6 Stochastic control over an infinite time horizon

In order to extend the previous control framework from finite time horizons to infinite time horizons, we need to introduce new optimality criteria for control policies. This entails defining modified expected cost functions to be optimized over possible policies.

6.1 Infinite time horizon optimality criteria

First, we review the framework for control of a Markov decision process with full observations having general state and action spaces over a finite time horizon.

6.1.1 Finite time horizon framework

- Let $\mathcal{X}$ and $\mathcal{U}$ represent the (Borel) state and action spaces.
- Let $0 \leq T < \infty$ represent the finite time horizon.
- We assume the system is time invariant and evolves according to dynamics given by

$$X_{t+1} = f(X_t, U_t, W_t), \quad t = 0, 1, \ldots \tag{1}$$

  where the disturbances $(W_t)_{t \geq 0}$ are i.i.d. and independent from X_0.

- Note that the dynamics can be equivalently expressed using a probability kernel $P_u(\mathrm{d}x'|x)$ on $\mathcal{X}$ such that for all $A \subset \mathcal{X}$ the transition probabilities are given by

$$\mathbf{P}(X_{t+1} \in A \mid X_t = x, U_t = u) = P_u(A|x) = \int_A P_u(\mathrm{d}x'|x) = \mathbf{P}(f(x, u, W_0) \in A \mid U_t = u) \tag{2}$$

- For each $t \in \{0, \ldots, T-1\}$, we have single-step costs $c_t : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$.
- For $t = T$, we have the terminal cost $c_T : \mathcal{X} \to \mathbb{R}$.
- The control policy g is a sequence of functions $(g_t)_{t=0}^{T-1}$ with $g_t : \mathcal{X}_0^t \times \mathcal{U}_0^{t-1} \to \mathcal{U}$.
- The expected cost is defined as

$$J_T(g) := \mathbf{E}^g[\sum_{t=0}^{T-1} c_t(X_t, U_t) + c_T(X_T)] \tag{3}$$

For infinite time horizons, we assume that the single-step costs are time-invariant ($c_t(x, u) = c(x, u)$) and that there is no terminal cost. Now we briefly introduce four new infinite horizon optimality criteria.

6.1.2 Discounted cost optimality criterion

The main idea of the discounted cost criterion is to consider the future costs with less importance than the current costs by multiplying them by successively smaller weights for times further into the future, but in a way that future costs are not totally negligible.

Introduce a factor $\beta \in (0,1)$ which is called the *discount factor*. The discounted expected cost is defined by

$$J^{\beta}(g) := \mathbf{E}^{g}[\sum_{t=0}^{\infty} \beta^{t} c(X_t, U_t)] \tag{4}$$

The discounted expected cost depending on the initial state is defined similarly but with the conditional expectation

$$J^{\beta}(x; g) := \mathbf{E}^{g}[\sum_{t=0}^{\infty} \beta^{t} c(X_t, U_t) \mid X_0 = x] \tag{5}$$

6.1.3 Long term average cost optimality criterion

The long run average cost criterion modifies the original expected cost for the finite time horizon framework by dividing the cost by the time horizon, then taking the limit supremum of this quantity as the time horizon goes to infinity. Note that if we attempt to simply take the limit of the finite horizon cost without dividing by the horizon, there is no guarantee that the sum inside the expectation will converge.

Given the initial state, the finite time horizon expected cost is given by

$$J_T(x; g) := \mathbf{E}^{g}[\sum_{t=0}^{T-1} c(X_t, U_t) \mid X_0 = x] \tag{6}$$

and the long run average cost is given by

$$\overline{J}(x; g) := \limsup_{T \to \infty} \frac{J_T(x; g)}{T} \tag{7}$$

As we will see, the solution to both the discounted cost and the long-run average cost criterion will not only be a Markov policy, but it will be *stationary* as well.

Definition 6.1 (Stationary Policy)**.** *A policy $g = (g_t)_{t \geq 0}$ is* : stationary *if g is Markov and $g_t = \overline{g} : \mathcal{X} \to \mathcal{U}$ does not depend on t.*

6.1.4 Optimal control up to exit time optimality criterion

Fix a measurable subset $S \subset \mathcal{X}$ and assume $X_0 = x \in S$. Suppose we want to control the system optimally until it leaves the subset S. The modified cost in this case is nearly identical to the finite horizon cost except the upper bound on the sum inside the expectation is replaced with the time at which X leaves S (minus one time step). This cost is given by

$$J^{S}(x; g) := \mathbf{E}^{g}[\sum_{t=0}^{\tau-1} c(X_t, U_t) \mid X_0 = x] \tag{8}$$

Note that the time τ at which the event that X leaves S occurs, given by

$$\tau := \inf\{t > 0 \mid X_t \notin S\} \tag{9}$$

is a random variable. Furthermore, τ is a stopping time (i.e., for each t, the event $\{\tau \leq t\}$ is determined only by the information available up to time t).

6.1.5 Risk sensitive control optimality criterion

The risk sensitive criterion is structured similarly to the long run average cost criterion except that the risk sensitive cost is defined in such a way to penalize large averages and also large excursions from averages. Define the risk sensitive cost for a finite time horizon T by

$$R_T^\lambda(x; g) := \frac{1}{\lambda} \log(\mathbf{E}^g[e^{\sum_{t=0}^{T-1} c(X_t, U_t)} \mid X_0 = x]) \tag{10}$$

and define the long run average risk sensitive cost by

$$R^\lambda(x; g) := \limsup_{T\to\infty} \frac{R_T^\lambda(x; g)}{T} \tag{11}$$

Now we are going to take a more in-depth look at the discounted cost optimality criterion

7 Discounted cost optimality criterion

Fix a policy $g = (g_t)_{t\geq 0}$ and recall that the discounted expected cost is given by

$$J^\beta(x; g) := \mathbf{E}^g[\sum_{t=0}^{\infty} \beta^t c(X_t, U_t) \mid X_0 = x] \tag{12}$$

Assume that the single-step costs are bounded. That is, $0 \leq c(x, u) \leq M < \infty$ for all $x \in \mathcal{X}, u \in \mathcal{U}$ for some M. Then the discounted expected cost of any policy g is also bounded because

$$0 \leq J^\beta(x; g) \leq M \sum_{t=0}^{\infty} \beta^t = \frac{M}{1-\beta} \tag{13}$$

This sum converges because $\beta \in (0, 1)$, hence it forms a geometric series. We start by looking at the finite horizon discounted costs. Note that $J_T^\beta(x; g)$ is nondecreasing in T because $c(x, u) \geq 0$. Then, since it is a bounded monotone sequence, the following limit exists for all x, g:

$$J^\beta(x; g) = \lim_{T\to\infty} J_T^\beta(x; g) \tag{14}$$

7.0.1 Discounted cost optimality equation (DCOE) and value iteration (VI)

Define the value function as the infimum of the previous limit over all admissible policies:

$$V^\beta(x) := \inf_g J_\beta(x; g) \tag{15}$$

We know that $0 \leq V^\beta(x) \leq \frac{M}{1-\beta}$. We want to prove the following.

- V^β is the unique bounded solution to the discounted cost optimality equation (DCOE), given by
$$V^\beta(x) = \inf_{u \in \mathcal{U}} \{c(x,u) + \beta \int_{\mathcal{X}} V^\beta(x') P_u(\mathrm{d}x'|x)\} \tag{16}$$
- Since the DCOE is a fixed point equation, V^β can be found via value iteration (VI), which is a set of recursions given by
$$V_0^\beta(x) \equiv 0 \ \forall x \in \mathcal{X} \tag{17}$$
$$V_{t+1}^\beta(x) := \inf_{u \in \mathcal{U}} \{c(x,u) + \beta \int_{\mathcal{X}} V_t^\beta(x') P_u(\mathrm{d}x'|x)\}, \quad t \geq 0 \tag{18}$$
$$V^\beta(x) = \lim_{t \to \infty} V_t^\beta(x) \tag{19}$$
We will show that, when c is bounded, $V_t^\beta(x)$ will converge to $V^\beta(x)$ at an exponential rate.
- Under some regularity conditions on $\mathcal{X}$, $\mathcal{U}$, and P_u, the policy
$$g^*(x) := \arg\min_{u \in \mathcal{U}} \{c(x,u) + \beta \int_{\mathcal{X}} V^\beta(x') P_u(\mathrm{d}x'|x)\} \tag{20}$$
will be a measurable function from $\mathcal{X}$ to $\mathcal{U}$ and will give the optimal stationary policy. This obviously holds when $\mathcal{X}$ and $\mathcal{U}$ are finite. We need to guarantee that this minimization can be performed for all initial states x and that g^* can be chosen to be measureable.

Consider a finite time horizon T and let the sequence $(W_t^\beta)_{t=0}^T$ represent the value functions in the usual dynamic programming recursion:

$$W_T(x) \equiv 0 \tag{21}$$
$$W_t(x) = \inf_{u \in \mathcal{U}} \{\beta^t c(x,u) + \int_{\mathcal{X}} W_{t+1}(x') P_u(\mathrm{d}x'|x)\}, \quad t \in \{T-1, T-2, \ldots, 0\} \tag{22}$$

We know that $W_0(x) = \inf_g J_T^\beta(x; g)$. Introduce the discounted value functions

$$V_T(x) := \beta^{t-T} W_{T-t}(x), \quad t \in \{0, \ldots, T\} \tag{23}$$

The following result will be proved in the homework:

Lemma 7.1. *The following recursion holds:*

$$V_{t+1}(x) = \inf_{u\in\mathcal{U}}\{c(x,u) + \beta\int_{\mathcal{X}} V_t(x')P_u(\mathrm{d}x'|x)\}, \quad t \geq 0 \tag{24}$$

To streamline further analysis, we need to introduce the following definition:

Definition 7.1 (Dynamic Programming Operator)**.** *Let $V : \mathcal{X} \to \mathbb{R}$ be a measurable function. The* dynamic programming operator $\mathbb{T}^\beta$ *which maps $V \mapsto \mathbb{T}^\beta V$ is defined according to*

$$[\mathbb{T}^\beta V](x) := \inf_{u\in\mathcal{U}}\{c(x,u) + \beta\int_{\mathcal{X}} V_t^\beta(x')P_u(\mathrm{d}x'|x)\} \tag{25}$$

The previous value iteration is now given by the following equations using simpler notation.

$$V_0^\beta \equiv 0 \tag{26}$$

$$V_{t+1}^\beta := \mathbb{T}^\beta V_t^\beta, \quad t \geq 0 \tag{27}$$

$$V^\beta(x) = \lim_{T\to\infty} V_T^\beta(x;g) = \inf_g J^\beta(x;g) \tag{28}$$

Definition 7.2 (Contraction)**.** *A* contraction *on a metric space (X,d) is a map $T : X \to X$ with the property that there exists a $\beta \in [0,1)$ such that for all $x, y \in X$, we have $d(T(x),T(y)) \leq \beta d(x,y)$.*

Lemma 7.2. *$\mathbb{T}^\beta$ is a contraction on $(\mathcal{V},\|\cdot\|_\infty)$, where $\mathcal{V}$ is the space of bounded measureable functions $V : \mathcal{X} \to \mathbb{R}$ and $\|\cdot\|_\infty$ is the supremum norm given by $\|V\|_\infty := \sup_{x\in\mathcal{X}}|V(x)|$.*

Proof: Fix any $x \in \mathcal{X}$. Then for any $V, V' \in \mathcal{V}$,

$$\mathbb{T}^\beta V(x) - \mathbb{T}^\beta V'(x) \tag{29}$$

$$= \inf_{u\in\mathcal{U}}\{c(x,u) + \beta\int_{\mathcal{X}} V(x')P_u(\mathrm{d}x'|x)\} - \inf_{u\in\mathcal{U}}\{c(x,u) + \beta\int_{\mathcal{X}} V'(x')P_u(\mathrm{d}x'|x)\} \tag{30}$$

$$= \inf_{u\in\mathcal{U}}\sup_{u'\in\mathcal{U}}\{c(x,u) - c(x,u') + \beta\int_{\mathcal{X}} V(x')P_u(\mathrm{d}x'|x) - \beta\int_{\mathcal{X}} V'(x')P'_u(\mathrm{d}x'|x)\} \tag{31}$$

$$\leq \sup_{u\in\mathcal{U}}\{\beta\int_{\mathcal{X}}[V(x') - V'(x')]P_u(\mathrm{d}x'|x)\} \tag{32}$$

$$\leq \beta\sup_{x\in\mathcal{X}}\left|V(x') - V'(x')\right|. \tag{33}$$

Interchanging the roles of V and V' and taking the supremum over all $x \in \mathcal{X}$ proves the claim. ■

Theorem 7.1 (Contraction mapping principle, Banach fixed point theorem)**.** *Let $\mathcal{V}$ be a complete metric space and $\mathbb{T} : \mathcal{V} \to \mathcal{V}$ a contraction, then the following hold.*

(i) $\mathbb{T}$ has a unique fixed point $V^* \in \mathcal{V}$ such that $\mathbb{T}V^* = V^*$.

(ii) For all $V_0 \in \mathcal{V}$, the sequence $(V_t)_{t\geq 0}$ defined according to the recursion $V_t := \mathbb{T}V_{t-1}$ converges to V^*.

(iii) Furthermore, the sequence above converges exponentially in norm to its limit. That is,

$$\|V_t - V^*\| \leq \beta^t \|V_0 - V^*\| \tag{34}$$

By the contraction mapping principle, there exists a unique bounded measurable function $V^* : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{T}^\beta V^* = V^*$ (i.e., V^* solves the DCOE) and $V_T^\beta \to V^*$ exponentially in norm as $T \to \infty$. We need to show that V^* is the desired value function. That is, $V^* \equiv V^\beta$.

Lemma 7.3. *(i) Let $\phi : \mathcal{X} \to \mathcal{U}$ be given, $\phi^\infty := (\phi, \phi, \dots)$. Let $w(x) := J^\beta(x, \phi^\infty)$. Then it satisfies equation $w(x) = c(x, \phi(x)) + \beta P w(x, \phi(x))$, where, for any $x \in \mathcal{X}$, $u \in \mathcal{U}$, and measurable $w : \mathcal{X} \to \mathbb{R}$,*

$$Pw(x, u) := \int_{\mathcal{X}} w(x') P_u(\mathrm{d}x'|x).$$

(ii) Given ϕ^∞, there exists a unique bounded $w : \mathcal{X} \to \mathbb{R}$ such that $w(x) = c(x, \phi(x)) + \beta P w(x, \phi(x))$ and $w(x) = J^\beta(x, \phi^\infty)$.

Proof: Homework. ■

We now recall the regularity assumptions:

Assumption 7.1. *(i) $0 \leq c(x, u) \leq M$ for some $0 < M < \infty$.*

(ii) existence of measurable selectors for any bounded $V : \mathcal{X} \to \mathbb{R}$ i.e.,

$$f(x) := \arg\min_{u \in \mathcal{U}} \left\{ c(x, u) + \beta \int V(x') P_u(\mathrm{d}x'|x) \right\} \quad \textit{exists}$$

and $f(x)$ is a measurable function of x.

Theorem 7.2. *Under above assumption,*

(i) V^β is the unique bounded solution of the discounted cost optimality equation (DCOE)

$$V^*(x) = \min_{u \in \mathcal{U}} \left\{ c(x, u) + \beta \int V^*(x') P_u(\mathrm{d}x'|x) \right\} = \mathbb{T}^\beta V^*(x)$$

(ii) V^* can be found by value iteration, by which we mean $V_{t+1} := \mathbb{T}_\beta V_t$ with $V_0 = 0$: $V_t \leq V_{t+1} \leq \dots$ and $\lim_{t \to \infty} V_t = V^\beta$.

(iii) The optimal policy is stationary: there exists $\phi^* : \mathcal{X} \to \mathcal{U}$ such that

$$V^\beta(x) = c(x, \phi^*(x)) + \beta \int V^\beta(x') P_{\phi^*(x)}(\mathrm{d}x'|x) = \mathbb{T}^\beta V^\beta(x) \quad \forall x$$

Thus, the stationary policy $g^* = (\phi^*, \phi^*, \dots)$ is optimal, and $J^\beta(x, \phi^*) = V^\beta(x)$ for any $x \in \mathcal{X}$.

Proof: For covenience of discussion, we will use the following notation.

$$Ph(x,u) := \int_{\mathcal{X}} h(x')P_u(\mathrm{d}x'|x) = \mathbf{E}[h(X_1) \mid X_0 = x, U_0 = u]$$

Accordingly, $\mathbb{T}^\beta h(x) = \min_{u\in\mathcal{U}} \{c(x,u) + \beta Ph(x,u)\}$ Let $V_0^\beta = 0$, $V_{t+1}^\beta = \mathbb{T}^\beta V_t^\beta$ Since $\mathbb{T}^\beta$ is a contraction, by the contraction mapping principle, there exists a unique bounded $V^* : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{T}^\beta V^* = V^*$, and the recursion $V_t^\beta = \mathbb{T}^\beta V_{t-1}^\beta$ converges to V^*. Thus, it suffices to show $V^* \equiv V^\beta$.

Fix any policy $g = (g_0, g_1, \dots)$. We observe that $J_T^\beta(x;g) = \mathbf{E}^g\left[\sum_{t=0}^{T-1} \beta^t c(X_t, U_t) \mid X_0 = x\right] \geq V_T^\beta(x)$ (proof is left as an exercise). Then

$$J^\beta(x;g) = \mathbf{E}^g\left[\sum_{t=0}^{\infty} \beta^t c(X_t,U_t) \mid X_0 = x\right] \tag{35}$$

$$= \mathbf{E}^g\left[\sum_{t=0}^{T-1} \beta^t c(X_t,U_t) \mid X_0 = x\right] + \mathbf{E}^g\left[\sum_{t=T}^{\infty} \beta^t c(X_t,U_t) \mid X_0 = x\right] \tag{36}$$

$$= J_T^\beta(x;g) + \mathbf{E}^g\left[\sum_{t=T}^{\infty} \beta^t c(X_t,U_t) \mid X_0 = x\right] \tag{37}$$

Therefore, $J_\beta(x;g) \geq J_T^\beta(x;g)$. Taking the infimum over g on both sides, we obtain $V^\beta(x) \geq V_T^\beta(x)$ (Recall that $\inf_g J_T^\beta(x;g) = V_T^\beta(x)$, as you will prove in the homework.) Using similar reasoning, with the fact that $J_T^\beta(x;g) \leq J_{T+1}^\beta(x;g)$, we have $V_t \leq V_{t+1}$ for all t. By sending T to ∞, we get $V^\beta(x) \geq V_T^*(x)$. For the converse,

$$J_\beta(x;g) = J_T^\beta(x;g) + \mathbf{E}^g\left[\sum_{t=T}^{\infty} \beta^t c(X_t,U_t) \mid X_0 = x\right] \tag{38}$$

$$= J_T^\beta(x;g) + \beta^T \mathbf{E}^g\left[\sum_{t=T}^{\infty} \beta^{t-T} c(X_t,U_t) \mid X_0 = x\right] \tag{39}$$

$$\leq J_T^\beta(x;g) + \beta^T M \sum_{t=0}^{\infty} \beta^t \tag{40}$$

Again taking the infimum over g on both sides, we obtain $V^\beta(x) \leq V_T^\beta(x) + \frac{M\beta^T}{1-\beta}$. Taking $T \to \infty$ again gives $V^\beta(x) \leq V_T^*(x)$.

Now we show part (iii) of the theorem. We can extract the optimal action for each state based on the value function to form policy ϕ^*, which is $\phi^* := \arg\min_{u\in\mathcal{U}} \left\{c(x,u) + \beta PV^\beta(x,u)\right\}$ and let $g^*(x) = (\phi^*, \phi^*, \dots)$. Then we can write the DCOE as $V^\beta(x) = c(x,\phi^*(x)) + \beta PV^\beta(x,\phi^*(x))$ for all x by substituting each u with $g^*(x)$. We can see V^β is a solution to equation $w(x) = c(x,\phi^*) + \beta Pw(x,\phi^*)$ and $V^\beta = J^\beta(x;g^*)$. Furthermore, Proposition (7.3) guarantees the uniqueness of such solution.

■

In fact, Assumption (7.1) can be relaxed to more general conditions. We will not prove the result under these more general assumptions, but will simply state them and give a illustrative exmaple.

Assumption 7.2. *(i) $c : \mathcal{X} \to \mathbb{R}$ is nonnegative, continuous and inf-compact, or equivalently, for all $x \in \mathcal{X}$, $a \geq 0$, the set $\{u \in \mathcal{U} : c(x,u) \leq a\}$ is compact. Recall that, in a finite-dimensional Euclidean space, a set is compact iff it is closed and bounded.*

(ii) $P_u(\mathrm{d}x'|x)$ is strongly continuous, or equivalently, for any bounded measurable function $h : \mathcal{X} \to \mathbb{R}$, $Ph(x,u) = \int_{\mathcal{X}} h(x')P_u(\mathrm{d}x'|x)$ is a bounded and continuous function of x and u.

Example 7.1. *Let $\mathcal{X} = \mathbb{R}^n, \mathcal{U} = \mathbb{R}^n$. $X_{t+1} = Ax_t + Bu_t + W_t$, where $W_t \overset{i.i.d}{\sim} \mathcal{N}(0,\Sigma)$, $\det(\Sigma) \neq 0$. $c(x,u) = x^\top Qx + u^\top Ru$, where $Q^\top = Q \succeq 0$, $R^\top = R \succ 0$*

(i) We check inf-compactness of $c(x,u)$: Fix $x \in \mathbb{R}^n$, $a \geq 0$, $c(x,u) = x^\top Qx + u^\top Ru \leq a$. Let $S_{x,a} = \left\{u \in \mathbb{R}^n : u^\top Ru \leq a - x^\top Qx\right\}$. If $x^\top Qx > a$, $S_{x,a} = \varnothing$, which is compact by definition. If $x^\top Qx \leq a$, $S_{x,a} = \left\{u^\top Ru \leq c\right\}$ which is an ellipsoid in $\mathbb{R}^n$, which is a closed and bounded subset of $\mathbb{R}^n$, hence compact.

(ii) We check strong continuity of $P_u(\mathrm{d}x'|x)$. Let h be measurable and bounded.

$$Ph(x,u) = \mathbf{E}_W[h(Ax + Bu + W)] \tag{41}$$

$$= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int_{\mathbb{R}^n} h(x') \exp(-\frac{1}{2}(x' - Ax - Bu)^\top \Sigma^{-1}(x' - Ax - Bu))\,\mathrm{d}x'. \tag{42}$$

Then integral is bounded since $h(x')$ is bounded and $Ph(x,u)$ is continuous since e^{-x} is continuous.

7.1 Policy Iteration

The above discussion focused on finding the discount-optimal value function. We now present an iterative procedure for finding the discount-optimal policy. Assume cost function is nonnegative and bounded for this section. We start with an arbitrary measurable map $\phi_0 : \mathcal{X} \to \mathcal{U}$. Then, for each $t = 0, 1, \ldots$, we carry out the following steps:

Step 1: Define $w_t(x) := J^\beta(x; \phi_t^\infty)$. Then it satisfies $w_t(x) = c(x, \phi_t(x)) + \beta P w_t(x, \phi_t(x))$.

Step 2: Let $\phi_{t+1} : \mathcal{X} \to \mathcal{U}$ be such that, for any x,

$$c(x, \phi_{t+1}(x)) + \beta P w_t(x, \phi_{t+1}(x)) = \mathbb{T}^\beta w_t(x).$$

then go back to Step 1.

Based on the above algorithm, we have the following statement.

Theorem 7.3. *For policy iteration, we have*

(i) for all $x \in \mathcal{X}$, $w_{t+1}(x) \leq w_t(x)$ for all t. Then $w(x) := \lim_{t\to\infty} w_t(x) = V^\beta(x)$ and $\phi^* : \mathcal{X} \to \mathbb{R}$ such that $V^\beta(x) = c(x, \phi^*) + \beta P V^\beta(x, \phi^*)$ is optimal.

(ii) if $w_{t+1} = w_t$ for some t, then $w_t = V^\beta$ and $\phi_{t+1} = \phi^*$

Proof: For part (i), we prove $w_{t+1} \leq w_t$ for all x.

$$
\begin{aligned}
w_t(x) &= c(x, \phi_t(x)) + \beta P w_t(x, \phi_t(x)) && (43)\\
&\geq \min_{u\in\mathcal{U}} \{c(x,u)) + \beta P w_0(x,u)\} && (44)\\
&= \mathbb{T}^\beta w_t(x) && (45)\\
&= c(x, \phi_{t+1}(x)) + \beta P w_t(x, \phi_{t+1}(x)) && (46)\\
&= c(x, \phi_{t+1}(x)) + \beta \mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_1) \mid X_0 = x] && (47)\\
&\geq c(x, \phi_{t+1}(x)) + \beta \mathbf{E}^{\phi_{t+1}^\infty}\Big[c(X_1, \phi_{t+1}(X_1)) + \beta \mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_2) \mid X_1] \mid X_0 = x\Big] && (48)\\
&= J_2^\beta(x; g_t^\infty) + \beta^2 \mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_2) \mid X_0 = x] && (49)\\
&\dots && (50)\\
&\geq J_T^\beta(x; g_t^\infty) + \beta^T \mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_T) \mid X_0 = x] && (51)
\end{aligned}
$$

Since $w_t(x) = J^\beta(x; \phi_t^\infty) < \infty$, $\mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_T) \mid X_0 = x] < \infty$ for all T. Then $\beta^T \mathbf{E}^{\phi_{t+1}^\infty}[w_t(X_T) \mid X_0 = x] \to 0$ as $T \to 0$. Therefore, $w_t(x) \geq \lim_{T\to\infty} J_T^\beta(x; \phi_{t+1}) = w_{t+1}(x)$.

On the other hand, for all x, $w_t(x)$ is lower bounded by 0. Therefore, for each x, limit of sequence $(w_t(x))_t$ exists. By uniqueness of the limit in the contraction mapping principle, $w(x) = V^\beta(x)$. Optimality of the stationary policy ϕ^* is shown in Theorem (7.2).

Assume now that $w_{t+1} = w_t$ for some t. We know w_{t+1} satisfies the equation $w_{t+1}(x) = c(x, \phi_{t+1}(x)) + \beta P w_{t+1}(x, \phi_{t+1}(x))$. So $w_{t+1}(x) = c(x, \phi_{t+1}(x)) + \beta P w_{t+1}(x, \phi_{t+1}(x)) = \mathbb{T}^\beta w_t(x)$, which means $w_{t+1} = w_t$ is the solution of DCOE. By Theorem (7.2), $w_t = V^\beta = w_{t+1}$. Then ϕ_{t+1} is the solution to $V^\beta(x) = c(x, \phi_{t+1}) + \beta P V^\beta(x, \phi_{t+1})$. Again, by Theorem (7.2) $\phi_{t+1} = g^*$.

■