# Discrete Dynamic Programming*

Peifan Wu

October 8, 2020

## 1 Notations and Definitions

A discrete dynamic programming is basically a maximization problem with an objective function of the form

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t r\left(s_t, a_t\right) \tag{1}$$

where

- $s_t$ is the **state variable**
- $a_t$ is the **action**, and $s_{t+1} = g\left(s_t, a_t\right)$.
- $\beta$ is a discount factor, $\beta < 1$
- $r\left(s_t, a_t\right)$ is the current reward function when we have state $s_t$ and action $a_t$.

Each pair $\left(s_t, a_t\right)$ pins down the **transition probabilities** $Q\left(s_t, a_t, s_{t+1}\right)$ for the next period state $s_{t+1}$.

Therefore, $a_t$ influences not only the current rewards, but also the future states.

The essence of dynamic programming: *a trade-off between current rewards and favorable future states.* Macroeconomics is, overall, a trade-off in time.

## 2 Formal Definition

Formally, a discrete dynamic program consists of the following components:

1. A finite set of states $S = \{0, \ldots, n-1\}$
2. A finite set of feasible actions $A\left(s\right)$ for each state $s \in S$, and a corresponding set of feasible state-action pairs
$$SA \equiv \left\{\left(s, a\right) | s \in S, a \in A\left(s\right)\right\}$$
3. A reward function $r : SA \to \mathbb{R}$
4. A transition probability function $Q : SA \to \Delta\left(S\right)$ where $\Delta\left(S\right)$ is the set of probability distributions over $S$

*Prepared for UBC ECON 407 Fall 2020

5. A discount factor $0 \leqslant \beta < 1$

We use the notation $A \equiv \bigcup_{s \in S} A(s) = \{0, \ldots, m-1\}$ and call this set the action space.

A policy is a function $\sigma : S \to A$.

A policy is called feasible if it satisfies $\sigma(s) \in A(s)$ for all $s \in S$. Denote the set of all feasible policies by $\Sigma$. Therefore for policy $\sigma \in \Sigma$,

- the current reward at time $t$ is $r(s_t, \sigma(s_t))$
- the probability that $s_{t+1} = s'$ is $Q(s_t, \sigma(s_t), s')$

For each $\sigma \in \Sigma$, define

- $r_\sigma(s) \equiv r(s, \sigma(s))$
- $Q_\sigma(s, s') \equiv Q(s, \sigma(s), s')$

# 3 Value, Policy, and Optimality

Let $v_\sigma(s)$ denote the discounted sum of expected reward flows from policy $\sigma$ when the initial state is $s$. Therefore,

$$v_\sigma(s) = \sum_{t=0}^{\infty} \beta^t \left(Q_\sigma^t r_\sigma\right)(s)$$

The **value function** is the function $V^* : S \to \mathbb{R}$,

$$V^*(s) = \max_{\sigma \in \Sigma} v_\sigma(s)$$

i.e. it's the expected maximum of all different action choices.

This value function $V^*$ is the unique solution to the **Bellman equation,**

$$V^*(s) = \max_{a \in A(s)} \left\{ r(s, a) + \beta \sum_{s' \in S} V^*\left(s'\right) Q\left(s, a, s'\right) \right\} \tag{2}$$

Intuitively, we rewrite Equation (1),

$$\begin{aligned}
\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t r(s_t, a_t) &= r(s_0, a_0) + \mathbb{E}_0 \sum_{t=1}^{\infty} \beta^t r(s_t, a_t) \\
&= r(s_0, a_0) + \mathbb{E}_0 \beta \left( \mathbb{E}_1 \sum_{t=0}^{\infty} \beta^t r(s_t, a_t) \right)
\end{aligned}$$

and you can see the similarity of the structures. We replace the structure by $V^*$.

This means the **policy function** will be

$$\sigma^*(s) \in \arg\max_{a \in A(s)} \left\{ r(s, a) + \beta \sum_{s' \in S} V^*\left(s'\right) Q\left(s, \sigma(s), s'\right) \right\}$$

In a word: value function characterized the maximum (discounted) sum, and the policy function records the action to achieve the corresponding value.