**Lecture Note 3**

# 1 Value Iteration

Using value iteration, starting at an arbitrary $J_0$, we generate a sequence of $\{J_k\}$ by

$$J_{k+1} = TJ_k, \forall \text{ integer } k \geq 0.$$

We have shown that the sequence $J_k \to J^*$ as $k \to \infty$, and derived the error bounds

$$||J_k - J^*||_\infty \leq \alpha^k ||J_0 - J^*||_\infty$$

Recall that the greedy policy $u_J$ with respect to value $J$ is defined as $TJ = T_{u_J} J$. We also denote $u_k = u_{J_k}$ as the greedy policy with respect to value $J_k$. Then, we have the following lemma.

**Lemma 1** *Given* $\alpha \in (0,1)$,

$$||J_{u_k} - J_k||_\infty \leq \frac{1}{1-\alpha} ||TJ_k - J_k||_\infty$$

**Proof:**

$$\begin{aligned}
J_{u_k} - J_k &= (I - \alpha P_{u_k})^{-1} g_{u_k} - J_k \\
&= (I - \alpha P_{u_k})^{-1} \left( g_{u_k} + \alpha P_{u_k} J_k - J_k \right) \\
&= (I - \alpha P_{u_k})^{-1} \left( TJ_k - J_k \right) \\
&= \sum_{t=0}^{\infty} \alpha^t P_{u_k}^t (TJ_k - J_k) \\
&\leq \sum_{t=0}^{\infty} \alpha^t P_{u_k}^t e ||TJ_k - J_k||_\infty \\
&= \sum_{t=0}^{\infty} \alpha^t e ||TJ_k - J_k||_\infty \\
&= \frac{e}{1-\alpha} ||TJ_k - J_k||_\infty
\end{aligned}$$

where $I$ is an identity matrix, and $e$ is a vector of unit elements with appropriate dimension. The third equality comes from $TJ_k = g_{u_k} + \alpha P_{u_k} J_k$, i.e., $u_k$ is the greedy policy w.r.t. $J_k$, and the forth equality holds because $(I - \alpha P_{u_k})^{-1} = \sum_{t=0}^{\infty} \alpha^t P_{u_k}^t$. By switching $J_{u_k}$ and $J_k$, we can obtain $J_k - J_{u_k} \leq \frac{e}{1-\alpha} ||TJ_k - J_k||_\infty$, and hence conclude

$$|J_{u_k} - J_k| \leq \frac{e}{1-\alpha} |TJ_K - J_K|$$

or, equivalently,

$$||J_{u_k} - J_k||_\infty \leq \frac{1}{1-\alpha} ||TJ_k - J_k||_\infty.$$

□

**Theorem 1**

$$||J_{u_k} - J^*||_\infty \leq \frac{2}{1-\alpha}||J_k - J^*||_\infty$$

**Proof:**

$$\begin{aligned}
||J_{u_k} - J^*||_\infty &= ||J_{u_k} - J_k + J_k - J^*||_\infty \\
&\leq ||J_{u_k} - J_k||_\infty + ||J_k - J^*||_\infty \\
&\leq \frac{1}{1-\alpha}||TJ_k - J^* + J^* - J_k||_\infty + ||J_k - J^*||_\infty \\
&\leq \frac{1}{1-\alpha}\left(||TJ_k - J^*||_\infty + ||J^* - J_k||_\infty\right) + ||J_k - J^*||_\infty \\
&\leq \frac{2}{1-\alpha}||J_k - J^*||_\infty
\end{aligned}$$

The second inequality comes from Lemma 1 and the third inequality holds by the contraction principle. □

## 2 Optimality of Stationary Policy

Before proving the main theorem of this section, we introduce the following useful lemma.

**Lemma 2** *If $J \leq TJ$, then $J \leq J^*$. If $J \geq TJ$, then $J \geq J^*$.*

**Proof:** Suppose that $J \leq TJ$. Applying operator $T$ on both sides repeatedly $k-1$ times and by the monotonicity property of $T$, we have

$$J \leq TJ \leq T^2 J \leq \cdots \leq T^k J.$$

For sufficiently large $k$, $T^k J$ approaches to $J^*$. We hence conclude $J \leq J^*$. The other statement follows the same argument. □

We show the optimality of the stationary policy by the following theorem.

**Theorem 2** *Let $u = (u_1, u_2, \ldots)$ be any policy and let $u^* \equiv u_{J^*}$*[1]*. Then,*

$$J_u \geq J_{u^*} = J^*.$$

*Moreover, let $u$ be any stationary policy such that $T_u J^* \neq TJ^*$.*[2] *Then, $J_u(x) > J^*(x)$ for at least one state $x \in \mathcal{S}$.*

**Proof:** Since $g_\cdot$ and $J$ are finite, there exists a real number $M$ satisfying $||g_u||_\infty \leq M$ and $||J^*||_\infty \leq M$. Define

$$J_u^k = T_{u_1} T_{u_2} \ldots T_{u_k} J^*.$$

[1] That is, $J^* = TJ^* = T_{u^*} J^*$.

[2] That is to say that $u$ is not a greedy policy w.r.t. $J^*$.

Then

$$||J_u^k - J_u||_\infty \leq M(1 + \frac{1}{1-\alpha})\alpha^k \to 0 \text{ as } k \to \infty.$$

If $u = (u^*, u^*, \dots)$, then

$$||J_{u^*} - J_{u^*}^k||_\infty \to 0 \text{ as } k \to \infty.$$

Thus, we have $J_{u^*}^k = T_{u^*}^k J^* = T_{u^*}^{k-1}(TJ^*) = T_{u^*}^{k-1} J^* = J^*$. Therefore $J_{u^*} = J^*$. For any other policy, for all $k$,

$$\begin{aligned}
J_u &\geq J_u^k - M\left(1 + \frac{1}{1-\alpha}\right)\alpha^k \\
&= T_{u_1} \dots \underbrace{T_{u_k} J^*}_{\geq TJ^*} - M\left(1 + \frac{1}{1-\alpha}\right)\alpha^k \\
&\geq T_{u_1} \dots T_{u_{k-1}} \underbrace{TJ^*}_{=J^*} - M\left(1 + \frac{1}{1-\alpha}\right)\alpha^k \\
&\geq \dots \geq J^* - M\left(1 + \frac{1}{1-\alpha}\right)\alpha^k
\end{aligned}$$

Therefore $J_u \geq J^*$. Take a stationary policy $u$ such that $T_u J^* \neq TJ^*$, i.e. $T_u J^* \geq TJ^*$, and $\exists$ at least one state $x \in \mathcal{S}$ such that $(T_u J^*)(x) > (TJ^*)(x)$. Observe

$$J^* = TJ^* \leq T_u J^*$$

Applying $T_u$ on both sides and by the monotonicity property of $T$, or applying Lemma 2,

$$J^* \leq T_u J^* \leq T_u^2 J^* \leq T_u^k J^* \to J_u$$

and $J^*(x) < J_u(x)$ for at least one state $x$. □

# 3 Policy Iteration

The policy iteration algorithm proceeds as follows.

1. Start with policy $u_0$, k=0;
2. Evaluate $J_{u_k} = g_{u_k} + \alpha P_{u_k} J_{u_k}$;
3. Let $u_{k+1} = u_{J_{u_k}}$;
4. If $u_{k+1} = u_k$ stop; otherwise, go back to Step 2.

Note that Step 2 aims at getting a better policy for each iteration. Since the set of policies is finite, the algorithm will terminate in finite steps. We state this concept formally by the following theorem.

**Theorem 3** *Policy iteration converges to $u^*$ after a finite number of iterations.*

**Proof:** If $u_k$ is optimal, then we are done. Now suppose that $u_k$ is not optimal. Then

$$TJ_{u_k} \leq T_{u_k} J_{u_k} = J_{u_k}$$

with strict inequality for at least one state $x$. Since $T_{u_{k+1}} J_{u_k} = TJ_{u_k}$ and $J_{u_k} = T_{u_k} J_{u_k}$, we have

$$J_{u_k} = T_{u_k} J_{u_k} \geq TJ_{u_k} = T_{u_{k+1}} J_{u_k} \geq T^n_{u_{k+1}} J_{u_k} \to J_{u_{k+1}} \text{ as } n \to \infty.$$

Therefore, policy $u_{k+1}$ is an improvement over policy $u_k$. $\square$

In step 2, we solve $J_{u_k} = g_{u_k} + \alpha P_{u_k} J_{u_k}$, which would require a significant amount of computations. We thus introduce another algorithm which has fewer iterations in step 2.

## 3.1 Asynchronous Policy Iteration

The algorithm goes as follows.

1. Start with policy $u_0$, cost-to-go function $J_0$, $k = 0$
2. For some subset $\mathcal{S}_k \subseteq \mathcal{S}$, do one of the following

$$\begin{aligned} &\text{(i) value update} \qquad && (J_{k+1})(x) = (T_{u_k} J_k)(x), \forall x \in \mathcal{S}_k, \\ &\text{(ii) policy update} \qquad && u_{k+1}(x) = u_{J_k}(x), \forall x \in \mathcal{S}_k \end{aligned}$$

3. $k = k + 1$; go back to step 2

**Theorem 4** *If $T_{u_0} J_0 \leq J_0$ and infinitely many value and policy updates are performed on each state, then*

$$\lim_{k \to \infty} J_k = J^*.$$

**Proof:** We prove this theorem by two steps. First, we will show that

$$J^* \leq J_{k+1} \leq J_k, \quad \forall\, k.$$

This implies that $J_k$ is a nonincreasing sequence. Since $J_k$ is lower bounded by $J^*$, $J_k$ will converge to some value, i.e., $J_k \searrow \bar{J}$ as $k \to \infty$. Next, we will show that $J_k$ will converge to $J^*$, i.e., $\bar{J} = J^*$.

**Lemma 3** *If $T_{u_0} J_0 \leq J_0$, the sequence $J_k$ generated by asynchronous policy iteration converges.*

**Proof:** We start by showing that, if $T_{u_k} J_k \leq J_k$, then $T_{u_{k+1}} J_{k+1} \leq J_{k+1} \leq J_k$. Suppose we have a value update. Then,

$$\left.\begin{aligned} &\forall x \in \mathcal{S}_k, \quad J_{k+1}(x) = (T_{u_k} J_k)(x) \leq J_k(x) \\ &\forall x \notin \mathcal{S}_k, \quad J_{k+1}(x) = J_k(x) \end{aligned}\right\} J_{k+1} \leq J_k$$

Thus,

$$(T_{u_{k+1}} J_{k+1})(x) = (T_{u_k} J_{k+1})(x) \leq (T_{u_k} J_k)(x) \begin{cases} = J_{k+1}(x), & \forall x \in \mathcal{S}_k \\ \leq J_k(x) = J_{k+1}(x), & \forall x \notin \mathcal{S}_k \end{cases}$$

Now suppose that we have a policy update. Then $J_{k+1} = J_k$. Moreover, for $x \in \mathcal{S}_k$, we have

$$\begin{aligned}
(T_{u_{k+1}} J_{k+1})(x) &= (T_{u_{k+1}} J_k)(x) \\
&= (TJ_k)(x) \\
&\leq (T_{u_k} J_k)(x) \\
&\leq J_k(x) \\
&= J_{k+1}(x).
\end{aligned}$$

The first equality follows from $J_k = J_{k+1}$, the second equality and first inequality follows from the fact that $u_{k+1}(x)$ is greedy with respect to $J_k$ for $x \in \mathcal{S}_k$, the second inequality follows from the induction hypothesis, and the third equality follows from $J_k = J_{k+1}$. For $x \notin \mathcal{S}_k$, we have

$$\begin{aligned}
(T_{u_{k+1}} J_{k+1})(x) &= (T_{u_k} J_k)(x) \\
&\leq J_k(x) \\
&= J_{k+1}(x).
\end{aligned}$$

The equalities follow from $J_k = J_{k+1}$ and $u_{k+1}(x) = u_k(x)$ for $x \notin \mathcal{S}_k$, and the inequality follows from the induction hypothesis.

Since by hypothesis $T_{u_0} J_0 \leq J_0$, we conclude that $J_k$ is a decreasing sequence. Moreover, we have $T_{u_k} J_k \leq J_k$, hence $J_k \geq J_{u_k} \geq J^*$, so that $J_k$ is bounded bellow. It follows that $J_k$ converges to some limit $\bar{J}$. □

**Lemma 4** *Suppose that $J_k \searrow \bar{J}$, where $J_k$ is generated by asynchronous policy iteration, and suppose that there are infinitely many value and policy updates at each state. Then $\bar{J} = J^*$.*

**Proof:** First note that, since $TJ_k \leq J_k$, by continuity of the operator $T$, we must have $T\bar{J} \leq \bar{J}$. Now suppose that $(T\bar{J})(x) < \bar{J}(x)$ for some state $x$. Then, by continuity, there is an iteration index $\bar{k}$ such that $(TJ_k)(x) < \bar{J}(x)$ for all $k \geq \bar{k}$. Let $k'' > k' > \bar{k}$ correspond to iterations of the asynchronous policy iteration algorithm such that there is a policy update at state $x$ at iteration $k'$, a value update at state $x$ at iteration $k''$, and no updates at state $x$ in iterations $k' < k < k''$. Such iterations are guaranteed to exist since there are infinitely many value and policy update iterations at each state. Then we have $u_{k''}(x) = u_{k'+1}(x)$, $J_{k''}(x) = J_{k'}(x)$, and

$$\begin{aligned}
J_{k''+1}(x) &= (T_{u_{k''}} J_{k''})(x) \\
&= (T_{u_{k'+1}} J_{k''})(x) \\
&\leq (T_{u_{k'+1}} J_{k'})(x) \\
&= (TJ_{k'})(x) \\
&< \bar{J}.
\end{aligned}$$

The first equality holds because there is a value update at state $x$ at iteration $k''$, the second equality holds because $u_{k''}(x) = u_{k'+1}(x)$, the first inequality holds because $J_k$ is decreasing and $T_{u_{k'+1}}$ is monotone and the third equality holds because there is a policy update at state $x$ at iteration $k'$.

We have concluded that $J_{k''+1} < \bar{J}$. However by hypothesis $J_k \downarrow \bar{J}$, we have a contradiction, and it must follow that $T\bar{J} = \bar{J}$, so that $\bar{J} = J^*$. $\square$