5 Controlled Markov processes with partial observations

Recall the setting of a controlled Markov process with partial observations. Consider the following system dynamics:

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

where X_0, $(W_t)_{t\geq 0}$, $(V_t)_{t\geq 0}$ are mutually independent primitive random variables. We assume that states, actions and observations are all finite, i.e., $\mathcal{X} = \{1, ..., n\}$, $\mathcal{Y} = \{1, ..., p\}$, $\mathcal{U} = \{1, ..., m\}$. We also define the following objects pertaining to the (controlled) state transitions and (noisy) observations:

$$P_u^{(t)}(x, x') := \mathbf{P}[X_t = x' | X_t = x, U_t = u]$$
$$M^{(t)}(x, y) := \mathbf{P}(Y_t = y | X_t = x),$$

Given a finite time horizon T, we have the stepwise cost functions $c_t : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, $t = 0, \ldots, T-1$, and the terminal cost, $c_T : \mathcal{X} \to \mathbb{R}$. A T-step policy $g = (g_t)_{t=0}^{T-1}$ is a collection of mappings $g_t : \mathcal{Y}_0^t \times \mathcal{U}_0^{t-1} \to \mathcal{U}$, i.e., the action at time t is a function of all information available at time t:

$$u_t = g_t(y_0^t, u_0^{t-1}).$$

The expected cost of a policy g is given by

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

and the goal of the optimal control is to find the g that minimizes $J(g)$.

Consider the space of all probability distributions,

$$\Delta_n = \{\pi = (\pi(1), ..., \pi(n)) : \pi(i) \geq 0, \sum_{i=1}^{n} \pi(i) = 1\},$$

Instead of $(\mathcal{X}, \mathcal{Y}, \mathcal{U})$, we work with $(\Delta_n, \mathcal{U})$. The belief state at time t is the posterior distribution of state X_t given all the observations up to time t, y_0^t and actions up to time $t-1$, u_0^{t-1}.

$$\pi_t(x) := \mathbf{P}[X_t = x | Y_0^t = y_0^t, U_0^{t-1} = u_0^{t-1}],$$

which is independent of the policy. In the previous lecture, we have shown that the belief state evolves according to the deterministic update

$$\pi_{t+1} = \mathsf{F}_t(\pi_t, u_t, y_{t+1}),$$

where F_t is a composition of a prediction step P_t and a correction step C_t:

$$\pi_t \xrightarrow[\mathsf{P}_t]{\text{prediction}} \pi_{t+1|t} \xrightarrow[\mathsf{C}_t]{\text{correction}} \pi_{t+1},$$

where $\pi_{t+1|t} := \mathbf{P}[X_{t+1} = x|y_0^t, u_0^t]$ is the predictive distribution of the state X_t. In the previous lecture, we have shown that the prediction step takes the form

$$\pi_{t+1|t}(x) = \sum_{x'} \pi_t(x') P_{u_t}^{(t)}(x', x),$$

while the correction step is

$$\pi_{t+1}(x) = \frac{\pi_{t+1|t}(x) M^{(t)}(x, y_{t+1})}{\pi_{t+1|t} M^{(t)}(y_{t+1})}.$$

We now come to the following crucial result:

Proposition 5.1 *$\{\pi_t\}_{t\geq 0}$ is a controlled Markov process: For any measurable $A \subseteq \Delta_n$ and any policy g, we have*

$$\begin{aligned}\mathbf{P}^g[\pi_{t+1} \in A|\pi_0^t, u_0^t] &= \mathbf{P}[\pi_{t+1} \in A|\pi_t, u_t] \\ &= \sum_{y\in\mathsf{Y}} \mathbf{1}_{\{\mathsf{F}_t(\pi_t,u_t,y)\in A\}} \pi_t P_{u_t}^{(t)} M^{(t+1)}(y), \end{aligned} \tag{1}$$

where the right-hand side is independent of policy.

Proof: Fix a policy g. Then, using the definition of $\pi_{t+1|t}$, for any t and $y \in \mathsf{Y}$ we have

$$\begin{aligned}\mathbf{P}^g[Y_{t+1} = y|\pi_0^t, u_0^t] &= \sum_x \mathbf{P}^g[X_{t+1} = x, Y_{t+1} = y|\pi_0^t, u_0^t] \\ &= \sum_x \mathbf{P}[Y_{t+1} = y|X_{t+1} = x]\pi_{t+1|t}(x) \\ &= \sum_x M^{(t+1)}(x, y)\pi_t P_{u_t}^{(t)}(x) \\ &= \pi_t P_{u_t}^{(t)} M^{(t+1)}(y) \end{aligned} \tag{2}$$

It follows then that

$$\begin{aligned}\mathbf{P}^g[\pi_{t+1} \in A, Y_{t+1} = y|\pi_0^t, u_0^t] &= \mathbf{1}_{\{\mathsf{F}_t(\pi_t,u_t,y)\in A\}} \mathbf{P}[Y_{t+1} = y|\pi_0^t, u_0^t] \\ &= \mathbf{1}_{\{\mathsf{F}_t(\pi_t,u_t,y)\in A\}} \pi_t P_{u_t}^{(t)} M^{(t+1)}(y)\end{aligned}$$

where we have used Eq. (2) in the last step. Marginalizing over Y_{t+1} gives

$$\mathbf{P}^g[\pi_{t+1} \in A|\pi_0^t, u_0^t] = \sum_{y\in\mathsf{Y}} \mathbf{1}_{\{\mathsf{F}_t(\pi_t,u_t,y)\in A\}} \pi_t P_{u_t}^{(t)} M^{(t+1)}(y),$$

where the right-hand side depends only on π_t and u_t. Therefore, $\{\pi_t\}_{t\geq 0}$ is a controlled Markov process. ■

The formula (1) for the controlled transition probabilities of $\{\pi_t\}$ looks formidable. Its main use is in the following *canonical computation step*:

Let $V : \Delta_n \to R$ be given. The goal is to compute the conditional expectation $\mathbf{E}[V(\pi_{t+1})|\pi_t, u_t]$ (note that this conditional expectation is policy-independent, as it should be). Using (1), we can write

$$\mathbf{E}[V(\pi_{t+1})|\pi_t, u_t] = \sum_{y \in \mathcal{Y}} V(\mathsf{F}_t(\pi_t, u_t, y))\pi_t P_{u_t}^{(t)} M^{(t+1)}(y).$$

6 Optimal policies for POMDPs

We are now ready to present the dynamic programming recursion for POMDPs. First, recall that we can replace the any state-action cost function $c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ with a belief-action cost function

$$\tilde{c} : \Delta_n \times \mathcal{U} \to \mathbb{R}, \qquad \tilde{c}(\pi, u) := \sum_{x \in \mathcal{X}} \pi(x) c(x, u).$$

As shown in the previous lecture, for any policy $g = (g_t)_{t \geq 0}$, we have

$$\begin{aligned} J(g) &= \mathbf{E}^g \left[\sum_{t=0}^{T} c_t(X_t, U_t) + c_T(X_T) \right] \\ &= \mathbf{E}^g \left[\sum_{t=0}^{T-1} \tilde{c}_t(\pi_t, U_t) + \tilde{c}_T(\pi_T) \right]. \end{aligned}$$

Thus, we can consider belief-based policies $g = (g_t)_{t \geq 0}$,

$$\tilde{g} = (\tilde{g}_t)_{t \geq 0}, \qquad u_t = \tilde{g}_t(\pi_0^t, u_0^{t-1}),$$

and belief-based Markov policies:

$$\tilde{g} = (\tilde{g}_t)_{t \geq 0}, \qquad u_t = \tilde{g}_t(\pi_t)$$

The optimality of Markov policies of the above form follows directly from the same arguments that were used for the case of complete observations, and the optimal policy g^* can be found using DP.

As before, the value functions $V_t : \Delta_n \to \mathbb{R}$, $t = 0, 1, \ldots, T$, are constructed via the DP recursion:

- For $t = T$,

$$V_T(\pi) = \tilde{c}_T(\pi) = \sum_{x \in \mathcal{X}} c_T(x)\pi(x)$$

- For $t = T-1, T-2, ..., 0$,

$$\begin{aligned} Q_t(\pi, u) &:= \tilde{c}_t(\pi, u) + \mathbf{E}[V_{t+1}(\pi_{t+1})|\pi_t = \pi, u_t = u] \\ V_t(\pi) &:= \min_{u \in \mathcal{U}} Q_t(\pi, u) \\ \tilde{g}_t^* &= \arg\min_{u \in \mathcal{U}} Q_t(\pi, u) \end{aligned}$$

θ_t	Y_t	U_t	θ_{t+1}	Y_{t+1}
1	*	1	1	$W_t(1)$
1	*	2	1	$W_t(2)$
2	*	1	2	$W_t(2)$
2	*	2	2	$W_t(1)$

Table 1: Table of evolution of state $X_t = (\theta_t, Y_t)$ over time.

where the conditional expectation can be written out using (1) as

$$\mathbf{E}[V_{t+1}(\pi_{t+1})|\pi_t = \pi, u_t = u] = \sum_{y \in \mathsf{Y}} V(\mathsf{F}_t(\pi, u, y))\pi P_u^{(t)} M^{(t+1)}(y).$$

7 Two-armed bandit problem

We now give an example of a stochastic control problem with partial observations, the *two-armed bandit problem*. For now, we will derive the solution from first principles, primarily to illustrate the ideas from the preceding sections; later on, we will see how the results below follow from a more general theory of multi-armed bandit problems.

Suppose that you have access to two slot machines. You know that one has higher chances of winning among the two, but you don't know which one. In each time step, you can play (or pull the lever for) either of the slot machines and observe the reward. You are allowed to play a maximum of T times. Our goal is to maximize the reward over time and eventually play the better slot machine.

We assume there are two arms $\{1, 2\}$. Let the optimal arm be the same over time: we denote this by $\theta_t = \ldots = \theta_0 \in \{1, 2\}$. We are allowed to pull one of the two arms in a time step – this is our action. Let U_t denote the action taken at time t. Now the observation at time t is the reward returned by the arm U_t. Let Y_t be the stochastic process which denotes the observation.

Let us take an example to better understand the process Y_t. If the optimal arm is 1, then pulling arm 1 returns higher reward on average that pulling arm 2. Suppose the observation space is $\{1, 2\}$ (This has nothing to do with the arm number $1, 2$). Each arm returns an observation which is a sample from a distribution over the observation space. If the optimal arm is 1, then arm 1 is likely to return observation 2 with higher probability than observation 1. The sub optimal arm 2 is going to return more number of observation 1 than observation 2 on average. To characterize this mathematically, consider two stochastic processes, $\{W_t(1), W_t(2)\}_{t \geq 0}$. $W_t(1)$ generates observation 2 with higher probability than $W_t(2)$. So the optimal arm generates observation from the distribution of $W_t(1)$ and the worse arm generates distribution from process $W_t(2)$. More formally,

$$\mathbf{P}[W_t(i) = 2] = 1 - \mathbf{P}[W_t(i)] = p_i, \quad i = \{1, 2\} \text{ and } p_1 > p_2$$

The state space X_t is (θ_t, Y_t). From Table 1 we can see that the state X_{t+1} evolves as some function of state X_t action U_t and process W_t. Let

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

We are allowed to pull the arms for a finite amount of time T. The cost function that we use to characterize the rewards is,

$$c(X,U) = c((\theta, y), u) = \begin{cases} -\alpha & \text{if } \theta = u \\ \alpha & \text{if } \theta \neq u \end{cases} \text{ (Terminal cost is 0).}$$

Let the probability vector $\tilde{\pi}_t(\theta)$ denote our guess for the better arm. We are going to look at the evolution of $\tilde{\pi}_t(\theta)$ and how we can update it with new observations over time.

Theorem 7.1 *Let $\tilde{\pi}_t(\theta) = \mathbf{P}[\theta_t = \theta | Y_0^t, U_0^{t-1}]$. Then for any t, the optimal policy,*

$$g_t^*(\tilde{\pi}_t) = \begin{cases} 1 & if \;\; \tilde{\pi}_t(1) > \tilde{\pi}_t(2) \\ 2, & otherwise \end{cases}.$$

Before proving this theorem we will do the following tasks:

- Write the evolution of $\tilde{\pi}_t$.
- Write down the value functions.

Let us familiarize ourselves with the parameters of the Markov model for X_t.

- State transition probabilities for X_t:

$$P_u((\theta, y), (\theta', y')) = \mathbb{1}_{\{\theta \neq \theta'\}} \cdot [p_1(y')\mathbb{1}_{u=\theta} + p_2(y')\mathbb{1}_{u \neq \theta}],$$

$$\text{where } p_1(y) = \mathbf{P}[W_t(1) = y] = \begin{cases} 1 - p_1, & y = 1 \\ p_1, & y = 2 \end{cases}$$

$$p_2(y) = \mathbf{P}[W_t(2) = y] = \begin{cases} 1 - p_2, & y = 1 \\ p_2, & y = 2 \end{cases}$$

- Observation model:

$$M((\theta, y), y') = \mathbb{1}\{y \neq y'\}.$$

Now, $\tilde{\pi}_t(\theta)$ is the probability at time t with which we can say that arm θ is optimal given observations up to time t and all our actions up to time $t-1$. We can write $\tilde{\pi}_{t+1}(\theta) = \sum_y \pi_{t+1}(\theta, y)$ where

$$\pi_t(\theta, y) = \mathbf{P}[(\theta_t, Y_t) = (\theta, y) | Y_0^t, U_0^{t-1}]$$

is the belief state. We first need to get the evolution of $\pi_t(\theta, y)$. Define

$$\pi_{t+1|t}(\theta, y) := \mathbf{P}[(\theta_{t+1}, Y_{t+1}) = (\theta, y) | Y_0^t, U_0^t].$$

Computing $\pi_{t+1|t}$ will be useful in computing π_t. So,

$$\begin{aligned}
\pi_{t+1|t}(\theta, y) &= \sum_{\theta', y'} P_{u_t}((\theta', y'), (\theta, y))\pi_t(\theta', y') \\
&= \sum_{y'} P_{u_t}((\theta, y'), (\theta, y))\pi_t(\theta, y') \qquad (P_{u_t}\text{is only non-zero when } \theta = \theta') \\
&= \sum_{y'} (p_1(y)\mathbb{1}_{\{u_t=\theta\}} + p_2(y)\mathbb{1}_{\{u_t\neq\theta\}})\pi_t(\theta, y') \\
&= \tilde{\pi}_t(\theta)[p_1(y)\mathbb{1}_{\{u_t=\theta\}} + p_2(y)\mathbb{1}_{\{u_t\neq\theta\}}] \qquad (3)
\end{aligned}$$

Now we compute π_{t+1} as a function of $\pi_{t+1|t}$ using Bayes rule:

$$\begin{aligned}
\pi_{t+1}(\theta, y) &= \frac{\pi_{t+1|t}(\theta, y)M((\theta, y), Y_{t+1})}{\sum_{\theta', y'} \pi_{t+1|t}(\theta', y')M((\theta', y'), Y_{t+1})} \\
&= \frac{\tilde{\pi}_t(\theta)[p_1(Y_{t+1})\mathbb{1}_{\{u_t=\theta\}} + p_2(Y_{t+1})\mathbb{1}_{\{u_t\neq\theta\}}]\mathbb{1}_{\{Y_{t+1}=y\}}}{\tilde{\pi}_t(u_t)p_1(Y_{t+1}) + \tilde{\pi}_t(\bar{u_t})p_2(Y_{t+1})}
\end{aligned}$$

where $\bar{u}$ is given by

$$\bar{u} = 3 - u = \begin{cases} 2, & \text{if } u = 1 \\ 1, & \text{if } u = 2 \end{cases} .$$

Finally we can compute $\tilde{\pi}_{t+1}$ as

$$\begin{aligned}
\tilde{\pi}_{t+1}(\theta) &= \sum_{y} \tilde{\pi}_{t+1}(\theta, y) \\
&= \frac{\tilde{\pi}_t(\theta)[p_1(Y_{t+1})\mathbb{1}_{\{u_t=\theta\}} + p_2(Y_{t+1})\mathbb{1}_{\{u_t\neq\theta\}}]}{\tilde{\pi}_t(u_t)p_1(Y_{t+1}) + \tilde{\pi}_t(\bar{u_t})p_2(Y_{t+1})}(\textbf{??}) \qquad (4)
\end{aligned}$$

Instead of keeping track of $\tilde{\pi}_{t+1}(\theta) = \{\tilde{\pi}_{t+1}(1), \tilde{\pi}_{t+1}(2)\}$ which has two quantities to update, we can update a single quantity $h_t \in [-1, 1]$ if we decompose $\tilde{\pi}_{t+1}(\theta)$ as

$$\tilde{\pi}_{t+1} = (\tilde{\pi}_{t+1}(1), \tilde{\pi}_{t+1}(2)) = \left(\frac{1 - h_{t+1}}{2}, \frac{1 + h_{t+1}}{2}\right)$$

Using this and (**??**), we can derive an update rule for $\{h_t\}$ in the form $h_{t+1} = \mathsf{F}(h_t, U_t, Y_t)$ for a suitable function F. This update rule is summarized in Table 2.

We call h_t as the belief process. Next we are going to characterize some evolution properties of the value function which is a function of the belief process.

u	y	$\mathsf{F}(h,u,y)$	
1	1	$\frac{R+Q_1h}{Q_1+Rh}$	
1	2	$\frac{-R+Q_2h}{Q_2-Rh}$	$R := \frac{p_1-p_2}{2} > 0$
2	1	$\frac{-R+Q_1h}{Q_1-Rh}$	$Q_y := \frac{p_1(y)+p_2(y)}{2}$
2	2	$\frac{R+Q_2h}{Q_2+Rh}$	

Table 2: Table of evolution of the belief process H_t over time.

7.1 Belief process for two-armed bandit

The belief process is $\{H_t\}_{t\geq 0}$, where H_t takes values in $[-1,1]$. For any function $V:[-1,1]\to\mathbb{R}$, we can use Table 2 to compute $\mathbf{E}[V(H_{t+1})|H_t=h,U_t=u]$ as follows:

$$\mathbf{E}[V(H_{t+1})|H_t=h,U_t=1] = (Q_1+Rh)V\left(\frac{R+Q_1h}{Q_1+Rh}\right)+(Q_2-Rh)V\left(\frac{-R+Q_2h}{Q_2-Rh}\right),$$
$$\mathbf{E}[V(H_{t+1})|H_t=h,U_t=2] = (Q_1-Rh)V\left(\frac{-R+Q_1h}{Q_1-Rh}\right)+(Q_2+Rh)V\left(\frac{R+Q_2h}{Q_2+Rh}\right),$$

The cost function $c(x,u)$ defined for the bandit problem gets transformed to the analogous cost function $\tilde{c}(h,u)$ for the belief process:

$$\tilde{c}(h,u) = \begin{cases} h\alpha, & u=1 \\ -h\alpha & ,u=2 \end{cases}, \qquad \tilde{c}(h,u) = \frac{1-h}{2}c(1,u)+\frac{1+h}{2}c(2,u) \tag{5}$$

Now we define the dynamic programming recursion for V through operators $\mathbb{T}_1$ and $\mathbb{T}_2$.

$$\mathbb{T}_1V(h) := \tilde{c}(h,1)+\mathbf{E}[V(H_1)|H_0=h,U_0=1]$$
$$\mathbb{T}_2V(h) := \tilde{c}(h,2)+\mathbf{E}[V(H_1)|H_0=h,U_0=2]$$
$$\mathbb{T}V(h) = \min\{\mathbb{T}_1V(h),\mathbb{T}_2V(h)\}$$

With the initial condition of $V_T(h)=\tilde{c}_T(h)=0$, the value functions $\{V_t\}_{t\geq 0}^T$ are computed as $V_t=\mathbb{T}V_{t+1}$. Thus, $V_0=\mathbb{T}^TV_T=\mathbb{T}^T0$ and, more generally, $V_t=\mathbb{T}^t0$, i.e., the value function V_t is obtained by the t-fold application of the operator $\mathbb{T}$ to the zero function $h\mapsto 0$.

Next we prove a property about the difference of the operators $\mathbb{T}_1$ and $\mathbb{T}_2$:

Theorem 7.2 *Let us define*

$$V_t(h) = \begin{cases} \mathbb{T}_1V_{t+1}(h), & h<0 \\ \mathbb{T}_2V_{t+1}(h), & h\geq 0 \end{cases}.$$

and

$$\Delta_t(h) := \mathbb{T}_1V_{t+1}(h)-\mathbb{T}_2V_{t+1}(h).$$

Then h and $\Delta_t(h)$ have the same sign, i.e.

$$h\Delta_t(h) \geq 0.$$

Proof: We will use the following two facts in the proof:

- Fact A: $\mathbb{T}_1\mathbb{T}_2$ operator is same as $\mathbb{T}_2\mathbb{T}_1$: for any function $V : [-1, 1] \to \mathbb{R}$ and any $h \in [-1, 1]$, $\mathbb{T}_1\mathbb{T}_2V(h) = \mathbb{T}_2\mathbb{T}_1V(h)$.
- Fact B: If V is non-decreasing, then $\mathbb{T}_1V$ is also non-decreasing.

Facts A and B will be established in the homework.

We use backward induction to the result.

- Base case: From (5) we can see that for $t = T - 1$, $\mathbb{T}_1V_T(h) = \alpha h, \mathbb{T}_2V_T(h) = -\alpha h$. So $\Delta_T(h) = 2\alpha h$ and $h\Delta_T(h) = 2\alpha h^2 \geq 0$. Also, $\Delta_T(h)$ is non decreasing and $\Delta_T(0) = 0$.
- Assume that the following facts are true for time $t + 1$.
  - $\Delta_{t+1}(h)$ is non-decreasing.
  - $\Delta_{t+1}(0) = 0$.

  We will prove that the above conditions are also true for time t. Note that this implies that $h\Delta_t(h) \geq 0$. Consider,

$$\begin{aligned}
\Delta_t(h) &= \mathbb{T}_1(V_{t+1}(h)) - \mathbb{T}_2(V_{t+1}(h)) \\
&= \mathbb{T}_1\mathbb{T}V_{t+2}(h) - \mathbb{T}_2\mathbb{T}V_{t+2}(h) \\
&= \mathbb{T}_1\mathbb{T}V_{t+2}(h) - \mathbb{T}_1\mathbb{T}_2V_{t+2}(h) + \mathbb{T}_2\mathbb{T}_1V_{t+2}(h) - \mathbb{T}_2\mathbb{T}V_{t+2}(h) \text{ from FACT A} \\
&= \mathbb{T}_1(\mathbb{T}V_{t+2}(h) - \mathbb{T}_2V_{t+2}(h)) + \mathbb{T}_2(\mathbb{T}_1V_{t+2}(h) - \mathbb{T}V_{t+2}(h)) \qquad (6)
\end{aligned}$$

  Now define $\Phi(h) := \mathbb{T}V_{t+2}(h) - \mathbb{T}_2V_{t+2}(h)$ and $\Psi(h) := \mathbb{T}_1V_{t+2}(h) - \mathbb{T}V_{t+2}(h)$ where $\mathbb{T}V_{t+2}(h) = \mathbb{T}_1V_{t+2}(h)$ if $h < 0$ or $\mathbb{T}_2V_{t+2}(h)$ if $h \geq 0$. Hence,

$$\Phi(h) = \begin{cases} \Delta_{t+1}(h), h < 0 \\ 0 \;\; h \geq 0 \end{cases}, \Psi(h) = -\Phi(-h)$$

  From (6),

$$\begin{aligned}
\Delta_t(h) &= \mathbb{T}_1\Phi(h) + \mathbb{T}_2\Psi(h) \\
&= \mathbb{T}_1\Phi(h) - \mathbb{T}_1\Phi(-h) \\
&= \begin{cases} \mathbb{T}_1\Delta_{t+1}(h) \text{ if } h < 0 \\ -\mathbb{T}_1\Delta_{t+1}(-h) \text{ if } h \geq 0 \end{cases}
\end{aligned}$$

  Since $\Delta_{t+1}(h)$ non decreasing by the inductive assumption, $\mathbb{T}_1\Delta_{t+1}(h)$ is also non decreasing (using FACT B). As $-\mathbb{T}_1\Delta_{t+1}(-h)$ is the mirror of $\mathbb{T}_1\Delta_{t+1}(h)$ at $h = 0$ followed by mirroring over $h - axis$, $-\mathbb{T}_1\Delta_{t+1}(-h)$ is also non decreasing. For $h = 0$, $\Delta_t(0) = \mathbb{T}_10 - T_20 = 0$. Hence we arrive at a proof for time t.

■