2. The dynamic optimality equation

(also known as the Bellman equation
or the dynamic programming equation)

2.1 Consider a controllable dynamical system

$$f : \mathbb{Z}^+ \times S \times A \longrightarrow S.$$

Recall that a control $u : \mathbb{Z}^+ \longrightarrow A$ determines a controlled sequence $(x_n)_{n \geq k}$ starting from (k, x) by

$$x_k = x, \qquad x_{n+1} = f(n, x_n, u_n), \quad n \geq k.$$

Introduce a <u>cost function</u>

$$c : \mathbb{Z}^+ \times S \times A \longrightarrow \mathbb{R}^+ \quad (\text{or } \mathbb{R}^-)$$

We interpret $c(n, x, a)$ as the cost incurred when we choose action a in state x at time n.

Define

$$V^u(k, x) = \sum_{n=k}^{\infty} c(n, z_n, u_n),$$

$$V(k, x) = \inf_u V^u(k, x) \qquad \underline{\text{infimal cost function}}.$$

The case $c \leq 0$ allows us to cover the maximization of rewards, by taking $r = -c$.

The values $u_0, \dots, u_{k-1}$ are irrelevant to the controlled process $(x_n)_{n \geq k}$ starting from (k, x).

Suppose we take $u_k = a$, then $x_{k+1} = f(k, x, a)$ and

$$V^u(k,x) = c(k,x,a) + V^u(k+1, f(k,x,a)).$$

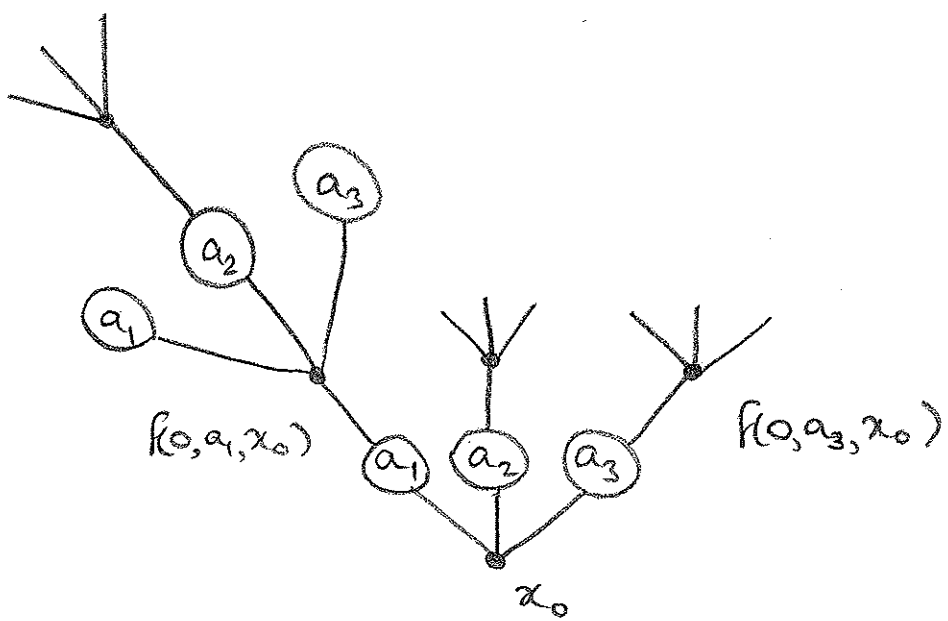
On taking the infimum over $\tilde{u} = (u_{k+1}, u_{k+2}, \dots)$, we get

$$\inf_{\tilde{u}} V^u(k,x) = c(k,x,a) + V(k+1, f(k,x,a))$$

Then, taking the infimum over a, we obtain the <u>dynamic optimality equation</u>

$$V(k,x) = \inf_a \{ c(k,x,a) + V(k+1, f(k,x,a)) \}.$$

Each control specifies a path through the tree of all possible choices. The optimality equation expresses the



fact that we can optimize over these paths iteratively

$$\inf_{u} = \inf_{a} \inf_{\tilde{u}}$$

2.2 Consider now a stochastic controllable dynamical system

$$P : \mathbb{Z}^+ \times S \times A \longrightarrow \mathrm{Prob}(S),$$

A <u>control</u> is a map

$$u : S^* \longrightarrow A$$

where

$$S^* = \bigcup_{k \geq 0} S_k^*, \quad S_k^* = \bigcup_{x \in S} S_{(k,x)}^*.$$

$$S_{(k,x)}^* = \{(n, x, x_{k+1}, \ldots, x_n) : n \geq k,\ x_{k+1}, \ldots, x_n \in S\}$$

We think of u as a set of instructions : if we start at x at time k, and if we then move through states $x_{k+1}, \ldots, x_n$, then we take action $u_n(x, x_{k+1}, \ldots, x_n)$ at time n.

- We use the values of u on $S^*_{(k,x)}$ when we start at time k and state x.

- We met already

  Markov controls — no dependence on $x_k, \ldots, x_{n-1}$

  Stationary Markov controls — no dependence on n either

Let $(X_n)_{n \geq k}$ be a random process with values in S.

Say that $(X_n)_{n \geq k}$ is a (P,u)-<u>controlled process</u> starting from (k,x) if, for all $n \geq k$ and all $x_k, x_{k+1}, \dots, x_n \in S$,

$$\mathbb{P}(X_k = x_k, X_{k+1} = x_{k+1}, \dots, X_n = x_n)$$

$$= \delta_{x x_k} P(k, x_k, u_k(x_k))_{x_{k+1}} P(k+1, x_{k+1}, u_{k+1}(x_k, x_{k+1}))_{x_{k+2}} \cdots P(n-1, x_{n-1}, u_{n-1}(x_k, \dots, x_{n-1}))_{x_n}.$$

Equivalently, $\mathbb{P}(X_k = x) = 1$ and, for all $n \geq k$ and all $x_k, \dots, x_{n+1} \in S$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_k = x_k, \dots, X_n = x_n) = P(n, x_n, u_n(x_k, \dots, x_n))_{x_{n+1}}.$$

We often write $\mathbb{P}^u_{(k,x)}$ in place of $\mathbb{P}$ to indicate the dependence of the distribution of $(X_n)_{n \geq k}$ on (k,x) and u.

The action taken at time n is

$$U_n = u_n(X_k, \ldots, X_n).$$

This is a random variable. The value of U_n is determined by the past values of the process, which is reasonable, and does not require knowledge of the future, which would be unreasonable.

Define

$$V^u(k,x) = \mathbb{E}^u_{(k,x)} \sum_{n=k}^{\infty} c(n, X_n, U_n),$$

$$V(k,x) = \inf_u V^u(k,x) \qquad \underline{\text{infimal cost function}}$$

Say that u is <u>optimal</u> <u>for (k,x)</u> if $V^u(k,x) = V(k,x)$

In the stochastic case, we use different pieces of u for different choices of (k,x), so these optimizations can be done independently.

Conditioning on the first step

Given a control u and a state x, define a new control $u^{(x)}$ by

$$u^{(x)}_n(x_{k+1}, \dots, x_n) = u_n(x, x_{k+1}, \dots, x_n), \quad n \geq k+1.$$

If $(X_n)_{n \geq k}$ is a (P, u)-controlled process starting from (k, x), then, conditional on $X_{k+1} = y$, $(X_n)_{n \geq k+1}$ is a $(P, u^{(x)})$-controlled process starting from $(k+1, y)$. To see this, compute

$$\begin{aligned}
&\mathbb{P}^u_{(k,x)}(X_{k+1} = x_{k+1}, \dots, X_n = x_n \mid X_{k+1} = y) \\
&= \delta_{y x_{k+1}} P(k+1, x_{k+1}, u^{(x)}_{k+1}(x_{k+1}))_{x_{k+2}} \cdots P(n-1, x_{n-1}, u^{(x)}_{n-1}(x_{k+1}, \dots, x_{n-1}))_{x_n} \\
&= \mathbb{P}^{u^{(x)}}_{(k+1,y)}(X_{k+1} = x_{k+1}, \dots, X_n = x_n).
\end{aligned}$$

Proposition 2.1

The infimal cost function satisfies the dynamic optimality equation

$$V(k,x) = \inf_a (C + PV)(k,x,a), \quad k \in \mathbb{Z}^+, \; x \in S.$$

Proof Fix (k,x) and set $U_n^{(x)} = u_n^{(x)}(X_{k+1}, \ldots, X_n)$. By conditioning on the first step,

$$\mathbb{E}^u_{(k,x)}\left(\sum_{n=k+1}^{\infty} c(n, X_n, U_n) \,\middle|\, X_{k+1} = y \right) = \mathbb{E}^{u^{(x)}}_{(k+1,y)}\left(\sum_{n=k+1}^{\infty} c(n, X_n, U_n^{(x)}) \right) = V^{u^{(x)}}(k+1, y),$$

so

$$V^u(k,x) = c(k,x,a) + \sum_y P(k,x,a)_y \, V^{u^{(x)}}(k+1,y),$$

where $a = u_k(x)$. As we vary u over all controls, so $(a, u^{(x)})$ ranges over all pairs with $a \in A$ and $u^{(x)}$ a control. Hence

$$V(k,x) = \inf_a \inf_{u^{(x)}} V^u(k,x) = \inf_a \left\{ c(k,x,a) + \sum_y P(k,x,a)_y V(k+1,y) \right\}. \quad \square$$