Lecture 4

Dynamic Programming and Operator Theory

Goals of this lecture

1. Introduce operator-based formalism for reasoning about value functions and policies.
2. Illustrate how policy evaluation and policy improvement can be expressed as operators.
3. Prove key properties of these operators: monotonicity and contraction.
4. Provide rigorous proofs of the policy improvement theorem and Bellman optimality principle.
5. Present and analyze the Value Iteration algorithms as practical instantiations of this theory.

4.1 The Bellman Operator for Policy Evaluation

Motivation. Previously, we saw that the value function v^π satisfies the recursive Bellman expectation equation:

$$v^\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v^\pi(S_{t+1}) \mid S_t = s].$$

Rather than viewing this purely as a fixed-point identity, we now define an operator that maps any function $v : \mathcal{S} \to \mathbb{R}$ to a new function. This operator perspective is both conceptually elegant and practically powerful.

Definition (Bellman Operator for Policy Evaluation). Given a Markov policy π, define the Bellman operator T_π acting on value functions $v : \mathcal{S} \to \mathbb{R}$ as:

$$[T_\pi v](s) := \mathbb{E}_\pi \left[R_{t+1} + \gamma v(S_{t+1}) | S_t = s\right].$$

In finite MDPs, this expression becomes:

$$[T_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) \left[r + \gamma v(s')\right].$$

Fixed Point Characterization. It is immediate from the Bellman expectation equation that:

$$v^\pi = T_\pi v^\pi.$$

That is, the value function v^π is a fixed point of the operator T_π. In fact, it is easy to show that under mild conditions, $\gamma \in (0, 1)$, the fixed point is unique.

Remarks.

- The operator T_π maps value functions to value functions: $T_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$.
- Intuitively, $T_\pi v$ gives the expected return when we perform one step of policy π, collect immediate reward, and continue with value v.
- Computing v^π amounts to finding the fixed point of T_π, which we can compute exactly (via matrix inversion) or approximately (via iterative updates).

4.2 The Bellman Optimality Operator

Motivation. Recall that the optimal value function v^* satisfies the Bellman optimality equation:

$$v^*(s) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r \mid s, a)\,[r + \gamma v^*(s')].$$

This naturally suggests defining an operator that captures this maximization. We now introduce the Bellman optimality operator, which plays a central role in algorithms for computing optimal policies.

Definition (Bellman Optimality Operator). Define the Bellman optimality operator T_* acting on value functions $v : \mathcal{S} \to \mathbb{R}$ as:

$$[T_* v](s) := \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r \mid s, a)\,[r + \gamma v(s')].$$

This operator aggregates, for each state s, the best expected return over all possible actions, assuming future values are given by v.

Fixed Point Characterization. The optimal value function v^* is the unique fixed point of the Bellman optimality operator:

$$v^* = T_* v^*.$$

Moreover, any policy π that is greedy with respect to v^* is an optimal policy:

$$\pi(s) \in \arg\max_{a\in\mathcal{A}} \sum_{s',r} p(s', r \mid s, a)[r + \gamma v^*(s')].$$

Remarks.

- The operator T_* maps value functions to value functions: $T_* : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$.
- Unlike T_π, which corresponds to a fixed policy, T_* selects the best action at each state—introducing a nonlinearity through the maximization.
- Solving $v^* = T_* v^*$ gives the optimal state values; the corresponding greedy policy yields optimal behavior.
- This motivates algorithms such as *value iteration*, which apply T_* repeatedly to converge to v^*.

4.3 Operator Theory in Reinforcement Learning

Motivation. Many core reinforcement learning problems can be cast as solving a fixed-point equation involving an operator on value functions:

$$v^\pi = T_\pi v^\pi, \qquad v^* = T_* v^*.$$

Understanding the behavior of such operators is essential for designing and analyzing algorithms such as value iteration and policy iteration.

4.3.1 Operators and Fixed Points

We now formalize a few basic concepts that underpin value function updates and convergence. Throughout this section, we assume a finite state space $\mathcal{S}$, so the space of real-valued functions on states is $\mathbb{R}^{\mathcal{S}} \cong \mathbb{R}^n$.

Definition (Operator). An *operator* is a function that maps value functions to value functions:

$$T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}.$$

Operators like T_π and T^* (defined earlier) are central to dynamic programming and reinforcement learning.

Definition (Fixed Point). A vector $v \in \mathbb{R}^{\mathcal{S}}$ is a *fixed point* of an operator T if:

$$Tv = v.$$

That is, applying T to v returns v itself. The Bellman expectation and optimality equations are examples of fixed-point equations.

Definition (Norm). A *norm* assigns a metric to each function $v \in \mathbb{R}^{\mathcal{S}}$. Common examples include:

$$\|v\|_\infty := \max_{s\in\mathcal{S}} |v(s)|, \qquad \|v\|_1 := \sum_{s\in\mathcal{S}} |v(s)|, \qquad \|v\|_2 := \sqrt{\sum_{s\in\mathcal{S}} v(s)^2}.$$

Each norm induces a notion of distance and convergence, which allows us to study iterative algorithms like value iteration using tools from metric fixed-point theory.

4.3.2 Monotonicity and Contraction

Definition (Monotonicity). An operator $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is said to be *monotone* if for any $v, w \in \mathbb{R}^{\mathcal{S}}$,

$$v(s) \le w(s) \quad \forall s \in \mathcal{S} \quad \Rightarrow \quad [Tv](s) \le [Tw](s) \quad \forall s \in \mathcal{S}.$$

Definition (Contraction). Let $\|\cdot\|$ be a norm on $\mathbb{R}^{\mathcal{S}}$, and let $\gamma \in [0, 1)$. The operator T is a γ-*contraction* with respect to this norm if:

$$\|Tv - Tw\| \le \gamma \|v - w\| \quad \text{for all } v, w \in \mathbb{R}^{\mathcal{S}}.$$

Theorem 4.1 (Banach Fixed Point Theorem). *Let $T : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ be a γ-contraction with respect to a norm $\|\cdot\|$, where $0 \le \gamma < 1$. Then:*

1. *There exists a unique fixed point $v^* \in \mathbb{R}^{\mathcal{S}}$ such that $Tv^* = v^*$.*
2. *For any initial value $v_0 \in \mathbb{R}^{\mathcal{S}}$, the sequence defined by:*

$$v_{k+1} := Tv_k$$

   converges to v^*.
3. *Moreover, the convergence is geometric:*

$$\|v_k - v^*\| \leq \gamma^k \|v_0 - v^*\| \quad \textit{for all } k \geq 0.$$

The proof of theis theorem relies on showing that the sequence generated by T is a Cauchy sequence.

Definition (Cauchy Sequence). A sequence $(v_k)_{k\geq 0} \subset \mathbb{R}^{\mathcal{S}}$ is called a *Cauchy sequence* if for every $\varepsilon > 0$, there exists an integer N such that:

$$\|v_k - v_\ell\| < \varepsilon \quad \text{for all } k, \ell \geq N.$$

In a complete normed vector space (like $\mathbb{R}^{\mathcal{S}}$), every Cauchy sequence converges to a limit in the space.

Proof. As mentioned before, we will show that the sequence $(v_k)_{k\geq 0}$, defined recursively by $v_{k+1} := Tv_k$, is Cauchy.

First, note that by the contraction property:

$$\|v_{k+1} - v_k\| = \|Tv_k - Tv_{k-1}\| \leq \gamma \|v_k - v_{k-1}\|.$$

By induction, this implies:

$$\|v_{k+1} - v_k\| \leq \gamma^k \|v_1 - v_0\|.$$

Now consider, for $k > \ell$,

$$\|v_k - v_\ell\| \leq \sum_{j=\ell}^{k-1} \|v_{j+1} - v_j\| \leq \|v_1 - v_0\| \sum_{j=\ell}^{k-1} \gamma^j \leq \frac{\gamma^\ell}{1-\gamma} \|v_1 - v_0\|.$$

This shows that (v_k) is Cauchy, hence converges to some limit v^* since $\mathbb{R}^{\mathcal{S}}$ is complete. Taking the limit in $v_{k+1} = Tv_k$ shows $v^* = Tv^*$, i.e., v^* is a fixed point.

Uniqueness follows from the contraction property: if v^* and w^* are fixed points, then

$$\|v^* - w^*\| = \|Tv^* - Tw^*\| \leq \gamma \|v^* - w^*\|,$$

which implies $v^* = w^*$ since $\gamma < 1$. □

Corollary 4.1 (Termination Criterion)**.** *Let T be a γ-contraction and suppose that for some $k \geq 1$,*

$$\|v_k - v_{k-1}\| \leq \varepsilon.$$

Then the distance to the fixed point v^* is bounded by:

$$\|v_k - v^*\| \leq \frac{\gamma}{1-\gamma} \|v_k - v_{k-1}\| \leq \frac{\gamma\varepsilon}{1-\gamma}.$$

Proof. We begin by creating a telescopic series for the term $\|v_k - v^*\|$:

$$\|v_k - v^*\| \leq \sum_{j=0}^{\infty} \|v_{k+j+1} - v_{k+j}\| \leq \sum_{j=0}^{\infty} \gamma^{j+1} \|v_k - v_{k-1}\| = \|v_k - v_{k-1}\| \gamma \sum_{j=0}^{\infty} \gamma^j.$$

The geometric series sums to $1/(1-\gamma)$, so we obtain:

$$\|v_k - v^*\| \leq \frac{\gamma}{1-\gamma} \|v_k - v_{k-1}\|.$$

Substituting the upper bound $\|v_k - v_{k-1}\| \leq \varepsilon$ completes the proof:

$$\|v_k - v^*\| \leq \frac{\gamma\varepsilon}{1-\gamma}.$$

□

4.4 Properties of the Bellman Operator T_π

Setup. Recall the Bellman operator for a fixed Markov policy π, defined for any $v : \mathcal{S} \to \mathbb{R}$ as:

$$[T_\pi v](s) := \mathbb{E}_\pi \left[R_{t+1} + \gamma v(S_{t+1}) | S_t = s \right].$$

We now study the properties of this operator.

Theorem 4.2 (Properties of T_π)**.** *Let T_π be the Bellman operator associated with a fixed Markov policy π and assume $\gamma \in [0,1)$. Then:*

1. ***(Monotonicity)** If $v(s) \leq w(s)$ for all $s \in \mathcal{S}$, then $[T_\pi v](s) \leq [T_\pi w](s)$ for all s.*
2. ***(Contraction)** For the norm $\|v - w\|_\infty := \max_s |v(s) - w(s)|$, T_π is a γ-contraction:*

$$\|T_\pi v - T_\pi w\|_\infty \leq \gamma \|v - w\|_\infty.$$

3. ***(Unique Fixed Point)** T_π has a unique fixed point, denoted v^π, and it satisfies $v^\pi = T_\pi v^\pi$.*

Proof.
(1) Monotonicity. Suppose $v(s) \leq w(s)$ for all s. Then for any s,

$$\begin{aligned} [T_\pi v](s) &= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v(s') \right] \\ &\leq \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma w(s') \right] = [T_\pi w](s). \end{aligned}$$

(2) Contraction. For any $v, w : \mathcal{S} \to \mathbb{R}$, and *any* $s \in \mathcal{S}$,

$$\begin{aligned} |[T_\pi v](s) - [T_\pi w](s)| &= \left| \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \, \gamma [v(s') - w(s')] \right| \\ &\leq \gamma \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) \, |v(s') - w(s')| \\ &\leq \gamma \|v - w\|_\infty \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a) = \gamma \|v - w\|_\infty. \end{aligned}$$

Taking the maximum over s yields:

$$\|T_\pi v - T_\pi w\|_\infty \le \gamma \|v - w\|_\infty.$$

(3) Unique fixed point. The Banach fixed point theorem (see previous section) guarantees that since T_π is a γ-contraction on the complete metric space $(\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$, it admits a unique fixed point, and iterating the operator:

$$v_{k+1} := T_\pi v_k$$

converges to v^π for any initial v_0. □

Value Iteration for Policy Evaluation. The monotonicity and contraction properties of the Bellman operator T_π yield a practical algorithm for computing the value function v^π of a fixed policy π. Starting from any initial guess $v_0 : \mathcal{S} \to \mathbb{R}$, we define a sequence:

$$v_{k+1} := T_\pi v_k.$$

By the Banach Fixed Point Theorem, this sequence converges geometrically to the unique fixed point v^π:

$$\|v_k - v^\pi\| \le \gamma^k \|v_0 - v^\pi\|.$$

Early Termination Rule. Suppose that at iteration k we observe:

$$\|v_k - v_{k-1}\| \le \bar{\varepsilon}.$$

Then, the actual distance to the true value function is bounded by:

$$\|v_k - v^\pi\| \le \frac{\gamma}{1-\gamma} \bar{\varepsilon}.$$

Hence, to guarantee that $\|v_k - v^\pi\| \le \varepsilon$, it suffices to stop when:

$$\|v_k - v_{k-1}\| \le \bar{\varepsilon} := \frac{(1-\gamma)}{\gamma} \varepsilon.$$

Summary. This yields a simple yet principled procedure for evaluating a fixed policy in a finite MDP:

- Initialize v_0 arbitrarily.
- Iterate $v_{k+1} := T_\pi v_k$.
- Terminate when $\|v_k - v_{k-1}\|$ falls below a predefined threshold.
- Guaranteed approximation quality: $\|v_k - v^\pi\| \le \varepsilon$.

4.5 Properties of the Optimal Bellman Operator T^*

Definition. The optimal Bellman operator T^* is defined as:

$$[T^* v](s) := \max_{a \in \mathcal{A}} \sum_{s', r} p(s', r \mid s, a) \left[r + \gamma v(s') \right].$$

This operator corresponds to acting greedily in the one-step lookahead based on the current value function v. It represents the best possible expected return starting from state s, assuming optimal decisions are made at each step from then on.

Theorem 4.3 (Properties of the Optimal Bellman Operator T^*)**.** *Let T^* be the optimal Bellman operator defined by*

$$[T^*v](s) := \max_{a\in\mathcal{A}} \sum_{s',r} p(s', r \mid s, a) \; \left[r + \gamma v(s')\right],$$

and let $\gamma \in [0, 1)$. Then:

1. ***(Monotonicity)*** *If $v(s) \le w(s)$ for all $s \in \mathcal{S}$, then $[T^*v](s) \le [T^*w](s)$ for all s.*
2. ***(Contraction)*** *T^* is a γ-contraction under the sup-norm:*

$$\|T^*v - T^*w\|_\infty \le \gamma\|v - w\|_\infty.$$

3. ***(Unique Fixed Point)*** *T^* has a unique fixed point v^*, and it satisfies $v^* = T^*v^*$.*

Proof. (1) Monotonicity. Assume $v(s) \le w(s)$ for all $s \in \mathcal{S}$. For any $s \in \mathcal{S}$ and any $a \in \mathcal{A}$,

$$\sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v(s')\right] \le \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma w(s')\right],$$

since $v(s') \le w(s')$ and r is the same in both. Taking the maximum over a preserves the inequality:

$$[T^*v](s) = \max_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v(s')\right] \le \sum_{s',r} p(s', r \mid s, a^*) \left[r + \gamma w(s')\right] \tag{4.1}$$

$$\le \max_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma w(s')\right] = [T^*w](s). \tag{4.2}$$

where a^* is the maximizer of the first term.

(2) Contraction. Fix $s \in \mathcal{S}$. Let

$$q_v(s, a) := \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v(s')\right], \quad q_w(s, a) := \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma w(s')\right].$$

Then:

$$\begin{aligned} |[T^*v](s) - [T^*w](s)| &= \left|\max_a q_v(s, a) - \max_a q_w(s, a)\right| \\ &\le \max_a |q_v(s, a) - q_w(s, a)| \quad \text{(see lemma below)} \\ &= \max_a \left|\sum_{s',r} p(s', r \mid s, a)\, \gamma[v(s') - w(s')]\right| \\ &\le \gamma \max_a \sum_{s',r} p(s', r \mid s, a)|v(s') - w(s')| \\ &\le \gamma\|v - w\|_\infty \quad \left(\text{since } \sum_{s',r} p(s', r \mid s, a) = 1\right). \end{aligned}$$

Taking the maximum over s gives:

$$\|T^*v - T^*w\|_\infty \le \gamma\|v - w\|_\infty.$$

(3) Unique Fixed Point. Because T^* is a γ-contraction on the complete metric space $(\mathbb{R}^{\mathcal{S}}, \|\cdot\|_\infty)$, Banach's Fixed Point Theorem guarantees the existence and uniqueness of a fixed point v^* such that $T^*v^* = v^*$, and that iterative application of T^* converges to v^* from any initial value. □

Lemma 4.1 (Max Difference Inequality)**.** *Let $\phi, \psi : \mathcal{A} \to \mathbb{R}$ be two functions. Then:*

$$\left| \max_a \phi(a) - \max_a \psi(a) \right| \leq \max_a |\phi(a) - \psi(a)|.$$

Proof. Assume without loss of generality that $\max_a \phi(a) \geq \max_a \psi(a)$. Let $a^* = \arg\max_a \phi(a)$. Then:

$$\max_a \phi(a) - \max_a \psi(a) \leq \phi(a^*) - \psi(a^*) \leq |\phi(a^*) - \psi(a^*)| \leq \max_a |\phi(a) - \psi(a)|.$$

□

Value Iteration for Optimal Bellman Operator. The monotonicity and contraction properties of the optimal Bellman operator T^* provide a practical and theoretically grounded method for computing the optimal value function v^*. Starting from an arbitrary initial guess $v_0 : \mathcal{S} \to \mathbb{R}$, we define the sequence:

$$v_{k+1} := T^* v_k = \max_{a \in \mathcal{A}} \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_k(s') \right].$$

By the Banach Fixed Point Theorem, this sequence converges to the unique fixed point v^* at a geometric rate:

$$\|v_k - v^*\| \leq \gamma^k \|v_0 - v^*\|.$$

This iterative process is known as *value iteration for optimal control.*

Early Termination Rule. Suppose that at iteration k, the difference between successive value estimates satisfies:

$$\|v_k - v_{k-1}\| \leq \bar{\varepsilon}.$$

Then the distance to the true optimum is bounded by:

$$\|v_k - v^*\| \leq \frac{\gamma}{1-\gamma} \bar{\varepsilon}.$$

Hence, to guarantee that $\|v_k - v^*\| \leq \varepsilon$, it suffices to terminate when:

$$\|v_k - v_{k-1}\| \leq \bar{\varepsilon} := \frac{(1-\gamma)}{\gamma} \varepsilon.$$

Theorem 4.4 (Performance Guarantee for Greedy Policy from Approximate Value)**.** *Let v_k be an approximate value function such that*

$$\|v_k - v_{k-1}\|_\infty \leq \varepsilon \frac{1-\gamma}{\gamma}, \ \textit{which implies} \quad \|v_k - v^*\|_\infty \leq \varepsilon,$$

and define the greedy policy π_k with respect to v_k as:

$$\pi_k(s) \in \arg\max_{a \in \mathcal{A}} \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_k(s') \right].$$

Then the value function of π_k satisfies:

$$\|v^{\pi_k} - v^*\|_\infty \leq 2\varepsilon.$$

Proof. Using the triangle inequality, we have:

$$\|v^{\pi_k} - v^*\|_\infty \leq \|v^{\pi_k} - v_k\|_\infty + \|v_k - v^*\|_\infty.$$

Since $\|v_k - v^*\|_\infty \leq \varepsilon$, we only need to show $\|v^{\pi_k} - v_k\|_\infty \leq \varepsilon$.

Observe that v^{π_k} satisfies the fixed-point equation:

$$v^{\pi_k} = T_{\pi_k} v^{\pi_k}.$$

Thus:

$$\|v^{\pi_k} - v_k\|_\infty = \|T_{\pi_k} v^{\pi_k} - v_k\|_\infty.$$

By adding and subtracting $T_{\pi_k} v_k$ and using the contraction property, we get:

$$\begin{aligned}\|v^{\pi_k} - v_k\|_\infty &= \|T_{\pi_k} v^{\pi_k} - T_{\pi_k} v_k + T_{\pi_k} v_k - v_k\|_\infty \\ &\leq \gamma \|v^{\pi_k} - v_k\|_\infty + \|T_{\pi_k} v_k - v_k\|_\infty.\end{aligned}$$

Rearranging, we obtain:

$$(1-\gamma)\|v^{\pi_k} - v_k\|_\infty \leq \|T_{\pi_k} v_k - v_k\|_\infty.$$

Since π_k is greedy w.r.t. v_k, we have $T_{\pi_k} v_k = T^* v_k$. Thus:

$$\begin{aligned}\|T_{\pi_k} v_k - v_k\|_\infty = \|T^* v_k - v_k\|_\infty &\leq \|T^* v_k - T^* v_{k-1}\|_\infty \\ &\leq \gamma \|v_k - v_{k-1}\| \leq \gamma \varepsilon \frac{1-\gamma}{\gamma} = \varepsilon(1-\gamma)\end{aligned}$$

which finally implies that

$$\|v^{\pi_k} - v_k\|_\infty \leq \varepsilon. \tag{4.3}$$

Thus, combining (4.3) with $\|v_k - v^*\|_\infty \leq \varepsilon$ leads to:

$$\|v^{\pi_k} - v^*\|_\infty \leq \|v^{\pi_k} - v_k\|_\infty + \|v_k - v^*\|_\infty \leq 2\varepsilon.$$

□

Summary. *Value iteration* provides a general-purpose algorithm for computing optimal solutions in finite MDPs:

- It produces a sequence of value estimates v_k that converges geometrically to the optimal value function v^*.
- The contraction property yields an explicit stopping rule: to guarantee $\|v_k - v^*\|_\infty \leq \varepsilon$, it suffices to stop when $\|v_k - v_{k-1}\|_\infty \leq \frac{(1-\gamma)}{\gamma}\varepsilon$.
- A greedy policy π_k derived from v_k satisfies $\|v^{\pi_k} - v^*\|_\infty \leq 2\varepsilon$, ensuring near-optimal performance.
- This method is simple to implement and serves as a foundation for more advanced algorithms such as policy iteration and approximate dynamic programming.