Theoretical Guarantees in Value-Based Reinforcement Learning:

A Unifying Perspective Through Fixed-Point Theory

By Taha Majlesi

July 17, 2025

Abstract

Reinforcement Learning (RL) stands as a distinct paradigm within machine learning, distinguished by its focus on sequential decision-making under uncertainty. At its core, RL addresses the problem of an autonomous agent learning to achieve a goal through interaction with a dynamic environment. This report provides a comprehensive and rigorous exploration of the theoretical foundations of value-based RL. We address fundamental questions regarding the existence, uniqueness, and computability of optimal value functions.

The central thesis is that the problem of finding an optimal value function can be elegantly framed as a fixed-point problem. By defining the Bellman Optimality Operator, we demonstrate that the optimal value function is its unique fixed point. The theoretical linchpin of this framework is the Banach Fixed-Point Theorem. This powerful result not only guarantees the existence and uniqueness of the optimal value function but also validates the convergence of the Value Iteration algorithm, providing a constructive method for its computation. This report systematically builds this argument, bridging the gap between the practical algorithms of RL and the abstract mathematical machinery of functional analysis.

Contents

Part I

The Problem: Optimality in Sequential Decision-Making

Chapter 1

Foundations of Reinforcement Learning

1.1 Introduction

Reinforcement Learning (RL) addresses the problem of an autonomous agent learning to achieve a goal through interaction with a dynamic environment. The agent's objective is not to classify data or predict a static value, but to learn a strategy, or *policy*, that maximizes a cumulative reward signal over time. This process of learning through trial and error, guided by scalar feedback, is fundamental to developing intelligent systems capable of complex control tasks.

Central to many successful RL approaches is the concept of a *value function*. A value function serves as a predictive model of future reward, quantifying the long-term desirability of being in a particular state or taking a specific action. By learning an accurate value function, an agent can formulate an effective policy by simply selecting actions that lead to states of higher value. The ultimate goal, therefore, often becomes the discovery of the *optimal value function*—the one corresponding to the best possible policy.

This pursuit of optimality, however, raises fundamental theoretical questions:

1. **Existence:** Does an optimal value function, which we denote as q^*, even exist?

2. **Uniqueness:** If such an optimal value function exists, is it unique?

3. **Computability:** If an optimal and unique value function exists, can we design an algorithm that is guaranteed to find it?

This report will demonstrate that the answers to these questions are resoundingly affirmative for a broad class of problems, using the powerful framework of fixed-point theory.

To ensure clarity and precision, the primary mathematical notation is defined below.

Table 1.1: Primary Mathematical Notation

Symbol	**LaTeX**	**Description**
State	$s \in \mathcal{S}$	A specific configuration of the environment. $\mathcal{S}$ is the set of all states.
Action	$a \in \mathcal{A}$	A choice the agent can make. $\mathcal{A}$ is the set of all actions.
Reward	$r \in \mathcal{R}$	The immediate scalar feedback from the environment.
Discount Factor	$\gamma \in [0, 1)$	A scalar determining the present value of future rewards.
Policy	$\pi(a\|s)$	A mapping from states to probabilities of selecting each action.
Transition Prob.	$p(s', r\|s, a)$	Probability of transitioning to state s' with reward r, from state s and
State-Value Func.	$V^\pi(s)$	The expected return starting from state s and following policy π.
Action-Value Func.	$q^\pi(s, a)$	The expected return from state s, taking action a, then following polic
Optimal Funcs.	V^*, q^*	The optimal value functions, associated with the best possible policy.
Bellman Operator	$\mathcal{T}^*$	The Bellman optimality operator.
Infinity Norm	$\|\|q_1 - q_2\|\|_\infty$	The maximum absolute difference between two value functions.

1.2 The Markov Decision Process (MDP) Framework

An MDP provides a complete mathematical description of the environment. For problems with a finite number of states and actions, an MDP is defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$.

- **States ($\mathcal{S}$):** The set of all possible situations the agent can be in. The MDP framework assumes the *Markov Property*, meaning the future is independent of the past given the present state s_t. Formally:

$$P(S_{t+1} = s'|S_t = s, A_t = a) = P(S_{t+1} = s'|S_t = s, A_t = a, S_{t-1}, A_{t-1}, \dots)$$

- **Actions ($\mathcal{A}$):** The set of all possible choices the agent can make.

- **Transition Probability Function (P):** The environment's dynamics model, defined as:

$$p(s', r|s, a) = \Pr(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a)$$

  This joint probability must sum to one: $\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1$ for all $s \in \mathcal{S}, a \in \mathcal{A}$.

- **Reward Function (R):** Defines the goal. The expected immediate reward is:

$$R(s, a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a)$$

- **Discount Factor (γ):** A scalar $\gamma \in [0, 1)$ that balances immediate and future rewards. It ensures that infinite sums of rewards converge and models a preference for earlier rewards. A γ close to 1 implies a "farsighted" agent, while a γ close to 0 implies a "myopic" agent. It can also be interpreted as the probability of the process continuing to the next step.

1.3 The Pursuit of Optimality: Value Functions and Policies

The agent's behavior is formalized by a policy, π, which is a mapping from states to actions.

Definition 1.3.1 (Policy)**.** *A **policy** π specifies the action an agent takes in a given state.*

- *A **deterministic policy** is a function $\pi : \mathcal{S} \to \mathcal{A}$.*
- *A **stochastic policy** $\pi(a|s)$ gives the probability of taking action a in state s.*

The agent's goal is to maximize the *return*, G_t, which is the sum of discounted future rewards:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

We evaluate policies based on the expected return, which leads to the two fundamental value functions.

Definition 1.3.2 (State-Value Function)**.** *The **state-value function** $V^\pi(s)$ is the expected return when starting in state s and following policy π:*

$$V^\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

Definition 1.3.3 (Action-Value Function)**.** *The **action-value function** $q^\pi(s, a)$ (or Q-function) is the expected return after taking action a in state s and then following policy π:*

$$q^\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

These functions are related. The value of a state is the expected value of its action-values, weighted by the policy:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q^\pi(s, a)$$

Part II

The Bellman Equations: A Recursive Framework for Value

Chapter 2

Recursive Relationships for Value Functions

Richard Bellman showed that value functions satisfy a recursive property, decomposing the value of a state into the immediate reward and the discounted value of the successor state.

2.1 The Bellman Expectation Equation

This equation provides the recursive relationship for a value function under a *fixed* policy π. It is used for *policy evaluation*.

For $V^\pi(s)$, the derivation is as follows:

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi[G_t|S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} p(s', r|s, a) \left(r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma V^\pi(s')\right]
\end{aligned}$$

This gives a system of $|\mathcal{S}|$ linear equations in $|\mathcal{S}|$ variables.

Similarly, for the action-value function $q^\pi(s,a)$:

$$\begin{aligned}
q^\pi(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a) \left(r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right) \\
&= \sum_{s',r} p(s',r|s,a) \left(r + \gamma V^\pi(s')\right) \\
&= \sum_{s',r} p(s',r|s,a) \left[r + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s') q^\pi(s',a')\right]
\end{aligned}$$

2.2 The Bellman Optimality Equation: The Principle of Optimality

The ultimate goal is to find the best policy, π^*. An optimal policy is one that is better than or equal to all other policies, i.e., $V^{\pi^*}(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$ and all policies π.

Definition 2.2.1 (Optimal Value Functions)**.** *The* ***optimal state-value function*** $V^*(s)$ *is the maximum possible value from state s:*

$$V^*(s) = \max_\pi V^\pi(s)$$

The ***optimal action-value function*** $q^*(s,a)$ *is the maximum value from taking action a in state s and then acting optimally:*

$$q^*(s,a) = \max_\pi q^\pi(s,a)$$

The Bellman optimality equation defines a condition that these optimal value functions must satisfy. It embodies the *Principle of Optimality*: an optimal policy's sub-policies must themselves be optimal for the subproblems they address.

The derivation for $q^*(s,a)$ is critical:

$$\begin{aligned}
q^*(s,a) &= \max_\pi \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a) \left[r + \gamma \max_\pi \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\right] \\
&= \sum_{s',r} p(s',r|s,a) \left[r + \gamma V^*(s')\right]
\end{aligned}$$

Since an optimal policy will choose the best action in the next state s', we know that

$V^*(s') = \max_{a' \in \mathcal{A}} q^*(s', a')$. Substituting this gives the final form:

$$q^*(s,a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a' \in \mathcal{A}} q^*(s', a') \right] \tag{2.1}$$

The profound difference from the expectation equation is the replacement of the policy-weighted average ($\sum_{a'} \pi(a'|s')$) with a direct maximization ($\max_{a'}$). This turns the problem into solving a system of non-linear equations.

Part III

The Mathematical Machinery of Convergence: Fixed-Point Theory

Chapter 3

Fixed Points and Contraction Mappings

3.1 The Concept of a Fixed Point

To guarantee that a solution to the Bellman optimality equation exists and is unique, we turn to fixed-point theory.

Definition 3.1.1 (Fixed Point)**.** *For an operator $\mathcal{T}$ that maps a set X to itself ($\mathcal{T} : X \to X$), a point $x \in X$ is a **fixed point** if it satisfies the equation:*

$$\mathcal{T}(x) = x$$

Intuition. A fixed point is a point of equilibrium under a transformation. For a real function $f(x)$, it's where the graph of $y = f(x)$ intersects the line $y = x$. A common method for finding fixed points is *fixed-point iteration*, $x_{k+1} = \mathcal{T}(x_k)$. For this iterative method to be reliable, we need to impose stricter conditions on the operator $\mathcal{T}$.

3.2 Contraction Mappings: The Principle of Shrinking Spaces

The condition that guarantees convergence is that the operator must be a *contraction mapping.* This concept is defined within a *metric space.*

Definition 3.2.1 (Metric Space)*. A **metric space** (X, d) is a set X paired with a distance function (or metric) $d : X \times X \to \mathbb{R}$ that satisfies four properties for all $x, y, z \in X$:*

1. ***Non-negativity:*** *$d(x, y) \geq 0$.*

2. ***Identity of Indiscernibles:*** *$d(x, y) = 0 \iff x = y$.*

3. ***Symmetry:*** *$d(x, y) = d(y, x)$.*

4. ***Triangle Inequality:*** *$d(x, z) \leq d(x, y) + d(y, z)$.*

Definition 3.2.2 (Contraction Mapping). *A mapping $\mathcal{T} : X \to X$ on a metric space (X, d) is a **contraction mapping** if there exists a constant α, called the contraction constant, such that $0 \leq \alpha < 1$ and for all $x, y \in X$:*

$$d(\mathcal{T}(x), \mathcal{T}(y)) \leq \alpha d(x, y)$$

Intuition. A contraction mapping uniformly shrinks the distance between any two points in the space. The condition $\alpha < 1$ is critical; it ensures the mapping actively pulls all points closer together, which is essential for guaranteeing convergence to a unique fixed point.

3.3 The Banach Fixed-Point Theorem

This theorem is the central result that connects contraction mappings to the existence and uniqueness of fixed points. It requires one final concept: completeness.

Definition 3.3.1 (Complete Metric Space). *A metric space (X, d) is **complete** if every Cauchy sequence in X converges to a limit that is also in X.*

Intuition. A complete space has no "holes." The set of real numbers $\mathbb{R}$ is complete, but the set of rational numbers $\mathbb{Q}$ is not. The space of all bounded functions, where our value functions reside, is a complete metric space.

Theorem 3.3.1 (Banach Fixed-Point Theorem). *Let (X, d) be a non-empty complete metric space and let $\mathcal{T} : X \to X$ be a contraction mapping on X. Then, $\mathcal{T}$ has exactly one fixed point $x^* \in X$.*

Proof. The proof is constructive and proceeds in two parts.

Existence

1. **Construct a sequence:** Pick an arbitrary $x_0 \in X$ and define the sequence $x_{n+1} = \mathcal{T}(x_n)$ for $n \geq 0$.

2. **Show the sequence is Cauchy:** The distance between successive terms shrinks geometrically:

$$d(x_{n+1}, x_n) = d(\mathcal{T}(x_n), \mathcal{T}(x_{n-1})) \leq \alpha d(x_n, x_{n-1}) \leq \cdots \leq \alpha^n d(x_1, x_0)$$

   Using the triangle inequality for arbitrary $m > n$:

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (\alpha^{m-1} + \cdots + \alpha^n) d(x_1, x_0) \\ &\leq \alpha^n (\alpha^{m-n-1} + \cdots + 1) d(x_1, x_0) \\ &\leq \alpha^n \frac{1}{1-\alpha} d(x_1, x_0) \end{aligned}$$

   Since $\alpha < 1$, as $n \to \infty$, $\alpha^n \to 0$. Thus, the distance can be made arbitrarily small, proving (x_n) is a Cauchy sequence.

3. **Show the limit is a fixed point:** Since X is complete, the sequence converges to a limit $x^* \in X$. We show x^* is a fixed point:

$$\begin{aligned} d(x^*, \mathcal{T}(x^*)) &\leq d(x^*, x_{n+1}) + d(x_{n+1}, \mathcal{T}(x^*)) \\ &= d(x^*, x_{n+1}) + d(\mathcal{T}(x_n), \mathcal{T}(x^*)) \\ &\leq d(x^*, x_{n+1}) + \alpha d(x_n, x^*) \end{aligned}$$

   As $n \to \infty$, both terms on the right go to 0. Thus, $d(x^*, \mathcal{T}(x^*)) = 0$, which implies $x^* = \mathcal{T}(x^*)$.

Uniqueness

Assume there are two distinct fixed points, x^* and y^*. Then $d(x^*, y^*) > 0$.

$$\begin{aligned} d(x^*, y^*) &= d(\mathcal{T}(x^*), \mathcal{T}(y^*)) && \text{(since they are fixed points)} \\ &\leq \alpha d(x^*, y^*) && \text{(by definition of contraction)} \end{aligned}$$

This implies $(1-\alpha)d(x^*, y^*) \leq 0$. Since $\alpha < 1$, we have $(1-\alpha) > 0$. And since $d(x^*, y^*) > 0$, their product must be positive. This is a contradiction. Therefore, the fixed point must be unique. □

Remark. The constructive proof of existence is not merely abstract; it is the blueprint for the **Value Iteration** algorithm.

Part IV

Synthesis: Proving the Guarantees of Value-Based RL

Chapter 4

The Bellman Operator as a Contraction

4.1 The Bellman Optimality Operator as a Fixed-Point Problem

We now cast the problem of solving the Bellman optimality equation as a fixed-point problem.

- **The Space:** The set of all bounded action-value functions $q(s, a)$, which is a complete metric space under the infinity norm metric.
- **The Metric:** The infinity norm, which measures the largest difference between two value functions:

$$d(q_1, q_2) = \|q_1 - q_2\|_\infty = \max_{s \in \mathcal{S}, a \in \mathcal{A}} |q_1(s, a) - q_2(s, a)|$$

- **The Operator:** We define the **Bellman Optimality Operator**, $\mathcal{T}^*$, which takes a value function q and returns a new one, $(\mathcal{T}^* q)$:

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

The optimal value function q^* is, by definition, the unique fixed point of this operator: $\mathcal{T}^* q^* = q^*$.

4.2 Proof of Contraction for the Bellman Operator

The lynchpin of the entire argument is proving that $\mathcal{T}^*$ is a contraction mapping with respect to the infinity norm. We must show that $\|\mathcal{T}^* q_1 - \mathcal{T}^* q_2\|_\infty \leq \gamma \|q_1 - q_2\|_\infty$.

Proof. Let q_1 and q_2 be two arbitrary action-value functions.

$$\begin{aligned}
\|\mathcal{T}^* q_1 - \mathcal{T}^* q_2\|_\infty &= \max_{s,a} |(\mathcal{T}^* q_1)(s,a) - (\mathcal{T}^* q_2)(s,a)| \\
&= \max_{s,a} \left| \sum_{s',r} p(s',r|s,a) \left[r + \gamma \max_{a'} q_1(s',a') \right] - \sum_{s',r} p(s',r|s,a) \left[r + \gamma \max_{a'} q_2(s',a') \right] \right| \\
&= \max_{s,a} \left| \gamma \sum_{s',r} p(s',r|s,a) \left(\max_{a'} q_1(s',a') - \max_{a'} q_2(s',a') \right) \right| \\
&\leq \max_{s,a} \gamma \sum_{s',r} p(s',r|s,a) \left| \max_{a'} q_1(s',a') - \max_{a'} q_2(s',a') \right|
\end{aligned}$$

We use the general property that $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$.

$$\begin{aligned}
&\leq \max_{s,a} \gamma \sum_{s',r} p(s',r|s,a) \max_{a'} |q_1(s',a') - q_2(s',a')| \\
&\leq \max_{s,a} \gamma \sum_{s',r} p(s',r|s,a) \underbrace{\max_{s'',a''} |q_1(s'',a'') - q_2(s'',a'')|}_{\text{This is } \|q_1 - q_2\|_\infty} \\
&= \max_{s,a} \gamma \|q_1 - q_2\|_\infty \sum_{s',r} p(s',r|s,a)
\end{aligned}$$

Since $\sum_{s',r} p(s',r|s,a) = 1$, we have:

$$\begin{aligned}
\|\mathcal{T}^* q_1 - \mathcal{T}^* q_2\|_\infty &\leq \max_{s,a} \gamma \|q_1 - q_2\|_\infty \\
&= \gamma \|q_1 - q_2\|_\infty
\end{aligned}$$

Since $\gamma \in [0,1)$, we have proven that the Bellman Optimality Operator $\mathcal{T}^*$ is a γ-contraction.

□

Remark. Proving contraction in the infinity norm guarantees *uniform convergence.* This means the error shrinks across all state-action pairs simultaneously, which is a very strong and desirable property.

4.3 The Convergence of Value Iteration

We can now provide a full proof for the convergence of the Value Iteration algorithm.

Definition 4.3.1 (Value Iteration Algorithm). *1.* ***Initialization:*** *Initialize $q_0(s, a)$ arbitrarily for all $s \in \mathcal{S}, a \in \mathcal{A}$ (e.g., to all zeros).*

2. ***Iteration:*** *For $k = 0, 1, 2, \dots,$ update the value function for all state-action pairs:*

$$q_{k+1}(s, a) \leftarrow \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a' \in \mathcal{A}} q_k(s', a')\right]$$

3. ***Termination:*** *Stop when $\|q_{k+1} - q_k\|_\infty < \epsilon$ for some small tolerance ϵ.*

It is clear that the update rule is simply the repeated application of the Bellman Optimality Operator: $q_{k+1} \leftarrow \mathcal{T}^* q_k$.

We can now state the final, conclusive argument:

1. The space of bounded action-value functions with the infinity norm metric is a **complete metric space**.

2. The Bellman Optimality Operator, $\mathcal{T}^*$, is a **contraction mapping** on this space with contraction constant $\gamma < 1$.

3. **Conclusion from Banach's Theorem:** Therefore, $\mathcal{T}^*$ has **one and only one fixed point**, which is the optimal action-value function, q^*.

4. **Final Implication:** The theorem further guarantees that the sequence $q_{k+1} = \mathcal{T}^* q_k$, which is precisely the sequence generated by the Value Iteration algorithm, is **guaranteed to converge** to this unique fixed point q^*, regardless of the initial estimate q_0.

This provides a definitive and positive answer to all three questions posed at the outset: the optimal value function **exists**, it is **unique**, and it is **computable**.

Part V

Conclusion and Broader Implications

Chapter 5

Final Remarks

5.1 Summary of Theoretical Guarantees

This report has systematically constructed the theoretical foundation that guarantees the correctness and convergence of value-based reinforcement learning. The logical progression was as follows:

1. The problem was formalized as finding an optimal policy in a Markov Decision Process (MDP).

2. The value of this policy, q^*, was shown to satisfy the Bellman Optimality Equation.

3. Solving this equation was reframed as finding a fixed point for the Bellman Optimality Operator, $\mathcal{T}^*$.

4. We proved that $\mathcal{T}^*$ is a contraction mapping on the complete metric space of value functions.

5. The Banach Fixed-Point Theorem was invoked to guarantee that a unique fixed point (q^*) exists and that the iterative sequence generated by Value Iteration ($q_{k+1} = \mathcal{T}^* q_k$) converges to it.

5.2 Implications for Reinforcement Learning

These theoretical guarantees have profound implications for the field.

- **Foundation of Trust:** They assure us that methods like Value Iteration are not just heuristics but are principled algorithms grounded in solid mathematical theory, guaranteed to converge to the optimal solution.

- **Unambiguous Target:** The uniqueness of q^* provides a clear target for learning. Once q^* is found, the optimal policy π^* can be determined greedily and without ambiguity:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} q^*(s, a)$$

- **Bridge to Advanced Algorithms:** The core concepts of Bellman updates and contraction properties are central to model-free algorithms like Q-learning and SARSA, whose convergence proofs build upon this foundational analysis.

- **Generalization:** The abstract nature of the framework—operators on function spaces—lends itself to generalization beyond finite MDPs to continuous spaces and other dynamic optimization problems in economics, control engineering, and operations research.

This unifying perspective highlights the deep and elegant mathematical structure that underlies the quest for intelligent, autonomous decision-making.