# 12 Finite Horizon Optimization & Imperfect Information

## Stochastic Optimization with Perfect Information

- We consider a general stochastic system with perfect information, i.e.,

$$X_{t+1} = f_t(X_t, U_t, W_t), \quad Y_t = X_t$$

  We assume the basic random variables $X_0, W_0, W_1, \dots, W_t, \dots$ are independent.

- We consider a control law $g := (g_0, g_1, \dots, g_t, \dots)$ with perfect record, i.e.,

$$U_t = g_t(X_{0;t}, U_{0,t-1}) = \tilde{g}_t(X_{0:t})$$

- Given a control law $g$, its cost is defined by

$$J^g = E^g\left(\sum_{t=0}^{T-1} c_t(X_t, U_t) + c_T(X_T)\right)$$

  where the $c_t(X_t, U_t)$ is called the **immediate cost** and $c_T(X_T)$ is the **terminal cost**.

- We denote $\mathcal{X}$ is the state space, $\mathcal{U}$ is the action space and $\mathcal{G}$ is the set of all control laws. In particular, we denote $\mathcal{G}_M$ the set of all Markov policies. We say a control law $g^* \in \mathcal{G}$ is **optimal** if

$$J^{g^*} = J^* := \inf\{J^g \mid g \in \mathcal{G}\},$$

  where $J^*$ is called the **minimum (expected) cost**. Our objective is to find such $g^* \in \mathcal{G}$ to attain the minimum.

## The Cost-To-Go Function

- Suppose that we restrict our attention is Markov policies $\mathcal{G}_M$, i.e., $U_t = g_t(X_t), \forall t$, then we can define the **cost-to-go** at time $t$ as

$$V_t(x) = E^g\left(\sum_{l=t}^{T-1} c_l(X_l, U_l) + c_T(X_T) \mid X_t = x\right)$$

  Similar to our previous discussion for computing finite cost, we know that the cost-to-go function can be computed resuively by

$$\begin{cases} V_T(x) = c_T(x) \\ V_t(x) = c_t(x, g_t(x)) + E_{W_t}\left(V_{t+1}(f_t(x, g_t(x), W_t))\right) \end{cases},$$

- Note that for a Markov policy, the above defined cost-to-go is a actually a function of state $x$. For an arbitrary policy $g$, which may not be Markov, the **cost-to-go** at time $t$ due to $g$ is defined by

$$J_t^g = E^g\left(\sum_{l=t}^{T-1} c_l(X_l, U_l) + c_T(X_T) \mid X_t^g, X_{t-1}^g \dots, X_1^g, X_0\right)$$

## Dynamic Programming Algorithm

- We present the following main result for stochastic dynamic programming without proof. It tells us it is sufficient to consider the Markov policies for the purpose of optimization and the optimization problem can be solved stage-by-stage in a backwards manner.

**Theorem: Optimality Condition for Perfect Information**

Define the following functions recursively

$$\begin{cases} V_T(x) := c_T(x), \forall x \in \mathcal{X} \\ V_t(x) := \inf_{u \in \mathcal{U}} \{c_t(x,u) + E_{W_t}\left[V_{t+1}(f_t(x,u,W_t))\right]\}, \forall x \in \mathcal{X}, t = 0,1,\ldots,T-1 \end{cases}$$

Then we have the following results

1. Let $g \in \mathcal{G}$ be arbitrary policy. Then we have

$$V(X_t^g) \leq J_t^g \text{ w.p.1 } \forall t = 0,1,\ldots,T$$

   In particular, we have $E_{X_0}\left[V_0(X_0)\right] \leq J^g$.

2. A Markov policy $g \in \mathcal{G}_M$ is optimal if and only if the infimum is achieved at $g_t(x), \forall x \in \mathcal{X}, \forall t = 0,1,\ldots,T$

- The above theorem suggests the following **dynamic programming algorithm** for solving the stochastic optimization problem. First, we compute functions $V_T(x)$. By solving the optimization problem for $t = T-1$, we obtain $V_{T-1}(x)$ and policy $g_{T-1}^*$. We then proceed to obtain $g_{T-2}^*, \ldots, g_1^*, g_0^*$ and $V_{T-2}(x), \ldots, V_1(x), V_0(x)$. Furthermore, we have $J^{g^*} = E(V_0(X_0))$.

## Stochastic Optimization with Imperfect Information

- Now we move to the case of imperfect information, i.e., we consider the following model

$$X_{t+1} = f_t(X_t, U_t, W_t), \quad Y_t = h_t(X_t, V_t)$$

  where $X_0, W_0, W_1, \ldots, V_0, V_1, \ldots$ are mutually independent with given PDF.

- We still assume that the controller has perfect recall. Then the information available to the controller at time $t$ is

$$I_t = (Y_{0:t}, U_{0:t-1}) = (Y_0, Y_1, \ldots, Y_t, U_0, U_1, \ldots, U_{t-1})$$

  Then a control law under imperfect information $g = (g_0, g_1, \ldots, g_T)$ is of form $U_t = g_t(I_t)$.

- We still consider immediate cost $c_t(X_t, U_t)$ and terminal cost $c_T(X_T)$. Our objective is still to find $g \in \mathcal{G}$ that minimizes

$$J^g = E^g\left(\sum_{t=0}^{T-1} c_t(X_t, U_t) + c_T(X_T)\right)$$

## Information States

- For the perfect observation case, we use dynamic programming for actual state $X_t$. For the imperfect observation setting, we do not know $X_t$ due to the observation noise and the output mapping. The basic idea is to use the notion of **information states** or **beliefs**.

**Definition: Information States**

A sequence $Z_0, Z_1, \ldots, Z_{T-1}, Z_T$ is called an **information state** for the above formulated stochastic optimization problem if

(i) $Z_t = l_t(I_t)$; and

(ii) $Z_{t+1} = T_t(Z_t, Y_{t+1}, U_t)$; and

(iii) $E\left(c_t(X_t, U_t) \mid I_t = i_t, U_t = u_t\right) = E\left(c_t(X_t, U_t) \mid Z_t = z_t, U_t = u_t\right)$.

- Intuitively, the above definition says that (i) the information state should be a function of all information available $I_t$; (ii) the information state can be updated recursively using the new information available $Y_{t+1}$ and $U_t$; (iii) the information state should carry the same information as the complete information $I_t$ for the purpose of minimizing the cost.

## Candidate of Information States

- How to choose information state, in particular a minimal one, in general, is not unique, very difficult and problem-independent. For example $I_t$ itself is an information state, but not very interesting.

- Here we introduce a very widely used candidate of information state, which is the probability distribution of each state at time instant $t$, i.e.,

$$\pi_t(x) = P(X_t = x \mid Y_{0:t}, U_{0:t-1}), x \in \mathcal{X}$$

- One can verify the all three conditions of information states hold for $\pi_t(x)$. Clearly, it is a function of $I_t$ and the expected costs given $Z_t$ and $I_t$ are the same. For the recursive update, using formula $P(A, C \mid B) = P(A \mid C, B)P(C \mid B)$, we have

$$\pi_{t+1}(x) = P_{X_{t+1}|Y_{0:t+1},U_{0:t}}(x \mid y_{0:t+1}, u_{0:t}) = \frac{P(x, y_{t+1}, u_t \mid y_{0:t}, u_{0:t-1})}{\sum_{x'} P(x', y_{t+1}, u_t \mid y_{0:t}, u_{0:t-1})} = \frac{N}{D}$$

$$N = \sum_{\hat{x}_t} P(x, y_{t+1}, u_t, \hat{x}_t \mid y_{0:t}, u_{0:t-1}) = \mathbf{1}_{\{g_t(u_{0:t-1}, y_{0:t}) = u_t\}} \sum_{\hat{x}_t} P(y_{t+1} \mid x) P(x \mid \hat{x}_t, u_t) \pi_t(\hat{x}_t)$$

$$\Rightarrow \pi_{t+1}(x) = \frac{\sum_{\hat{x}_t} P(y_{t+1} \mid x) P(x \mid \hat{x}_t, u_t) \pi_t(\hat{x}_t)}{\sum_{x'} \sum_{\hat{x}_t} P(y_{t+1} \mid x') P(x' \mid \hat{x}_t, u_t) \pi_t(\hat{x}_t)}$$

- The above formula of $\pi_t(x)$ has an important implication, i.e., $\pi_t(x)$ *is independent of the control law* $g$. This property will be the basic for using DP for solving the imperfect observation problem

## Implications of Policy Independent of $\pi_t(\cdot)$

- For time instant $t = T-1$, where the last control decision is made, we have

$$E^g\left(c_{T-1}(X_{T-1}, U_{T-1}) + c_T(X_T) \mid Y_{0:T-1}, U_{0:T-2}\right),$$

which depends on $\underbrace{\pi_{T-1}(\cdot) = P(X_{T-1} \mid Y_{0:T-1}, U_{0:T-2})}_{\text{which is independent of } g}$ and $g_{T-1}$

- For time instant $t = T-2$, we have

$$\begin{aligned}&E^g\left(c_{T-2}(X_{T-2}, U_{T-2}) + c_{T-1}(X_{T-1}, U_{T-1}) + c_T(X_T) \mid Y_{0:T-2}, U_{0:T-3}\right)\\ =&E^g\left(c_{T-2}(X_{T-2}, U_{T-2}) + E^g\left(c_{T-1}(X_{T-1}, U_{T-1}) + c_T(X_T) \mid Y_{0:T-1}, U_{0:T-2}\right) \mid Y_{0:T-2}, U_{0:T-3}\right)\end{aligned}$$

which depends on $\underbrace{\pi_{T-2}(\cdot) = P(X_{T-2} \mid Y_{0:T-2}, U_{0:T-3})}_{\text{which is independent of } g}$, $g_{T-1}$ and $g_{T-2}$

- Therefore, we can conclude that the expected cost-to-go given all information we have so far, i.e., $E^g(\sum_{l=t}^{T-1} c_l(X_l, U_l) + C_T(X_T) \mid I_t, U_t)$, only depends on the future part of the policy, i.e., $g_{t+1}, g_{t+2}, \ldots, g_{T+1}$. This suggests the following main results for dynamic programming under imperfect information.

### Theorem: Optimality Condition under Imperfect Information

Define the following functions recursively

$$\begin{cases} V_T(\pi) := E\left(c_T(X_T) \mid \pi_T = \pi\right) \\ V_t(\pi) := \inf_{u \in \mathcal{U}}\{E\left(c_t(X_t, u) + V_{t+1}(T_t(\pi, Y_{t+1}, u)) \mid \pi_t = \pi\right)\}, t = 0, 1, \ldots, T-1 \end{cases}$$

Then we have the following results

1. Let $g \in \mathcal{G}$ be arbitrary policy. Then we have

$$V(\pi) \leq J_t^g \text{ w.p.1 } \forall t = 0, 1, \ldots, T$$

   In particular, we have $J^g \geq E_{X_0}\left[V_0(X_0)\right]$

2. A separated Markov policy $g \in \mathcal{G}_{SM}$ is optimal if and only if the infimum is achieved at $g_t(\pi), \forall \pi, \forall t = 0, 1, \ldots, T$, where $\mathcal{G}_{SM}$ denotes the set of separated control policies of form, $u_t = g_t(\pi_t)$.

- According to the above result, the finite-horizon stochastic optimization problem under imperfect information can be solved as follows:
  - Compute the optimal control decision $u \in \mathcal{U}$ for each information state $\pi$ at each time instant $t = 0, 1, \ldots, T$ by the dynamic programming equations
  - start from the initial information state $\pi_0(x) = P(X_0 = x)$ and choose an optimal control decision $u_0$ and make observation $y_t$
  - update the information state to $\pi_1 = T_t(\pi_0, y_1, u_0)$ and then repeat the above steps until the last time instant $t = T$.

## Example of Perfect Information: Gambler's Strategy

- Let us consider the following gambling process. Initially, you have certain amount of money. At each instant, we can bet an amount up to the money you have: you either loss you bet or win the same amount. The game stops at at time $T$. Your objective is to find a strategy $g$ to maximize $E^g(\ln X_T)$.
- Formally, we have the following model

$$X_{t+1} = X_t + U_t \cdot W_t$$

where $P(W_t = 1) = p$ and $P(W_t = -1) = 1 - p, 0.5 < p < 1$. The cost function is

$$C_T(X_T) = \ln X_t \text{ and } c_t(X_t, U_t) = 0, t = 0, 1, \ldots, T-1$$

- We can write the dynamic programming equations by:

$$\begin{cases} V_T(x) = \ln x \\ V_t(x) = \sup_{0\le u\le x}\{pV_{t+1}(x+u) + (1-p)V_{t+1}(x-u)\}, t = 0, 1, \ldots, T-1 \end{cases}$$

- For the above value functions, we conjecture that

$$V_t(x) = \ln x + A_t$$

It is clearly true for $T$ and we assume the case of $t+1, \ldots, T$. That is, for $l = t+1, \ldots, T$, we can write

$$V_l(x) = \ln x + A_l$$

Then for the case of $t$, we have

$$V_t(x) = \sup_{0\le u\le x}\left\{\underbrace{p\left[\ln(x+u) + A_{t+1}\right] + (1-p)\left[\ln(x-u) + A_{t+1}\right]}_{=:K_t}\right\}$$

Then take the derivative, we have

$$\frac{\partial}{\partial u}K_t = \frac{p}{x+u} - \frac{1-p}{x-u} \Rightarrow u = (2p-1)x$$

Therefore, by taking $u^* = (2p-1)x$, we have

$$\begin{aligned} V_t(x) =& p\left[\ln(x + (2p-1)x) + A_{t+1}\right] + (1-p)\left[\ln(x - (2p-1)x) + A_{t+1}\right] \\ =& \ln x + \underbrace{A_{t+1} + \ln[2p^p(1-p)^{1-p}]}_{A_t} \end{aligned}$$

- By proofing the above conjecture, we actually find the optimal law $g^*(x) = (2p-1)x$, which is not only Markov but also stationary.

## Example of Perfect Information: Linear Quadratic Optimal Control

- Let us consider the following very simple linear dynamic system

$$X_{t+1} = X_t + bU_t + W_t$$

where we assume $\{W_t\}$ is an i.i.d. sequence with zero mean and variance $\sigma^2$

- We consider the following quadratic cost over $T$ stages

$$X_T^2 + \sum_{t=0}^{T} (X_t^2 + rU_t^2)$$

where $r$ is a known non-negative weighting parameter. We assume no constraints on $X_t$ and $U_t$, i.e., both the state space and the decision space are $\mathbb{R}$

- We apply the DP algorithm, and derive the optimal cost-to-go functions $J_t^*$ and optimal policy. We have

$$\begin{aligned}
J_T^*(x_T) =& x_T^2 \\
J_{T-1}^*(x_{T-1}) =& \min_{u_{T-1}} \left\{ E\left( x_{T-1}^2 + ru_{T-1}^2 + J_T^*(x_{T-1} + bu_{T-1} + W_{T-1}) \right) \right\} \\
=& \min_{u_{T-1}} \left\{ E\left( x_{T-1}^2 + ru_{T-1}^2 + (x_{T-1} + bu_{T-1})^2 + 2W_{T-1}(x_{T-1} + bu_{T-1}) + W_{T-1}^2 \right) \right\} \\
=& x_{T-1}^2 + \sigma^2 + \min_{u_{T-1}} \left\{ ru_{T-1}^2 + (x_{T-1} + bu_{T-1})^2 \right\}
\end{aligned}$$

Therefore, the optimal policy for the last stage is

$$g_{T-1}^*(X_{T-1}) = -\frac{b}{r + b^2} X_{T-1}$$

and the optimal cost-to-go function is

$$J_{T-1}^*(X_{T-1}) = P_{T-1} X_{T-1}^2 + \sigma^2, \text{ where } P_{T-1} = \frac{r}{r + b^2} + 1$$

- We can now continue the DP algorithm to obtain $J_{T-2}^*$ from $J_{T-1}^*$. An important observation is that $J_{T-1}^*$ is quadratic (plus an inconsequential constant term), so with a similar calculation we can derive $g_{T-2}^*$ and $J_{T-2}^*$ in closed form, as a linear and a quadratic function of $X_{T-2}$, respectively. This gives us the following equations

$$\begin{aligned}
g_t^*(X_t) =& -\frac{bP_{t+1}}{r + b^2 P_{t+1}} X_t \\
J_t^*(X_t) =& P_t X_t^2 + \sigma^2 \sum_{l=t}^{T-1} P_{t+1}
\end{aligned}$$

where $P_t$ can be computed backwards recursively by

$$P_t = \frac{rP_{t+1}}{r + b^2 P_{t+1}} + 1 \quad \text{with} \quad P_T = 1$$

## Example of Imperfect Information: Sequential Hypothesis

- We consider the following sequential hypothesis problem. Suppose that a parameter $H$ is either 0 or 1 with prior information that $P(H = 0) = p$. At each instant, we can decide either to make an observation or make a guess. Each observation $Y_t = H + N_t$ is assumed to be the true value of $H$ adding a noise and we need to pay a constant cost $c$ for each observation. Once we make a guess, we stop the process and if we guess the value incorrectly, we pay a cost $K$. You have to make a guess up to instant $T$. The objective is to obtain an optimal sequential hypothesis strategy with minimum cost.

- Formally, we have the following model

$$H_{t+1} = H_t, \qquad Y_t = H_0 + N_t, \quad \text{with } H_0 \in \{0, 1\} \text{ and } P(H_0 = 0) = p$$

and $\mathcal{U} = \{0, 1, \text{continue}\}$ and $c_t(H_t, U_t) = \begin{cases} \mathbf{1}_{\{U_t \neq H_0\}} \cdot K & \text{if } U_t = 0, 1 \\ c & \text{if } U_t = \text{cont.} \end{cases}$.

- Since there are only two possible states, we can choose our information state as:

$$\pi_t = P(H = 0 \mid Y_{0:t}, U_{0:t-1} = (\text{cont.}, \dots, \text{cont.}))$$

This information state can be updated by

$$\pi_{t+1} = P(H{=}0 \mid Y_{0:t+1}) = \frac{P(Y_{t+1}, H{=}0 \mid Y_{0:t})}{P(Y_{t+1} \mid Y_{0:t})} = \frac{P(Y_{t+1} \mid H{=}0)\pi_t}{P(Y_{t+1} \mid H{=}0)\pi_t + P(Y_{t+1} \mid H{=}1)(1 - \pi_t)}$$

- The dynamic programming equation is as follows:

$$V_T(\pi) = \min\{\underbrace{K(1-\pi)}_{U_T=0}, \underbrace{K\pi}_{U_T=1}\}$$

$$V_t(\pi) = \min\{\underbrace{K(1-\pi)}_{U_t=0}, \underbrace{K\pi}_{U_t=1}, \underbrace{c + E(V_{t+1}(\pi_{t+1}) \mid \pi)}_{U_t=\text{cont.}}\}$$

- To see the above solution intuitively, let us consider the following special cases:
  - $T = 0$: since we need to make a guess immediately, we simple guess $H = 0$ if $p \geq 0.5$.
  - $c = 0$: since there is no observation cost, we will continue until the last instant $T$ and guess $H = 0$ if $\pi \geq 0.5$.
  - $K = 0$: since there is no guessing cost, we just randomly make a guess initially.
  - $N_t = 0$: since we can get precise value of $H$ by making an observation, we know that $E(V_1(\pi_1) \mid p) = 0$. Therefore, if $c \leq \min\{K(1-p), Kp\}$, then we make one observation and use the observed value to guess. Otherwise, we directly guess $H$ with higher probability.