Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 9(12): 33818–33842.
DOI: 10.3934/math.20241613
Received: 04 August 2024
Revised: 11 November 2024
Accepted: 19 November 2024
Published: 28 November 2024

Research article

A study of value iteration and policy iteration for Markov decision processes in Deterministic systems

Haifeng Zheng* **and Dan Wang**

School of Economics, Jinan University, Guangzhou 510632, Guangdong, China

* **Correspondence:** Email: 15892732432@163.com; Tel: +8615892732432.

Abstract: In the context of deterministic discrete-time control systems, we examined the implementation of value iteration (VI) and policy (PI) algorithms in Markov decision processes (MDPs) situated within Borel spaces. The deterministic nature of the system's transfer function plays a pivotal role, as the convergence criteria of these algorithms are deeply interconnected with the inherent characteristics of the probability function governing state transitions. For VI, convergence is contingent upon verifying that the cost difference function stabilizes to a constant k ensuring uniformity across iterations. In contrast, PI achieves convergence when the value function maintains consistent values over successive iterations. Finally, a detailed example demonstrates the conditions under which convergence of the algorithm is achieved, underscoring the practicality of these methods in deterministic settings.

Keywords: Markov decision processes; Deterministic system; value iteration; policy iteration; average cost criterion
Mathematics Subject Classification: 60J05, 60J10

1. Introduction

A Markov chain, introduced by the Soviet mathematician Andrey Markov in the early 20th century, is a class of stochastic processes characterized by the property that future states depend only on the current state, independent of the sequence of preceding states. Early applications of Markov decision theory concentrated on solving optimal decision problems in deterministic settings. However, decision-making in real-world scenarios typically involves significant uncertainty and risk, prompting the expansion of decision theory frameworks to encompass stochastic environments. Markov decision processes (MDPs) provide a structured approach for stochastic optimization in control systems. In cases where the transition probability is degenerate, resulting in a deterministic distribution, the control system is termed a deterministic system, a concept rooted in the work of Bellman [1].

Within deterministic systems, the transition dynamics can be explicitly defined by a transition function, situating these systems within a specialized domain of stochastic Markov decision theory where state evolution follows deterministic rules. The foundational theory of discrete-time Markov decision processes (MDPs), as outlined by Hernandez and Lasserre [2, 3], includes a comprehensive treatment of random optimization problems under both discounted and non-discounted criteria. Section 1.2 of their work provides a detailed exploration of deterministic systems, though certain critical conclusions may not universally hold, as noted in further discussions by Hernandez et al. [4]. Earlier work by Meyn and Tweedie [5] recharacterized deterministic systems as nonlinear systems, positing the transition function as a nonlinear mapping—a perspective demonstrated in examples provided by Hernandez et al. [4]. While many real-world MDPs involve optimization with stochastic (or "random") transition probabilities, problems featuring deterministic transition probabilities are applicable across various domains, including linear programming and certain economic models [6, 7]. These applications highlight the practical breadth of deterministic optimization within stochastic frameworks, underscoring its relevance in systems with known transition dynamics. Building on this framework, Hernandez et al. [4] examined the average cost optimization problem for discrete-time MDPs in deterministic systems. Their analysis focused on establishing the conditions for the existence of an average cost criterion, operating under the assumptions that both the state and action spaces are Borel spaces and that all costs are non-negative. To this end, they employed three distinct approaches: the average cost inequality, the stationary method, and the vanishing discount method. Notably, the literature does not extend to discussing two approximate processes applicable to dynamic programming within deterministic systems: the value iteration (VI) process and the policy iteration (PI) process.

In these three approaches, the average cost inequality method facilitates a comparative analysis of strategies by evaluating expected returns. This framework enables decision-makers to determine optimal strategies through the calculation of expected outcomes across various decisions. When state transitions and rewards are stochastic, the average cost inequality serves as a tool to assess the expected returns across potential future states, thereby addressing inherent uncertainties. For deterministic systems, however, the average cost inequality offers a direct method for comparing deterministic returns, leveraging the known state transition functions to provide clear strategic insights. The steady-state approach, or stable strategy method, aims to identify policies under which the system's state distribution converges to a steady state over time in specific Markov decision processes. In this equilibrium, returns across different states stabilize, emphasizing policy consistency and clarity within deterministic systems, where predictable state transitions are key to achieving stable returns in the long term. The vanishing discount method, on the other hand, incorporates time preference by discounting future returns, thus emphasizing their present value. Through a discount factor—typically set between 0 and 1—future returns gradually diminish in influence, aligning decisions with immediate priorities. In deterministic systems, a discount factor close to 1 underscores the value of long-term returns, promoting decision-making that is stable, consistent, and geared toward sustained effectiveness.

1.1. Where do the ways of implementation originate from?

The two primary approaches proposed here aim to address dynamic programming challenges by enhancing computational efficiency. Before establishing their validity, it is essential to review the underlying dynamic programming framework. Building on the work of Hernandez and Lasserre [2, 3] and Gaitsgory et al. [8], the computation of the cost function transitions from a sequence of integral functions to a summation of costs over specific state-action pairs. This reformulation significantly improves the efficiency of iterative procedures and provides a foundation for examining the validity of the value iteration (VI) process in deterministic systems.

Costa and Dufour [9] extended this approach by deriving new results for the average cost criterion under the Feller property of transition probabilities, thus ensuring policy iteration (PI) convergence under these conditions. However, they noted that the Feller property alone does not fully guarantee PI validity within deterministic frameworks. In their verification, Meyn [10] demonstrated that PI has a linear convergence rate, effectively reflecting its convergence properties. This prompts consideration of whether similar methods can be employed to validate these approaches fully.

When analyzing the VI and PI processes under the average cost criterion, they find application in both finite-level cost growth assessments and in operations research at the infinite-horizon level, as initially explored by Bellman [11] and Howard [12]. On the topic of discount factors, Dai and Menoukeu [13] investigated the optimal stopping problem for value functions under stochastic discounting, with relevance to pulse control problems. Feinberg and Liang [14] further explored scenarios where the discount factor approaches 1, thereby approximating deterministic behavior. Using a similar approach, Yu et al. [15] confirmed the existence of an optimal value in game-theoretic settings under these conditions.

In deterministic systems, discussions typically adopt an offline approach that applies to all cases where the transition function is deterministic, regardless of whether these functions are linear or nonlinear. This contrasts with the online policy iteration method proposed by He et al. [16], specifically targeting nonlinear systems. He et al.'s work emphasizes real-time, dynamic processes, using policy iteration to forecast future outcomes under uncertain conditions by approximating nonlinear problems as linear ones. Similarly, the online reinforcement learning study by Fang et al. [17] encounters comparable challenges. Although both approaches are based on Markov decision processes, online reinforcement learning methods do not require prior knowledge of state transitions that are better suited for scenarios where data collection and model training occur concurrently.

In another study by Fang et al. [18], an online fuzzy optimization algorithm based on Markov jump systems demonstrated convergence and introduced a completely model-free, offline policy iteration fuzzy reinforcement learning algorithm. This algorithm effectively controls without relying on system dynamics or transition probability information. This fully model-free approach is fundamentally different from deterministic systems that depend on known system dynamics and transition probabilities.

1.2. Key innovations

This study primarily investigates value iteration (VI) and policy iteration (PI) for deterministic Markov decision processes (DMDPs), introducing their application under the average cost (AC) criterion within deterministic systems for the first time. Section 2 examines the existence of dynamic programming processes within deterministic systems, noting that while Hernandez et al. [4] have extensively explored these systems, many foundational theorems in this specialized domain remain unverified.

In Section 3, we focus on establishing the validity of VI within deterministic systems. If the value function converges under deterministic conditions, this convergence confirms the existence of an optimal value function in DMDPs, thereby enabling the derivation of an optimal policy set. Section 4 addresses the convergence of PI, where each step of policy evaluation and improvement incorporates VI, iteratively refining policies toward optimality.

VI and PI serve as essential methods for analyzing the convergence of decision processes in stochastic systems. Therefore, verifying the validity of these iterations within deterministic contexts is vital to advancing research across the entire field of Markov decision processes.

2. Dynamic programming problems in deterministic systems

DMDPs, as a subfield of stochastic processes, have been discussed by Hernandez and Laura [4] to have the following form for their transition function:

$$x_{t+1} = F(x_t, a_t), t = 0, 1, 2... \tag{2.1}$$

Given the cost function for each stage, the total cost function $c(x, a)$ within time T is expressed as follows:

$$J_T(\lambda, x) := \sum_{t=0}^{T-1} c\left(x_t, a_t\right). \tag{2.2}$$

Equation (2.2) represents the total cost generated by the path taken in the decision-making process under a given initial state x_0 and strategy λ. Due to the infinity of time, the average cost is calculated by taking the limit over time:

$$J(\lambda, x) = \liminf_{T\to\infty} \frac{1}{T} J_T(\lambda, x). \tag{2.3}$$

The convergence of MDPs is discussed by using the convergence of the average cost function.

VI and PI are classic methods for solving dynamic programming problems, first introduced by Howard [12]. The following verifies whether the dynamic programming problem holds in deterministic systems. The specific definition of dynamic programming problems is as follows:

$$J_t(x) := \min\left[c(x, a) + J_{t+1}(F(x, a))\right],$$

for the above equation, the time is calculated by the backward $t = N, N - 1, ..., 1, 0$.

Theorem 1. *Let $J_0, J_1, ..., J_{T-1}, J_T$ be a function on the state space X and from $t = T$ to $t = 0$, if the value function is defined as:*

$$J_T(x) = c_T(x), t = N, N - 1, ..., 1, 0, \tag{2.4}$$

the momentary cost function is:

$$J_t(x) = \min_{a\in A(x)} \left[c(x, a) + J_{t+1}(F(x, a))\right].$$

Then, there must exist a set of strategies Λ_t that $\lambda_t(x) \in \Lambda(x)$ contains the minimum value in the above equation, i.e.:

$$J_t(x) = c(x, a) + J_{t+1}(F(x, a)),$$

and for each stage, there will exist a deterministic policy $\lambda^* = \{\lambda_1, \lambda_2, \ldots, \lambda_{t-1}\}$ such that the following equation holds:

$$J^*(x) = J_0(x) = J\left(f^*, x\right). \tag{2.5}$$

Proof of Theorem 1. Let $\lambda = \lambda_t$ be the set of fixed strategies, and C_t be the value of the cost from the time T to the time t, and given the state $x_t = x$ at the moment t, if $t = 0, 1, ..., T - 1$.

$$C_t(\lambda, x) = \sum_{n=t}^{T-1} c\left(x_n, a_n\right) + c_T\left(x_T\right), \tag{2.6}$$

$$C_T(\lambda, x) = c_T\left(x_T\right) = c_T(x), x_T = x, \tag{2.7}$$

$$\begin{aligned} J(\lambda, x) &= C_0(\lambda, x) \\ &= \sum_{n=0}^{T-1} c\left(x_n, a_n\right) + c_T\left(x_T\right). \end{aligned}$$

Assuming that for all $t = 0, 1, ..., T$

$$C_t(\lambda, x) \geq J_t(x) \tag{2.8}$$

holds. When $\lambda = \lambda^*$,

$$C_t\left(\lambda^*, x\right) = J_t(x) \tag{2.9}$$

holds. In particular when taken to $t = 0$, for any state x, there are

$$J(\lambda, x) \geq J_0(x) and J\left(\lambda^*, x\right) = J_0(x).$$

If the hypothesis holds, the desired result Eq (2.5) is obtained.

Now, verify Eqs (2.8) and (2.9): When $t = T$, it is obtained from Eqs (2.4) and (2.7):

$$C_T(\lambda, x) = J_T(x) = c_T(x).$$

The inductive hypothesis method: if for $t = T - 1, ..., 1, 0$:

$$C_{t+1}(\lambda, x) \geq J_{t+1}(x), x \in X$$

holds, then for Eq (2.7).

$$\begin{aligned} C_t(\lambda, x) &= c(x, a) + \sum_{n=t+1}^{T-1} c\left(x_n, a_n\right) + c_T\left(x_T\right) \\ &\geq \min_{a \in A(x)} \left[c(x, a) + J_{t+1}(F(x, a))\right] \\ &= J_t(x). \end{aligned}$$

So there is:

$$C_t(x) \geq J_t(x).$$

If $t = 0$, then there is:

$$\begin{aligned} J(\lambda, x) \geq J_0 :&= \min_{a \in A(x)} \left[c(x, a) + J_1(F(x, a))\right] \\ &= \min_{a \in A(x)} \left[c(x, a) + c\left(x_1, a_1\right) + J_2\left(F\left(x_1, a_1\right)\right)\right] \\ &= \min_{a \in A(x)} \left[\sum_{t=0}^{T-1} c\left(x_t, a_t\right) + J_T(x)\right]. \end{aligned}$$

The equation holds when the set of optimal strategies is taken. □

It follows from the above proof that DP holds under deterministic systems.

3. Value iteration in deterministic systems

In deterministic systems, dynamic programming problems offer a clearer and more efficient framework for applying value iteration and policy iteration, facilitating a faster determination of value functions or optimal policy sets. Section 3 centers on an in-depth examination of value iteration, exploring its convergence properties and implications for identifying optimal solutions.

3.1. The basic idea of value iteration

In the deterministic system, the cost function is shown in Eq (2.2), and the AC is shown in Eq (2.3), if we make:

$$J_n^*(x) =: \min_{a \in A(x)} J_n(\lambda, x).$$

For any $n \geq 1$, representing the optimal cost of the n-th stage and the:

$$J_0^*(x) = 0,$$

$$J_n^*(x) = \min_{a \in A(x)} \left[c(x, a) + J_{n-1}^*(F(x, a)) \right].$$

Definition 1.

$$\begin{aligned} x_{t+1} = F\left(x_t, a_t\right) &= F\left(F\left(x_{t-1}, a_{t-1}\right), a_t\right) \\ &= \underbrace{F\left(F\left(F \ldots (F(x, a), a_1)\right), a_t\right)}_{t}, \end{aligned}$$

$$F\left(F(F, \ldots (F(x, a))), a_t\right) =: F^{t+1}\left(x, (a_t, \ldots, a)\right).$$

Assumption 1. *The OCP (2.1)–(2.3) satisfies:*

(1) The value function for each stage t in sequence $\{J_n^*(x)\}$ is equicontinuous;
(2) There exists a continuous function F such that $x_{n+1} = F\left(x_n, a_n\right)$;
(3) If there exists a termination cost function $l(x)$ that satisfies:

$$-N \leq l(x) \leq b(x),$$

$b(x)$ is any bounded non-negative function with:

$$b\left(F\left(x, (\lambda_{n-1}, \lambda_{n-2}, \ldots, \lambda_0)\right)\right) \leq L(x),$$

$L(x)$ for a mappingX $\rightarrow R$ hold.

In the Borel space, consider a fixed state z and define the following equations:

$$p_n(x) := J_n^*(x) - J_n^*(z),$$

$$q_n(x) := J_n^*(x) - J_{n-1}^*(x).$$

The following conditions are given:

$$p_n(x) \rightarrow l(x), q_n(x) \rightarrow \rho^* \quad \forall x \in X,$$

where $(\rho^*, l(x))$ is the solution to the AC equation.

The purpose of VI is to find λ_n such that the following equation holds true:

$$J_n^*(x) = c(x, \lambda_n) + J_{n-1}^*(F(x, \lambda_n)).$$

Introduce the canonical triplet$(\rho^*, l(x), \lambda^*)$ to satisfy the following equation:

$$\rho^* + l(x) = c(x, a) + l^*(F(x, \lambda^*)). \tag{3.1}$$

In Eq (3.1), ρ^* denotes the average optimal cost, $l(x)$ denotes a termination cost function, and λ^* denotes the optimal strategy.

3.2. Value iteration processes

Theorem 2. *For the difference function:*

$$h(x) := J_n^* + l(x_n) - n\rho^*, x \in X. \tag{3.2}$$

For all $x \in X$ and $n \geq 0$, the following three inequalities hold:

(1) $h(x) \geq -N$;
(2) $h_n(F^m(x, (\lambda_{n+m-1}, \lambda_{n+m-2}, \dots, \lambda_n)) \leq h_{n+m}(x)$;
(3) $h_n(x) \leq h_{n-1}(F(x, \lambda_n))$.

Proof of Theorem 2. (1) By

$$\begin{aligned} J_t(\lambda, x, l) &= \sum_{t=0}^{T-1} c(x_t, a_t) + l(x_T) \\ &= J_t(\lambda, x) + l(x_T). \end{aligned}$$

Also due to $-N \leq l(x) \leq b(x)$:

$$J_n^*(x, l) := \min_{a \in A(x)} J_n(\lambda, x, l),$$

when $l(x) \equiv 0$ is available,

$$J_n^*(x) := \min_{a \in A(x)} J_n(\lambda, x)$$

is the general cost function:

$$J_n(\lambda^\infty, x, l) = J_n^*(x, l) = n\rho + l(x).$$

When $\rho = \rho^*$,:

$$n\rho^* + l(x) \geq J_n^*(x) + l(x) \geq J_n^*(x) - N,$$

therefore

$$n\rho^* + l(x) - J_n^*(x) \geq -N.$$

(2)

$$\begin{aligned} h_n(F(x, \lambda)) =& n\rho^* + l(F(x, \lambda)) - J_n^*(F(x, \lambda)) \\ &\leq n\rho^* + J(F(x, \lambda)) + c(x, \lambda) - J_{n+1}^*(x) \\ &\leq (n+1)\rho^* + l(x) - J_{n+1}^*(x). \end{aligned}$$

Therefore, we get $h_n(F(x, \lambda)) \leq h_{n+1}(x)$, which is the same as

$$h_n(F^m(x, (\lambda_{n+m-1}, \lambda_{n+m-2}, \dots, \lambda_t))) \leq h_{n+m}(x).$$

(3)

$$\begin{aligned}\rho^* + l(x) \leq & c\left(x, \lambda_n\right) + l\left(F\left(x, \lambda_n\right)\right) \\ & = J_n^*(x) + h\left(F\left(x, \lambda_n\right)\right) - J_{n-1}^*\left(F\left(x, \lambda_n\right)\right),\end{aligned}$$

$$\rho^* + l(x) - J_n^*(x) \leq h\left(F\left(x, \lambda_n\right)\right) - J_{n-1}^*\left(F\left(x, \lambda_n\right)\right),$$

$$n\rho^* + l(x) - J_n^*(x) \leq (n-1)\rho^* + h\left(F\left(x, \lambda_n\right)\right) - J_{n-1}^*\left(F\left(x, \lambda_n\right)\right)x,$$

$$h_n(x) \leq h_{n-1}\left(F\left(x, \lambda_\lambda\right)\right).$$

□

Theorem 3. *For the sequnce $\{h_n(x)\}$:*

(1) $\{h_n(x)\}$ is uniformly continuous and bounded;
(2) For each sub-sequence $\{h_{n_i}(x)\}$ there exists a constant k such that

$$\lim_{i\to\infty} h_{n_i}(x) = k, \quad \forall x \in X;$$

(3) If any subsequence $\{h_{n_i}(x)\}$ converges to the unique constant k, then $\{h_n(x)\}$ converges to the constant k.

Proof of Theorem 3. (1) For any $x_1, x_2 \in X$, from Eq (3.3) we can obtain:

$$\left|h_n\left(x_1\right) - h_n\left(x_2\right)\right| \leq \left|l\left(x_1\right) - l\left(x_2\right)\right| + \left|J_n^*\left(x_1\right) - J_n^*\left(x_2\right)\right|,$$

due to the uniform continuity of the function J_n^*, it is bounded. Furthermore,

$$h_n(x) \leq h_0\left(F^n\left(x, \lambda_{n-1}, \ldots, 0\right)\right),$$

since: h_0 is $l_0(x)$, and $l(x)$ is finite, we have

$$h_n(x) \leq h_0\left(F^n\left(x, \lambda_{n-1}, \ldots, 0\right)\right) \leq L(x),$$

thus:

$$-N \leq h_n(x) \leq L(x), \quad \forall x \in X.$$

(2) Assuming $\{h_{n_i}(x)\}$ is a subsequence of $\{h_n(x)\}$ by (1) and Ascoli's theorem [2, 3], it is known that as $i \to \infty$ converges to a continuous function θ,

$$h_{n_i}(x) \to \theta(x), \forall x \in X, \tag{3.3}$$

$$h_{n+m}\left(F^n\left(x, \left(\lambda_{n+m-1}, \ldots, \lambda_n\right)\right)\right) \leq h_{n+m}(x).$$

When n is fixed and $m \to \infty$,

$$h_{n+m}(x) \to \theta(x)$$

for deterministic systems, the probability P values are either 0 or 1. As $P\left(F^m(x, a)\right) = 1$, the VI converges and satisfies

$$\sum h_n P \leq \lim_X \min h_n\left(F^m\left(x_n, a_n\right)\right) \leq \theta(x). \tag{3.4}$$

By Fatou's lemma:

$$\sum \theta(x)P \leq \lim \min_X h_{n+m} \left(F\left(x_{n+m}, a_{n+m}\right)\right) \leq \theta(x),$$

$$\sum \theta(x)P \leq \theta(x),$$

let

$$k := \min_X \theta(x)$$

be such that

$$k \leq \sum \theta(x)P \leq k,$$

thus

$$\sum \theta(x)P = k.$$

Furthermore, due to the arbitrariness of P we know that $\theta(x) = k$. Since

$$\lim_{i\to\infty} h_{t_i}(x) = k,$$

if $\{h_{n_i}(x)\}$ is assumed not to converge to k, implying the existence of $\varepsilon > 0$ such that the subsequence $\{h_{n_i}(x)\}$ converges in the following manner:

$$\left|h_{n_i}(x) - k\right| > \varepsilon.$$

Contradicting (2).
(3) If (2) holds, for another sub-sequence $\left\{h_{n_j}(x)\right\}$ of $\{h_n(x)\}$, as in the proof of (2), if $\left\{h_{n_j}(x)\right\}$ converges to k', then from Eqs (3.3) and (3.4) we can deduce that $k \geq k'$, $k \leq k'$ and for all states x,

$$h_n(x) \to \theta(x)$$

holds. Otherwise, the sequence $\{h_n(x)\}$ and $\left\{x_j\right\}$ must converge to

$$\left|h_{n_j}\left(x_j\right) - k\right| > \varepsilon, \quad \forall i \geq 1$$

contradicts the condition of 3, so

$$\lim_{n\to\infty} h_n(x) = k.$$

□

Theorem 4. *(Theorem of VI Convergence): With the above assumptions holding, the value iteration processes are convergent, i.e., $p_n(x)$ converges to $l(x)$ and $q_n(x)$ converges to ρ^* and the following holds:*

$$\lim_{n\to\infty} p_n(x) = l(x),$$

$$\lim_{n\to\infty} q_n(x) = \rho^*.$$

Proof of Theorem 4. For the canonical triplet (ρ^*, l, λ) that solves the average optimal equation (3.1), and when $n \to \infty$, the value of the deviation function converges to a constant k . Hence, for

$$q_n(x) = \rho^* - h_n(x) + h_{n-1}(x),$$

$$p_n(x) = l(x) - h_n(x) + h_n(z),$$

satisfy:

$$\lim_{n\to\infty} p_n(x) = l(x),$$

$$\lim_{n\to\infty} q_n(x) = \rho^*.$$

Therefore, the VI converges. □

4. Policy iteration in deterministic systems

In Section 3, the convergence of value iteration (VI) within deterministic systems is confirmed under specific assumptions. Once the optimal value function is obtained in practical applications, the corresponding optimal policy can then be derived. Policy iteration (PI) is generally employed to identify the optimal policy through an iterative approach that alternates between policy evaluation and policy improvement, progressively refining the policy until optimality is achieved. Each step in the PI process involves evaluating the policy using the value function and improving the policy via the selection function. Section 4 delves into the application and convergence of PI within deterministic systems, emphasizing its effectiveness in obtaining optimal decision policies.

4.1. The basic idea of policy iteration

For the initial stage of the policy λ, there exists a cost function $V(\lambda_0, \cdot)$ related to the discount factor α:

$$V_\varepsilon(\lambda, x) := \sum_{t=0}^{T_\lambda} \varepsilon^t c\left(x_t, a_t\right),$$

the relationship between its cost function and the AC function Eq (2.3) without the discount factor is:

$$(1-\varepsilon)V_\varepsilon(\lambda, x) = J(\lambda, x) + (1-\varepsilon)\sum_{t=0}^{T_\lambda} \varepsilon^t \left(c\left(x_t, a_t\right) - J(\lambda, x)\right). \tag{4.1}$$

When $\alpha \to 1$, $V \to J$. The cost function is defined $V(\lambda_0, \cdot)$ as $\omega_0(\cdot)$ and it satisfies the following equation:

$$\omega_0(x) = c\left(x, \lambda_0\right) + \varepsilon\omega_0\left(F\left(x, \lambda_0\right)\right). \tag{4.2}$$

A kernel function is defined by Eq (4.1):

$$K_t(x, A) := (1-\varepsilon)\sum_{i=1}^{T_{\lambda_t}-1} \varepsilon^i \left(c\left(x_i, a_i\right) - \rho_t\right). \tag{4.3}$$

Equation (4.3) signifies iterating into the system with a policy λ_0, reaching time t and utilizing this kernel function to reflect the deviation in function values. In deterministic systems, the discount factor $\varepsilon \to 1$. The selection rules for policy iteration are introduced:

$$\lambda_t(x) := \arg\min_{a \in A(x)} \left(c(x,a) + \varepsilon\omega_{t-1}(F(x,a))\right), \tag{4.4}$$

$$\kappa_{t-1} = (x,a).$$

From Eq (4.4), it follows that the policy λ_1 at the first time step leads to:

$$c\left(x,\lambda_1\right) + \varepsilon\omega_0\left(F\left(x,\lambda_1\right)\right) = \min_{a \in A(x)} \left[c(x,a) + \varepsilon\, \omega_0(F(x,a))\right] \tag{4.5}$$

at this point

$$\omega_1(\cdot) := V\left(\lambda_1, \cdot\right).$$

Given the optimal policy λ_t at t-th, the cost function along the path is as follows:

$$V\left(\lambda_t, \cdot\right) =: \omega_t(\cdot),$$

the optimal policy λ_{t+1} at time $t+1$ satisfies the following equation:

$$c\left(x,\lambda_{t+1}\right) + \varepsilon\omega_0\left(F\left(x,\lambda_{i+1}\right)\right) = \min_{a \in A(x)} \left[c(x,a) + \varepsilon\omega_i(F(x,a))\right]. \tag{4.6}$$

Equations (4.2), (4.4)–(4.6) represent the processes of PI in MDPs, where in each iteration, the minimum cost for each stage needs to be calculated to choose the optimal policy based on this minimum cost.

Remark: The function presented in Eq (4.3) draws on concepts introduced by Meyn [10], where it is described as the sum of probability distribution functions in stochastic systems. In deterministic systems, however, the probability distribution values are restricted to either 0 or 1, allowing for integration with the discrepancy function. From Eq (4.3), it becomes evident that the core mechanism of policy iteration involves alternating between policy evaluation and policy improvement through the value iteration method. At each step, this approach enables the selection of an optimal policy based on iterative evaluation.

4.2. Policy iteration processes

Without considering the termination cost $l(x)$, the discrepancy value function $h(\cdot)$ for Eq (3.2) in Section 3 is redefined as follows:

$$H(x) = \sum_{t=0}^{T-1} \bar{c}\left(x_t\right),$$

$$\bar{c} = c - \rho.$$

Theorem 5. *Assuming there exists an initial policy λ_0, and the AC generated by the iteration processes at this point is ρ_0, if the following two conditions hold for the DMDPs:*

(1) The cost function is continuous in product space and there exists a cost function $\underline{c}(x)$ such that for any κ there exists $c(x,a) \geq \underline{c}(x)$;

(2) *There exists a state $\chi \in X$ and a continuous function $s : X \to (0,1)$ satisfying the following conditions:*

When $n \geq 1$, there exists a canonical triplet $(\lambda_{n-1}, H_{n-1}, \rho_{n-1})$, which is generated by the initial policy iteration λ_0 under Eq (4.4), further λ_n obtained by iteration in Eq (4.5) satisfies the following inequality:

$$K_n(x,\chi) \geq s(x), \forall x \in X, \tag{4.7}$$

then for stage n-th, the algorithm will have a solution (λ_n, H_n, ρ_n) and the related value function:

$$H_{\lambda_n}(x) = \sum_{n=0}^{T_{\lambda_n}-1} \bar{c}_{\lambda_n}(x_i, a_i), \quad n \geq 0,$$

is finite and satisfies Proposition 1.

Remark: In stochastic processes, defining a continuous function $s(x)$ [10] ensures that for any state x, the probability of transitioning to another state remains within the range $(0,1)$ until convergence is achieved in the iteration process. The kernel function used in these processes is given by:

$$K_n := (1-\varepsilon)\sum_{n=0}^{\infty} \varepsilon^n P^n,$$

where ε represents a discount factor. In contrast, within deterministic systems, the probability of states participating in the iteration path due to policy selection is precisely 1. Thus, this paper primarily employs the discrepancy value function in place of the transition function used in stochastic processes, as indicated in Eq (4.3).

In the context of deterministic systems, kernel functions play a critical role in assessing the stability and convergence of policy iteration. Specifically, when the result of a kernel function equals zero, it indicates that the selected set of policies fails to consistently follow the same iterative path throughout the policy iteration process. Thus, the kernel function value reflects the consistency and reliability of policies during iteration. Moreover, the construction of the kernel function is crucial for verifying the convergence of policy iteration. In stochastic processes, the kernel function helps determine whether each state can effectively converge to a corresponding state at the boundary, thus ensuring convergence of the entire policy iteration. However, in deterministic systems, the lack of randomness means there is no need to consider boundary state points. Instead, the focus shifts to constructing a kernel function that verifies whether the chosen policy function ensures that the optimal state consistently follows the same iterative path. This simplification renders kernel function analysis in deterministic systems more efficient and straightforward. Additionally, enhancing the construction methods of kernel functions—such as refining their definitions and computational approaches—can improve the performance of policy iteration, thereby offering stronger theoretical support for analyzing deterministic systems.

Property 1. *Difference function: The discrepancy cost function has the following properties:*

$$H_{n-1}\left(F_{\lambda_{n-1}}(x,a)\right) = H_{n-1}(x) - c(x,a) + \rho_{n-1}, \tag{4.8}$$

$$H_n\left(F_{\lambda_n}(x,a)\right) = H_n\left(x_n\right) - \bar{c}\left(x_n\right), \tag{4.9}$$

$$H_{n-1}\left(F_{\lambda_n}(x,a)\right) = H_{n-1}(x) - \bar{c}\left(x_n\right) - \gamma_n, \tag{4.10}$$

$$\bar{c}\left(x_n\right) = c\left(x_n\right) - \rho_n.$$

Equation (4.8) represents a degenerate form of the AC equation, where

$$\kappa_n = (x,a)$$

denotes the feasible state-action pair under the optimal policy at the n-th stage. H_{n-1} represents the deviation function obtained from the $n-1$th stage of PI, with the value of this deviation function being the sum of the differences between the cost function and the AC along the subsequent iteration path.

Equation (4.9) indicates that the discrepancy value under the optimal policy differs from the discrepancy value along any feasible path state at the n-th stage. Equation (4.10) introduces the sequence γ_n , which signifies that as the iteration process reaches the $n-1$-th stage, the current deviation function is evaluated, and the next optimal policy is applied to continue calculating the deviation function along the current path. If the current selection in the iteration process fails to lead to convergence, even after selecting the next optimal policy, deviations may persist that the system cannot accept. In this context, the discrepancy γ describes the potential for convergence within the iterative process.

From Eq (4.4):

$$c\left(F_{\lambda_n}(x,a)\right) + H_{n-1}\left(F_{\lambda_n}(x,a)\right) \leq c(x,a) + H_{n-1}(x,a). \tag{4.11}$$

The transformed form of Eq (4.8) is as follows:

$$c(x,a) + H_{n-1}\left(F_{\lambda_n}(x,a)\right) = H_{n-1}(x) + \rho_{n-1}, \tag{4.12}$$

By combining Eqs (4.11) and (4.12), we obtain:

$$H\left(F_{\lambda_n}(x,a)\right) + c\left(F_{\lambda_n}(x,a)\right) \leq H_{n-1}(x) + \rho_{n-1};$$

$$H(x) - \bar{c}\left(F_{\lambda_n}(x,a)\right) - \gamma_n + c\left(F_{\lambda_n}(x,a)\right) \leq H(x) + \rho_{n-1}.$$

Therefore, it is not difficult to derive:

$$\gamma_n \geq \rho_n - \rho_{n-1}. \tag{4.13}$$

If the sequence $\{\rho_n\}$ is monotonically decreasing, then there exists a lower bound 0 for $\{\gamma_n\}$ and also a lower bound for the sequence $\{H_n\}$.

Proposition 1. *(1) Uniform boundedness: For some constants* $0 < N < \infty$ *such that*

$$\inf_{x\in X, n\geq 0} H\left(x_n\right) > -N;$$

(2) Almost monotonicity:

There exists a sequence of functions $\{g_n : n \geq 0\}$ *such that:*

$$g_n(x) \leq g_{n-1}(x) \leq \ldots \leq g_0(x), x \in X, n \geq 0,$$

for some positive sequence satisfying the linear relationship

$$g_n(x) = \alpha_n h\left(x_n\right) + \beta_n. \tag{4.14}$$

When n increases $\alpha_n \downarrow 1$, $\beta_n \downarrow 0$.

In reference [10] the stability of the decision processes are reflected through the regression of linear functions. In this paper, the convergence of the cost function also satisfies this condition, where as n increases, the function $H(x_n)$ point of convergence to $g_n(x)$, i.e.:

$$H(x) := \lim_{n\to\infty} g_n(x).$$

Proof of Theorem 5. Let S be a compact set:

$$S := \left\{x : \underline{c}(x) \le 2\rho_0\right\}.$$

The upper bound for $\underline{c}(x)$ is not restricted to a specific value $2\rho_0$ and can also be any finite function that contains $c(x)$. If there exists $\delta > 0$ such that

$$K_n(x,\chi) \ge \delta. \tag{4.15}$$

The state χ that each λ_{n-1} can reach after the $n-1$th stage under the policy. The function $H(x)$ is almost everywhere bounded, and from Eqs (4.10) and (4.11), we have

$$H_{n-1}\left(F_{\lambda_n}(x,a)\right) \le H_{n-1}(x) - \frac{1}{2}c\left(x_n\right) + \rho_{n-1}P_S.$$

If the selected state falls within the set S, then $P_S = 1$, for the cost function at stage n the cost function is obvious, it is evident:

$$\{x : c\left(x_n\right) \le \rho_{n-1}\} \subset S.$$

By Fatou's lemma, it is obtained that

$$\sum_{i=0}^{T_\chi-1} c_{\lambda_n}\left(x_i, a_i\right) \le 2\left(H_{n-1}(x) - H_{n-1}(x) + \rho_{n-1}\sum_{i=0}^{T_\chi-1} P_s\left(x_i\right)\right).$$

The following inequality for the sum of probabilities falling into the set S ensures the validity of the following expression:

$$\sum_{i=1}^{T_\chi-1} P_s\left(x_i\right) \le \sum_{i=0}^{T_\chi-1} K_n\left(x_i, S\right)/\delta \le \varepsilon\pi/\delta.$$

Due to the boundedness of the cost function $c(\cdot)$, the following holds:

$$\sum_{i=0}^{T_\chi-1} K_n\left(x_i, S\right) = (1-\varepsilon)\sum_{i=0}^{T_\chi-1} \varepsilon^i\left(c\left(x_i, a_i\right) - \rho_t\right) \le (1-\varepsilon)\pi\frac{\varepsilon}{1-\varepsilon} = \pi\varepsilon,$$

π represents a constant for the difference between the maximum cost and the average cost. The original expression becomes:

$$\sum_{n=0}^{T_\chi-1} c\left(x_n, a_n\right) \le 2\left[H_{n-1}(x) + \rho_0\pi/\delta\right].$$

In particular, when:

$$N = \rho_0\pi/\delta,$$

we have:

$$H(x) \geq -N.$$

By Eq (4.14), the following inequality can be derived:

$$H_{n-1}\left(F_{\lambda_n}(x, a)\right) \leq H_{n-1}\left(x_n\right) - c\left(x_n\right) + \eta_{n-1},$$

$$\begin{aligned}
H_{\lambda_n}(x) &= \sum_{i=0}^{T_\chi - 1} \bar{c}\left(x_i\right) \\
&= \sum_{i=0}^{T_\chi - 1} \left[c\left(x_i\right) - \rho_n\right] \\
&= \sum_{i=0}^{T_\chi - 1} \left[\left(c\left(x_i\right) - \rho_{n-1}\right) + \left(\rho_{n-1} - \rho_n\right)\right] \\
&\leq H_{n-1}(x) + \left(\rho_{n-1} - \rho_n\right) \sum_{i=0}^{T_\chi - 1} c\left(x_i\right) \\
&\leq H_{n-1}(x)\left[1 + 2\left(\rho_{n-1} - \rho_n\right)\right] + 2\left(\rho_0 \pi / \delta\right)\left(\rho_{n-1} - \rho_n\right).
\end{aligned}$$

Thus, an upper bound for the discrepancy function is obtained as follows::

$$-\rho_0 \pi / \delta \leq H_n(x) \leq \left(1 + \xi_n\right) H_{n-1}(x) + \left(\rho_0 \pi / \delta\right) \xi_n,$$

$$\xi_n = 2\left(\rho_{n-1} - \rho_n\right),$$

$$H_n \leq \left(1 + \xi_n\right) H_{n-1}(x) + \left(\rho_0 \pi / \delta\right) \xi_n.$$

Likewise, it can be obtained that:

$$H_n \leq \left(1 + \xi_n\right) H_n(x) + \left(\rho_0 \pi / \delta\right) \xi_{n+1}.$$

Thus, there exists:

$$g_n(x) := \left(\prod_{i=n+1}^{\infty}\left(1 + \xi_i\right)\right)\left(H_n(x) + \left(\rho_0 \pi / \delta\right) \sum_{i=n+1}^{\infty} \xi_i\right).$$

Let

$$\alpha_n = \prod_{i=t+1}^{\infty}\left(1 + \xi_{n+1}\right), \beta_n = \left[\left(\rho_0 \pi / \delta\right) \sum_{i=n+1}^{\infty} \xi_i\right] \prod_{i=n+1}^{\infty}\left(1 + \xi_{n+1}\right)$$

satisfying Eq (4.14). □

This completes the proof. Theorem 5 primarily describes the characteristics that policy iteration exhibits in deterministic systems, laying the foundation for the subsequent theorems.

4.3. Convergence of the PI

In the previous section, the discussion of PI shifted focus to the discrepancy function, which has demonstrated that the cumulative cost discrepancy converges linearly to a value function. This result shows that in the decision-making process, the discrepancies between stages gradually stabilize, providing a preparatory condition for the convergence of PI. As a consequence, this approach relaxes the verification conditions for convergence. The following section will further validate the convergence of PI.

Assumption 2. *For*

(1) Each stage n, if the policy iteration generates a canonical triplet $(\lambda_{n-1}, H_{n-1}, \rho_{n-1})$ satisfying the Poisson equation (4.8):

$$H_{n-1}\left(F_{\lambda_{n-1}}(x,a)\right) = H_{n-1}(x) - c(x,a) + \rho_{n-1},$$

which results in a solution and for H_{t-1} being a bounded function

$$H_{n-1}\left(x_{\lambda_n}\right) \leq H_{n-1}\left(x_n\right) - c_{n-1}\left(x_n\right) + \eta_{\lambda_{n-1}},$$

$H_{n-1}(x)$ it can be minimized, utilizing the decision Eq (4.4)

$$\lambda_n(x) := \arg\min_{a\in A(x)} \left[c(x,a) + H_{n-1}(x,a)\right]$$

to find the next stage's policy $\lambda_n(x)$ for the next stage n.

(2) For a fixed x, the cost function $c(x,\cdot)$ and the function $\underline{c}(x)$ on the action space are functions resembling a norm , for any $x \in X$,$n \in Z^+$ for any satisfying:

$$\infty > K_n(x,A) \geq \underline{c}(x).$$

Under this assumption, the algorithm iteratively generates stable policies.

Theorem 6. *Strategy convergence theorem: If the above assumptions hold, for a certain stage n, the policy set $\{\lambda_i : i < n\}$ and the associated cumulative discrepancy functions $\{H_i : i < n\}$ are both defined through the policy iteration process, provided that:*

(1) The relevant value function has a lower bound $i \leq n-1$;
(2) Under tight sets, the level sets of the value functions of the MDPs obtained under the strategy as per assumption 2 are all bounded.

Then for the PI, there exists a solution (λ_n, H_n, ρ_n) such that

(1) The cumulative cost H_n discrepancy is bounded from below;
(2) When $0 \leq i \leq t$, the constant ρ_i is the average cost function at i stage, i.e.,

$$\rho_i = J\left(\lambda_i\right)$$

and the cost sequence decreases, i.e.:

$$\rho_0 \geq \rho_1 \geq \ldots \geq \rho_n.$$

If the cumulative cost discrepancy

$$H_{n-1} = H_n,$$

then the policy iteration converges.

Remark: In deterministic systems, the deterministic nature of the transition function and the validity of the average cost optimality inequality imply that the selection of policies during the iterative process effectively transforms into the selection of iterative paths. On one hand, the reduction of randomness contributes to enhancing the convergence of the algorithm. Throughout the iteration, there is an expectation that the optimal policy obtained at each stage will be included along the same iterative path. An increase in path determinacy signifies that policy selections will become more consistent, thereby improving the likelihood of converging to the optimal policy. On the other hand, the reduction of randomness can elevate the efficiency of the algorithm. When paths are more stable, the outcomes obtained during iteration will become more reliable, assisting in the reduction of unnecessary computations and accelerating convergence speed. The reduction of randomness reinforces the predictability of the decision-making processes, allowing the iterative algorithm to better assess the convergence of the policies. It is only when the randomness in path selection is diminished that the convergence conditions for the iteration can be satisfied, which is advantageous for both algorithm design and performance evaluation. In summary, as the number of iterations increases, the reduction of path randomness not only decreases the computational load but also provides a clearer indication of whether the iteration can converge, ultimately enhancing the performance of the algorithm in deterministic systems.

5. Application of VI and PI

In Sections 3 and 4, the VI and the PI are proved to hold under deterministic systems. The idea is to optimize the policy step by step until the optimal policy is found by performing the two steps of policy evaluation and policy improvement alternately. In the policy iteration process, the two steps of policy evaluation and policy improvement are performed at each iteration, but the value function is used in each step. In Section 4, the main focus is on the PI for deterministic systems.

Example: (Brock-Mirman model [6, 7]): For the infinite model, the criterion of discounting and non-discounting was introduced by Brock and Mirman [6, 7]. The state and control variables x_t, a_t denote capital and expenses at the time $t = 0, 1, 2, ...,$ respectively. The state space is $X = (0, \infty)$, and the action space $A = (0, \infty)$ are Borel space, $A(x) = \left(0, rx^\theta\right]$. The dynamic process of the system is as follows:

$$x_{t+1} = F(x, a) = rx_t^\theta - a_t, \quad t = 0, 1, 2 \ldots \tag{5.1}$$

Its initial state is x_0. Consider the objective function is to optimize the long-term AC:

$$J(\pi, x_0) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log(a_t). \tag{5.2}$$

Proof. From:

$$J(\pi, x_0) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \log(a_t),$$

it can be shown that the cost function is:

$$c(x,a)=\log(a), \tag{5.3}$$

for the AC inequality:

$$\rho^{*}+l(x)=c(x,a)+l^{*}\left(F\left(x,\lambda^{*}\right)\right). \tag{5.4}$$

5.1. Use of VI

These are the results obtained by Hernández et al. [4] on the AC criterion, which will be used as a basis to continue the discussion of the value iteration process in the following: assumption 1 is satisfied for Eqs (5.2)–(5.4), and for the difference function:

$$\begin{aligned}
&h_t(x)=J_t^{*}+l(x)-t\rho^{*}\\
&=\textstyle\sum_{i=0}^{t-1}\log\left(a_i\right)+l\left(x_t\right)-t\rho^{*}\\
&=\tfrac{\theta-\theta^{t}}{1-\theta}\log x_0+\left[\theta\tfrac{t-\frac{\theta-\theta^{t-1}}{1-\theta}}{1-\theta}+\tfrac{\theta}{1-\theta}\cdot\tfrac{1-\theta^{t-1}}{1-\theta}\right]\log r\theta-t\tfrac{\theta}{1-\theta}\log(r\theta).
\end{aligned}$$

If $t\to\infty$, then:

$$\lim_{t\to\infty}h_t=\frac{\theta}{1-\theta}\left[\log x_0+\log r\theta\right]=\frac{1}{1-\theta}\log x_0 r\theta.$$

Therefore, it can be seen that the discrepancy function is converging to a fixed value, and the iterative convergence of values is satisfied. □

5.2. Use of PI

Continuation of Brock-Mirman model. If the tth difference function:

$$h_{t-1}(x)=\sum_{i=0}^{T_{\lambda_{t-1}}-1}\left[\log a_{i-1}-\rho_{t-1}\right]$$

already satisfies Properties 1 and 2, then for the difference function at tth

$$\begin{aligned}
h_i(x)&=\sum_{i=0}^{T_{\lambda_t}-1}\left[\log a_i-\rho_t\right]\\
&=\sum_{i=0}^{T_{\lambda_t}-1}\left[\log a_i-\rho_{t-1}\right]+T_{\lambda_t}\left(\rho_{t-1}-\rho_t\right).
\end{aligned}$$

The T_{λ_t}th is equal to the $T_{\lambda_{t-1}}$; this is because the strategy λ_t at the tth is continued to be realized under the path generated by λ_t-1, so there is, when $t\to\infty$:

$$h_t(x)=h_{t-1}(x).$$

Satisfying the convergence of the PI. □

In the financial domain, many problems exhibit significant nonlinear characteristics, such as portfolio selection and option pricing. These nonlinear challenges often involve multiple factors, each contributing complex nonlinear effects. Such intricate dependencies not only complicate the direct resolution of the problems but also significantly increase the dimensionality of the state space.

In deterministic systems, the processes of value iteration and policy iteration provide an effective framework for addressing these challenges. In high-dimensional state spaces, these algorithms consolidate information from multiple related states into an abstract value by updating the state value function. This abstraction simplifies the decision-making process and reduces dimensionality, making it more feasible to handle high-dimensional decision problems. The Bellman equation plays a central role in this context. It recursively updates state values, effectively performing expected calculations of potential future returns for each state. Through each update, the algorithm uses existing information to assess the effectiveness of the current policy, thereby refining the strategy over time. Due to the dynamic nature of the system, the value dependencies among different states lead to the phenomenon of information sharing. This implies that the algorithm can efficiently synthesize and integrate information across iterations, utilizing prior learning to inform current decision-making. This mechanism of information sharing enables the derivation of effective strategies, even in the presence of complex, high-dimensional nonlinear problems.

Value iteration and policy iteration under deterministic systems not only offer a clear decision-making process but also provide theoretical support for tackling nonlinear problems. They leverage the interdependencies among states and the mechanisms of information sharing, thereby effectively addressing complex decision-making challenges.

5.3. Simulation

Algorithm1. Deterministic system value iteration.

Step1: Initialization

Set $k = 1$. Initialize the value function $V(0) = 0$ for all states. Choose a small constant $\epsilon > 0$ as the accuracy threshold, the iteration counter $k = 1$ and begin the iterative process.

Step2: Iteration

For each state $x \in S$, calculate the value of the system under each action $a \in A$. For each action a, calculate:

$$V(x) = c(x, a) + \gamma \cdot V(T(x, a)),$$

where $T(x, a)$ is the transition function, and $c(x, a)$ is the cost associated with action a at state x. Record the maximum change δ in the value function across all states:

$$\delta = \max_{x \in S} |V(x)^{new} - V(x)^{old}|.$$

Step3: Convergence Check

If $\delta < \epsilon$, stop the iteration and return the value function V as the solution. Otherwise, set $k = k + 1$ and return to Step 2 for the next.

Step4: Output the Results

Once the iteration has converged, output the final value function.

In the value iteration, with the number of states is set to 10, the number of actions is set to 3, and the threshold value is set to 1×10^{-3}; the following results are obtained:

- Final value function: $[0, 3, 3, 0, 0, 0, 3, 3, 0, 0]$.
- Final average cost: 1.2.
- Figure 1.

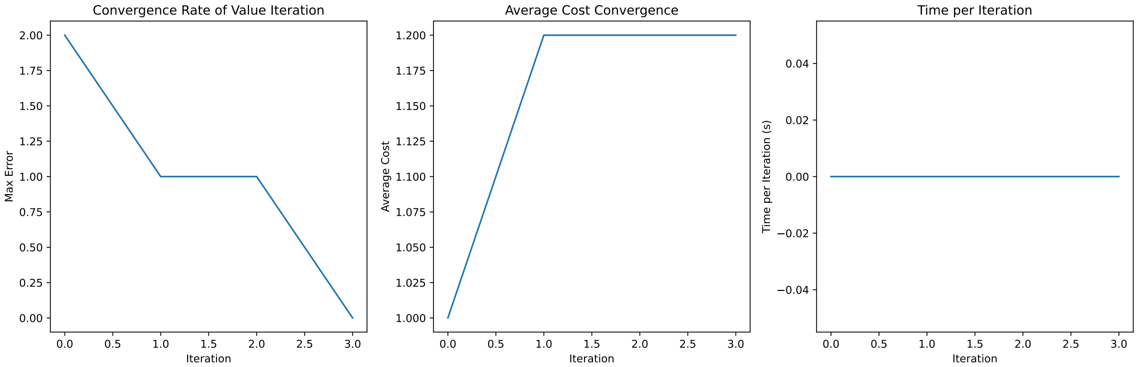

Figure 1. Value iteration.

From the above results, we can obtain the following results:

(1) Final value function: This vector represents the optimal value for each state, obtained by the convergence of the value iteration algorithm. Some states (positions 2, 3, 7, and 8) have a higher value of 3, while other states have a value of 0. This indicates a higher expected return in these particular states, possibly due to their ability to reach the target state faster or reduce the average cost more effectively. The distribution of the value function reflects the priority or preference in different states, where states with a value of 3 are considered more advantageous, due to higher rewards or lower costs in these states.

(2) Average cost: The average cost represents the expected cost per time step over an infinite operational period. The final average cost is 1.2, indicating that the current policy stabilizes the system's long-term operating cost.

(3) Convergence analysis:

  - Average cost convergence (middle plot): The average cost stabilizes at 1.2 after a few iterations, indicating that the current policy is near-optimal, as the average cost no longer fluctuates significantly. The stability of the average cost reflects the policy's effectiveness in balancing cost and return, allowing the system to operate at a low cost.
  - Policy change convergence (left plot): In the left plot, the policy change rate rapidly decreases to near zero after the initial few iterations. This fast convergence aligns with the final value function and average cost results, indicating that the policy iteration finds a stable set of policies within a few steps, maintaining the final average cost at 1.2.
  - Time per iteration (right plot): The time per iteration remains constant, showing that computation time did not increase with changes in state or policy. This is due to the simplification of the state space or the optimized handling of the algorithm.

Algorithm2. Deterministic system value iteration.

Step1: Initialization

Define the state space with size num_{states}, and the action space with size $num_{actions}$. Initialize a random policy and value function V. Set a convergence threshold δ to track error.

Step2: Iteration

Evaluate the value function V under the current policy. Initialize $\delta = 0$. For each state s, compute the updated value function:

$$V(s) = c(s, \pi(s)) + V(T(s, \pi(s))),$$

where $T(s, \pi(s))$ denotes the state transition given state s and action $\pi(s)$. Calculate the difference:

$$\delta = \max(\delta, |V_{\text{new}}(s) - V(s)|).$$

Continue until $\delta \leq \epsilon$.

Step3: Policy Improvement

For each state s, improve the policy by choosing the action that minimizes:

$$\pi(s) = \arg\min_{a} \left(c(s, a) + V(T(s, a)) \right).$$

Check if the policy π changes. If the policy has changed, continue to the next step; otherwise, terminate the iteration.

Step4: Record and Analyze

Track the policy change (as error), average cost, and time per iteration to assess the convergence and performance of the algorithm.

Step5: Check Convergence and Output

If the policy π remains stable (no changes), output the final policy and value function. Otherwise, return to Step 2 and continue iterating.

In the policy iteration, when the number of states is set to 10, the number of actions is set to 3, but the threshold value is set to 1×10^{-6} (policy iteration is faster), the following results are obtained:

- Final Policy: $[0, 0, 0, 2, 1, 0, 0, 0, 2, 1]$.
- Final Value Function:$[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$.
- Final average cost: 0.6.
- Figure 2.

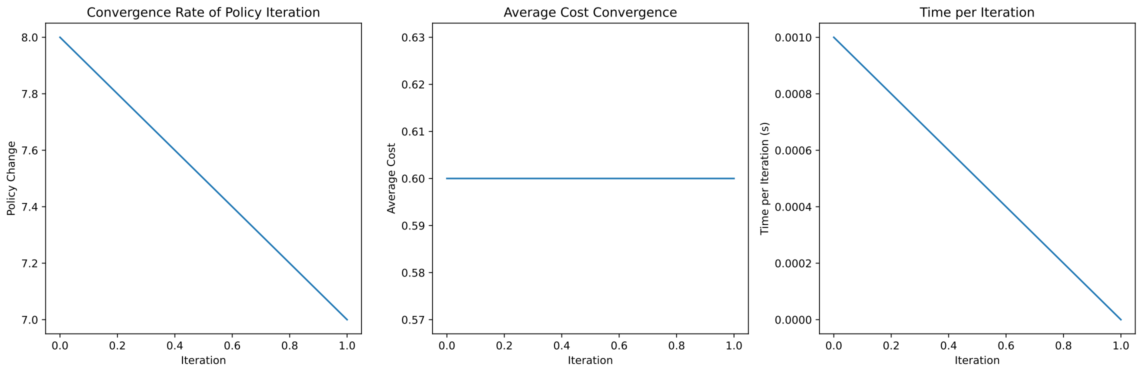

Figure 2. Policy iteration.

From the above results, we can obtain the following results:

(1) Final policy: The final policy vector represents the optimal action for each state in the state space. Each element corresponds to an action that minimizes the long-term cost for the specific state. In this case, the policy has converged, indicating the algorithm has found the preferred actions for each state based on the cost and transition functions.

(2) Final value function: The value function for each state has converged to zero, suggesting that each state's cost balances out over time, leading to zero additional value. This outcome aligns with the final average cost and implies a steady-state configuration.

(3) Final average cost: The average cost per time step, computed over an infinite horizon, converges to 0.6. This value indicates the expected cost that the system incurs per step under the optimal policy, reflecting a balance that minimizes the total cost per step.

(4) Convergence analysis:

- Convergence rate of policy iteration (left plot): The policy change rate decreases rapidly, converging toward zero within a few iterations, indicating that the policy stabilizes quickly.
- Average cost convergence (middle plot): The average cost converges smoothly to 0.6, maintaining stability across iterations, suggesting the system reaches a steady state with minimal fluctuation in cost.
- Time per iteration (right plot): The time per iteration remains consistent, indicating that each iteration takes a similar amount of computational time, reflecting the stability of the processes.

Based on the simulation results mentioned above, we observe that the iterative process in deterministic systems is very rapid. When the initial threshold is set to 1×10^{-3} , policy iteration can achieve convergence in a remarkably short time. Therefore, even with increased precision requirements during the policy iteration process, convergence can still be realized within a short period (1×10^{-6}).

As can be clearly seen from Figures 1 and 2, stability can be achieved in the iterative process without substantial fluctuations. This phenomenon is largely attributed to the fact that, in deterministic environments, the state transitions of the system are deterministic and the entire iterative processes are conducted offline. This presents a significant contrast to the work of He [16] and Fang [17, 18]. Thus, the degraded value iteration and policy iteration are more suitable for this specific subfield of deterministic systems.

6. Conclusions

This study primarily focuses on demonstrating the convergence of the value iteration and policy iteration under the average criterion in deterministic systems. The main areas of focus are divided into the following three parts:

(1) Verification of the validity of DP in deterministic systems;
(2) Convergence verification of the VI based on the AC criterion;
(3) Convergence verification of the PI based on the AC criterion.

In deterministic systems, value iteration (VI) can converge to the optimal value function, exhibiting a notably improved convergence speed. Furthermore, policy iteration (PI), based on the value function, also demonstrates enhanced convergence. A comparison between the contents of Sections 3 and 4 reveals that PI is more intricate than VI. In practical applications, PI is often easier to understand, and the policy derived at each update step is more meaningful compared to the policy obtained after VI convergence. As demonstrated in Section 5, both value iteration and policy iteration algorithms in deterministic systems can more effectively address certain nonlinear problems.

However, the limitations of existing theorems primarily lie in their lack of robustness. Specifically, the theorems regarding stochastic Markov processes necessitate considering the "boundary" surrounding each state point, which influences the selected state and typically requires integration to eliminate uncertainty. In contrast, deterministic systems do not face this uncertainty: given a current state and action, the subsequent state will be deterministic, not subject to random fluctuations. Therefore, the transition function in deterministic systems needs adjustment. The originally stochastic transition function, arising from multiple state points, can be simplified into an indicator function, rendering state transitions clear and stable. This transformation not only simplifies the complexity of the model but also allows for the avoidance of challenges and limitations related to randomness in the application of theorems.

During the derivation process in this study, it was further observed that the policy iteration algorithm in deterministic systems exhibits decreasing randomness in paths as the number of iterations increases, a result distinct from those observed in stochastic processes. Given that the field of deterministic systems has not yet been fully developed into a comprehensive framework, certain aspects still rely on existing theorems. For example, in Sections 2 and 3, dynamic programming processes discussed by Hernandez et al. [4] were employed, and in Section 5, the properties of the policy iteration algorithm were based on Meyn's work [10] There may be better methods for defining kernel functions in deterministic systems. Hernandez [2] raised several unresolved questions at the conclusion of their work, emphasizing the need for further development in addressing these challenges. Once a more comprehensive system is established, adjustments can be made to address the imperfections in the framework set forth in this study.

Author contributions

Conceptualization, Zheng.H;
Methodology, Zheng.H;
Validation, Wang.D;

Formal analysis, Zheng.H and Wang.D;
Writing—original draft preparation, Zheng.H and Wang.D;
Writing—review and editing, Zheng.H;
Visualization, Wang.D.
All authors have read and approved the final version of the manuscript for publication.

Conflict of interest

All authors disclosed no relevant relationship.

References

1. R. Bellman, The theory of dynamic programming, *Bull. Amer. Math. Soc.*, **1954**, 503–515.
2. O. Hernández-Lerma, J. B. Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, Vol. 30, New York: Springer Science & Business Media, 2012.
3. O. Hernández-Lerma, J. B. Lasserre, *Further topics on discrete-time Markov control processes*, Vol. 42, New York: Springer Science & Business Media, 2012.
4. O. Hernández-Lerma, L. R. Laura-Guarachi, S. Mendoza-Palacios, A survey of AC problems in deterministic discrete-time control systems, *J. Math. Anal. Appl.*, **522** (2023), 126906. https://doi.org/10.1016/j.jmaa.2022.126906
5. S. P. Meyn, R. L. Tweedie, *Markov chains and stochastic stability*, 1 Ed., New York: Springer London, 1993. https://doi.org/10.1007/978-1-4471-3267-7
6. W. A. Brock, L. J. Mirman, Optimal economic growth and uncertainty: the discounted case, *J. Econ. Theory*, **4** (1979), 479–513. https://doi.org/10.4337/9781782543046.00008
7. W. A. Brock, L. J. Mirman, Optimal economic growth and uncertainty: the no discounting case, *Int. Econ. Rev.*, **14** (1973), 560–573. https://doi.org/10.2307/2525969
8. V. Gaitsgory, A. Parkinson, I. Shvartsman, Linear programming based optimality conditions and approximate solution of a deterministic infinite horizon discounted optimal control problem in discrete time, *arXiv Preprint*, 2017. https://doi.org/10.48550/arXiv.1711.00801
9. O. L. V. Costa, F. Dufour, Average control of Markov decision processes with Feller transition probabilities and general action spaces, *J. Math. Anal. Appl.*, **396** (2012), 58–69. https://doi.org/10.1016/j.jmaa.2012.05.073
10. S. P. Meyn, The policy iteration algorithm for average reward Markov decision processes with general state space, *IEEE Trans. Automat. Control*, **42** (1997), 1663–1680. https://doi.org/10.1109/9.650016
11. R. Bellman, Dynamic programming, *Science*, **153** (1966), 34–37. https://doi.org/10.1126/science.153.3731.34
12. R. A. Howard, *Dynamic programming and markov processes*, MIT Press, 1960, 46–69. https://doi.org/10.1086/258477

13. S. Dai, O. Menoukeu-Pamen, An algorithm based on an iterative optimal stopping method for Feller processes with applications to impulse control, perturbation, and possibly zero random discount problems, *J. Comput. Appl. Mathe.*, **421** (2023), 114864. https://doi.org/10.1016/j.cam.2022.114864

14. E. A. Feinberg, Y. Liang, On the optimality equation for average cost Markov decision processes and its validity for inventory control, *Ann. Ope. Res.*, **317** (2022), 569–586. https://doi.org/10.1007/s10479-017-2561-9

15. Z. Yu, X. Guo, L. Xia, Zero-sum semi-Markov games with state-action-dependent discount factors, *Discrete Event Dyn. Syst.*, **32** (2022), 545–571. https://doi.org/10.1007/s10626-022-00366-4

16. S. He, H. Fang, M. Zhang, F. Liu, Z. Ding, Adaptive optimal control for a class of nonlinear systems: the online policy iteration approach, *IEEE Trans. Neural Netw. Learn. Syst.*, **31** (2019), 549–558. https://doi.org/10.1109/TNNLS.2019.2905715

17. H. Fang, M. Zhang, S. He, X. Luan, F. Liu, Z. Ding, Solving the zero-sum control problem for tidal turbine system: an online reinforcement learning approach, *IEEE Trans. Cybern.*, **53** (2023), 7635–7647. https://doi.org/10.1109/TCYB.2022.3186886

18. H. Fang, Y. Tu, H. Wang, S. He, F. Liu, Z. Ding, et al., Fuzzy-based adaptive optimization of unknown discrete-time nonlinear Markov jump systems with off-policy reinforcement learning, *IEEE Trans. Fuzzy Syst.*, **30** (2022), 5276–5290. https://doi.org/10.1109/TFUZZ.2022.3171844