textit{for } t = 0, \ldots, N-1 \end{aligned}$$

Moreover, there exists a deterministic policy that achieves the optimal cost. This policy is given by picking u_t as the $\arg\min$ *of the minimization for each t.*

The typical case of interest is when each $\boldsymbol{x}_{t+1}$ only depends on $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$. If $\boldsymbol{i}_t = \boldsymbol{x}_t$ (we observe the state at each timestep), this is called a *Markov Decision Process* (MDP). If instead, we observe

some measurement $\boldsymbol{y}_t$ of the state, then we call it a *Partially Observed Markov Decision Process*, (POMDP). However, the principle of optimality is far more general, and holds even when the states and actions do not satisfy a Markovian structure.

2 Proof of Theorem 1

Recall the value function defined in Eq. (1). Now define a recursive version as

$$W_N(i_N) := \mathbf{E}\left[g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_N = i_N\right] \tag{2a}$$

$$W_t(i_t) := \min_u \; \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \;\middle|\; \boldsymbol{i}_t = i_t, \boldsymbol{u}_t = u\right] \quad \text{for } t = 0, \ldots, N-1 \tag{2b}$$

Our goal is to prove that $W_t(i_t) = V_t(i_t)$ for all t and all i_t. We will proceed by induction. First, note that when $t = N$, there is no minimization at all in Eq. (1) and we immediately obtain $W_N(i_N) = V_N(i_N)$ for all i_N. Now, suppose that $W_k(i_k) = V_k(i_k)$ for $k = t+1$. We will prove that it holds for $t = k$ as well. Consider any fixed set of policies $\{K_t, \ldots, K_{N-1}\}$ and compute the cost-to-go for this set:

$$\begin{aligned}
V_t^{K_{t:N-1}}(i_t) &= \mathbf{E}\left[\sum_{k=t}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \mathbf{E}\left[\sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_{t+1}\right] \;\middle|\; \boldsymbol{i}_t = i_t\right] && \text{(3a)} \\
&= \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + V_{t+1}^{K_{t+1:N-1}}(\boldsymbol{i}_{t+1}) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&\geq \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + V_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{i}_t = i_t\right] && \text{(3b)} \\
&= \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{i}_t = i_t\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{u}_t, \boldsymbol{i}_t = i_t\right] \;\middle|\; \boldsymbol{i}_t = i_t\right] && \text{(3c)} \\
&= \sum_u \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{u}_t = u, \boldsymbol{i}_t = i_t\right] \mathbf{Prob}(\boldsymbol{u}_t = u \mid \boldsymbol{i}_t = i_t) \\
&= \sum_u \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{u}_t = u, \boldsymbol{i}_t = i_t\right] K_t(u, i_t) \\
&\geq \min_u \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{u}_t = u, \boldsymbol{i}_t = i_t\right] && \text{(3d)} \\
&= W_t(i_t)
\end{aligned}$$

In Eqs. (3a) and (3c), we used the *tower rule*[1], which relies on the fact that $\boldsymbol{i}_t \subseteq \boldsymbol{i}_{t+1}$. So, for any fixed $K_t, \ldots, K_{N-1}$, we have $W_t(i_t) \leq V_t^{K_{t:N-1}}(i_t)$. We can make this an equality by considering the inequalities (3b) and (3d) separately.

- For Eq. (3b), we have equality if we pick $K_{t+1}, \ldots, K_{N-1}$ to be the optimal policies for the cost-to-go from $t+1$. That is, $K_{t+1:N-1} = \arg\min_{K_{t+1:N-1}} V_{t+1}^{K_{t+1:N-1}}(i_{t+1})$. By the definition of V_{t+1}, this will produce $V_{t+1}^{K_{t+1:N-1}}(i_{t+1}) = V_{t+1}(i_{t+1})$.

[1] We have $\mathbf{E}\big[\,\mathbf{E}[\,\boldsymbol{x} \mid \boldsymbol{y}\,]\,\big] = \mathbf{E}[\,\boldsymbol{x}\,]$. More generally, whenever $\boldsymbol{y} \subseteq \boldsymbol{z}$, we have $\mathbf{E}\big[\,\mathbf{E}[\,\boldsymbol{x} \mid \boldsymbol{z}\,] \;\big|\; \boldsymbol{y}\,\big] = \mathbf{E}[\,\boldsymbol{x} \mid \boldsymbol{y}\,]$.

- For Eq. (3d), since $K_t(u, i_t)$ is a pdf over the space of actions u, we achieve equality by picking the deterministic policy $u_t = k_t(i_t) = \arg\min_u \mathbf{E}[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + W_{t+1}(\boldsymbol{i}_{t+1}) \mid \boldsymbol{u}_t = u, \boldsymbol{i}_t = i_t]$.

Therefore, $W_t(i_t) \le V_t^{K_{t:N-1}}(i_t)$, and there is a particular choice of $K_{t:N-1}$ that achieves equality. Consequently, $V_t(i_t) := \min_{K_{t:N-1}} V_t^{K_{t:N-1}}(i_t) = W_t(i_t)$. We also found a deterministic K_t, so if we use this choice for all t, the entire optimal policy will be deterministic. ■

3 Incorrect proof

The proof of this principle of optimality is more complicated than for the deterministic version. Let's see why the deterministic approach does not work here. It is tempting to start with the definition of V_t and try to split it up:

$$\begin{aligned}
V_t(i_t) &= \underset{K_t,\dots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[\sum_{k=t}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \underset{K_t,\dots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \underset{K_t,\dots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \mathbf{E}\left[\sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_{t+1}\right] \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \min_{K_t} \underset{K_{t+1},\dots,K_{N-1}}{\text{minimize}} \quad \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \mathbf{E}\left[\sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_{t+1}\right] \;\middle|\; \boldsymbol{i}_t = i_t\right]
\end{aligned}$$

Since the g_t term only depends on K_t and the inner expectation only depends on $K_{t+1:N-1}$, we would like to split the minimizations and write

$$\begin{aligned}
V_t(i_t) &= \min_{K_t} \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + \underset{K_{t+1},\dots,K_{N-1}}{\text{minimize}} \mathbf{E}\left[\sum_{k=t+1}^{N-1} g_k(\boldsymbol{x}_k, \boldsymbol{u}_k) + g_N(\boldsymbol{x}_N) \;\middle|\; \boldsymbol{i}_{t+1}\right] \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \min_{K_t} \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + V_{t+1}(\boldsymbol{i}_{t+1}) \;\middle|\; \boldsymbol{i}_t = i_t\right] \\
&= \min_{u} \mathbf{E}\left[g_t(\boldsymbol{x}_t, \boldsymbol{u}_t) + V_{t+1}(\boldsymbol{i}_{t+1}) \;\middle|\; \boldsymbol{i}_t = i_t, \boldsymbol{u}_t = u\right]
\end{aligned}$$

Unfortunately, this does not work. This is because in order to bring the minimization inside, we need to swap the order of the expectation and the minimization. We're effectively claiming that

$$\min_u \mathbf{E}\left[f(\boldsymbol{x}, u)\right] = \mathbf{E}\left[\min_u f(\boldsymbol{x}, u)\right].$$

But this is not true in general! For example, consider $f(x, u) = (x-u)^2$ with $\boldsymbol{x} \sim \mathcal{N}(0, 1)$. Then it's easy to check that $\min_u \mathbf{E}[f(\boldsymbol{x}, u)] = 1$ but $\mathbf{E}\left[\min_u f(\boldsymbol{x}, u)\right] = 0$. In fact, in general, we have

$$\begin{aligned}
& f(\boldsymbol{x}, u) \ge \min_u f(\boldsymbol{x}, u) && \text{(definition of minimum)} \\
\implies \quad & \mathbf{E}[f(\boldsymbol{x}, u)] \ge \mathbf{E}\left[\min_u f(\boldsymbol{x}, u)\right] && \text{(take expectation of both sides)} \\
\implies \quad & \min_u \mathbf{E}[f(\boldsymbol{x}, u)] \ge \mathbf{E}\left[\min_u f(\boldsymbol{x}, u)\right] && \text{(minimize both sides with respect to } u\text{)}
\end{aligned}$$

The reason the principle of optimality still holds, even though this proof approach is flawed, is because the formulation of the problem gives us great flexibility in picking policies. So although $W_t(i_t)^K \leq V_t^K(i_t)$ for any particular choice of policies K, there is a way to choose K so that we have equality.