International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(20s) 2026 pp. 170-178
https://doi.org/10.70917/ijcisim-2026-5162
Article

Mathematical Foundations of Reinforcement Learning and Stochastic Control Systems

Shirish Prabhakarrao Kulkarni[1], Dr. C. Ashwini[2], Dr.M. Buvanasankari[3] , Dr.K. Ramesh[4] , Dr. Rajat Verma[5] , Abhinav Jha[6]

[1]Assistant Professor, Ajeenkya D Y Patil University, lohagaon Pune, Maharashtra
[2]Assistant Professor, Department of Computer Science and Engineering, SRM Institute of science and Technology, Rampuram Campus, Chennai, Tamil Nadu
[3]Assistant Professor, Department of Mathematics, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, buvanasankari@gmail.com
[4]Associate Professor, Mathematics, Nehru Institute of Engineering and technology Coimbatore, Tamil Nadu, ramesh251989@gmail.com
[5]Associate Professor and Additional Head, Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur, Uttar Pradesh, India, Orcid Id: https://orcid.org/0000-0001-5707-2695
[6]Assistant Professor, Amity School of Engineering and Technology, Amity University Patna, Patna, Bihar, Orcid id: https://orcid.org/0000-0002-0085-583X

Abstract: Reinforcement learning (RL) and stochastic optimal control both address a single underlying mathematical problem: how an agent should select actions over time, under uncertainty, to optimize a cumulative reward or cost criterion. This paper reviews the shared mathematical scaffolding uniting these two fields, tracing the progression from Bellman's dynamic programming and the Markov decision process (MDP) formalism through stochastic approximation theory, temporal-difference learning, policy-gradient methods, actor-critic architectures, and the Hamilton–Jacobi–Bellman (HJB) equation governing continuous-time stochastic control. Particular attention is given to the convergence-theoretic results that justify RL algorithms as legitimate stochastic approximation procedures: Robbins and Monro's foundational stochastic approximation method, Jaakkola, Jordan, and Singh's convergence proof for stochastic iterative dynamic programming, Tsitsiklis and Van Roy's analysis of temporal-difference learning with linear function approximation, and the policy-gradient theorem of Sutton, McAllester, Singh, and Mansour. The paper further examines deterministic policy-gradient methods, deep reinforcement learning's departure from classical convergence guarantees, and the connection between the discrete-time Bellman equation and its continuous-time HJB counterpart via viscosity solution theory. Comparative tables map core mathematical structures onto their representation assumption and guarantee type, contrast convergence rigor across tabular, linear, and nonlinear function-approximation regimes, and set the discrete-time RL and continuous-time control literatures against their shared and divergent mathematical machinery. The paper concludes that convergence guarantees degrade in a predictable, representation-dependent order as function approximation grows more expressive, and identifies extending stochastic approximation theory to nonlinear function approximation as the central future research prospect.

Keywords: Reinforcement Learning, Stochastic Control, Markov Decision Processes, Bellman Equation, Stochastic Approximation, Temporal-Difference Learning, Policy Gradient Methods, Hamilton–Jacobi–Bellman Equation

1. Introduction

Sequential decision-making under uncertainty arises across robotics, finance, epidemiology, industrial process control, and game-playing agents, and in nearly every setting of practical interest the decision-maker cannot observe the future consequences of an action with certainty [1]. Classical stochastic optimal control, rooted in the calculus of variations and formalized by Bellman's principle of optimality, characterizes the optimal cost-to-go function as the solution of a functional equation, the Bellman equation in discrete time or the Hamilton–Jacobi–Bellman partial

differential equation in continuous time [1]. Reinforcement learning, emerging independently from computational learning theory and animal-learning psychology, addresses the same functional equation but under the additional constraint that the underlying transition dynamics and reward structure are unknown and must be estimated from sampled interaction with the environment [18].

The mathematical unification of these two traditions is neither accidental nor merely notational. Markov decision processes (MDPs), the formal object underlying nearly all of modern reinforcement learning, were themselves developed within operations research as a discrete-time stochastic control formalism before being adopted as the standard problem specification for RL [2]. The value function, the Bellman equation, and the principle of optimality that Bellman established for deterministic and stochastic control problems translate directly into the RL setting, with the critical difference that RL algorithms must estimate these quantities from data rather than compute them from a known model [1], [2].

This shared foundation has produced a body of convergence theory that borrows directly from stochastic approximation, the branch of probability theory concerned with the almost-sure convergence of stochastic iterative algorithms. Robbins and Monro's original stochastic approximation method established the conditions, principally a square-summable-but-not-summable step-size sequence, under which a noisy iterative root-finding procedure converges almost surely to the true root [8]. Watkins and Dayan's Q-learning algorithm, and its convergence proof due to Jaakkola, Jordan, and Singh, cast temporal-difference learning as precisely such a stochastic approximation procedure applied to the Bellman optimality operator, establishing that Q-learning converges to the optimal action-value function under the same step-size conditions Robbins and Monro identified decades earlier [4], [7], [8]. Borkar's subsequent dynamical-systems treatment of stochastic approximation generalized this convergence machinery to a far broader class of RL algorithms, including those with function approximation and two-timescale updates characteristic of actor-critic methods [9].

A second strand of the field addresses policy optimization directly rather than value estimation, formalized through the policy-gradient theorem, which expresses the gradient of expected cumulative reward with respect to policy parameters as an expectation computable from sampled trajectories, without requiring a model of the environment's transition dynamics [11]. Williams's REINFORCE algorithm was the earliest practical instantiation of this idea, and Sutton, McAllester, Singh, and Mansour's policy-gradient theorem later established the general mathematical form underlying REINFORCE and its variants, including the compatible function approximation conditions under which an approximate value function can replace the true value function inside the gradient estimator without introducing bias [10], [11] [19]. Actor-critic architectures, formalized rigorously by Konda and Tsitsiklis, combine both strands by maintaining a value-function estimate, the critic, that reduces the variance of a policy-gradient estimate, the actor, and Konda and Tsitsiklis established two-timescale stochastic approximation convergence conditions under which this coupled iteration converges to a local optimum [12].

The continuous-time analogue of the Bellman equation is the Hamilton–Jacobi–Bellman equation, a nonlinear partial differential equation whose solution is the optimal cost-to-go function for a continuous-time stochastic control problem, and who’s well-posedness in the general nonsmooth case is established through Fleming and Soner's viscosity solution theory [14]. Kalman's filtering theory, though addressing state estimation rather than control directly, established the linear-quadratic-Gaussian machinery that remains the principal closed-form solvable special case bridging discrete-time RL and continuous-time stochastic control [17].

This paper synthesizes this mathematical foundation, organizing the discussion around five core mathematical structures, the Bellman equation and dynamic programming, stochastic approximation theory, temporal-difference learning and its convergence guarantees, policy-gradient and actor-critic methods, and the continuous-time HJB formulation, before critically examining the mathematical challenges that arise once function approximation, particularly deep neural networks, replaces the tabular or linear representations for which most classical convergence theory was derived.

Research Objectives

- To formalize the mathematical structures, the Bellman equation, the MDP formalism, and the HJB equation, that jointly underlie reinforcement learning and stochastic control.
- To examine the stochastic approximation theory that establishes convergence guarantees for temporal-difference and Q-learning algorithms.
- To evaluate policy-gradient and actor-critic methods and their underlying variance-reduction and two-timescale convergence theory.
- To assess the connection between discrete-time RL and continuous-time stochastic control via the HJB equation and viscosity solutions.
- To identify open mathematical problems, particularly the absence of general convergence theory for RL with nonlinear function approximation.

2. Literature Review

2.1 Markov Decision Processes and the Bellman Equation

The Markov decision process formalizes sequential decision-making under uncertainty as a tuple of states, actions, transition probabilities, and a reward function, together with a discount factor that ensures the infinite-horizon cumulative reward remains well-defined [2]. Bellman's principle of optimality states that an optimal policy has the property that, whatever the initial state and initial decision, the remaining decisions constitute an optimal policy with respect to the state resulting from the first decision, a recursive structure that yields the Bellman equation expressing the optimal value function as the maximum, over actions, of the immediate reward plus the discounted expected value of the successor state [1] [19]. Puterman's comprehensive treatment formalizes the existence and uniqueness of this fixed point via the contraction-mapping property of the Bellman operator under the discounted-reward criterion, establishing that value iteration, the repeated application of this operator, converges geometrically to the unique optimal value function [2].

2.2 Dynamic Programming and Approximate Dynamic Programming

Classical dynamic programming computes the optimal value function directly when the transition and reward functions are fully known, via value iteration or policy iteration, both of which Bertsekas and Tsitsiklis's neuro-dynamic programming framework recast as special cases of a broader class of approximate dynamic programming algorithms designed for problems where exact computation is intractable due to the curse of dimensionality [3] [20]. Bertsekas and Tsitsiklis's central contribution was to demonstrate that replacing the exact Bellman operator with a sampled or function-approximated estimate preserves convergence under specific conditions on the approximation architecture and the sampling distribution, providing the theoretical bridge between classical dynamic programming and reinforcement learning's data-driven variants [3].

2.3 Stochastic Approximation Theory

Reinforcement learning algorithms that learn from sampled transitions rather than a known model are fundamentally stochastic approximation procedures, iterative algorithms that estimate the root or fixed point of a function using noisy observations rather than exact evaluations. Robbins and Monro established the foundational result that such an iteration converges almost surely to the true root provided the step-size sequence is square-summable but not summable, a condition that balances sufficient exploration of the iterate space against eventual convergence stability [8] [21]. Borkar's dynamical-systems approach reformulates stochastic approximation as a discretization of an underlying ordinary differential equation, showing that the stochastic iterate tracks the trajectory of this associated ODE with high probability, a framework that extends naturally to the two-timescale updates that appear throughout actor-critic reinforcement learning [9].

2.4 Temporal-Difference Learning and Convergence

Sutton's temporal-difference (TD) learning algorithm updates a value-function estimate using the difference between successive predictions rather than waiting for a final outcome, a bootstrapping mechanism that requires substantially less data than Monte Carlo estimation in many sequential prediction tasks [5]. Watkins and Dayan's Q-learning algorithm extended this bootstrapping idea to action-value estimation under an unknown policy, learning the optimal action-value function directly regardless of the behavior policy generating the data, an off-policy property with no direct analogue in on-policy TD prediction [4] [22]. Jaakkola, Jordan, and Singh provided the rigorous convergence proof underlying Q-learning, showing that the Q-learning update is a stochastic approximation of the Bellman optimality operator and that the same Robbins–Monro step-size conditions guarantee almost-sure convergence to the optimal action-value function under tabular representation [7]. Tsitsiklis and Van Roy subsequently extended this convergence analysis to TD learning with linear function approximation, establishing convergence to a well-defined fixed point under on-policy sampling while also demonstrating, through explicit counterexamples, that convergence can fail under off-policy sampling with linear function approximation, a result that identifies a precise mathematical boundary of the theory rather than a purely empirical limitation [6].

2.5 Policy Gradient Methods

Rather than estimating a value function and deriving a policy from it, policy-gradient methods parameterize the policy directly and optimize its parameters via gradient ascent on expected cumulative reward. Williams's REINFORCE algorithm established that the gradient of expected reward with respect to policy parameters can be estimated using only the log-derivative of the policy and the observed return, without requiring a model of the environment, at the cost of high variance in the gradient estimate [10]. Sutton, McAllester, Singh, and Mansour's policy-gradient theorem generalized this result, deriving the exact functional form of the policy gradient as an expectation over the state-action visitation distribution of the log-policy gradient weighted by the true action-value function, and established the compatible function approximation conditions under which substituting an approximate value function for the true one introduces no bias into the gradient estimate [11] [20].

2.6 Actor-Critic Architectures and Deterministic Policy Gradients

Actor-critic methods combine a policy-improvement step, the actor, with a value-function estimation step, the critic, using the critic's value estimate to reduce the variance of the actor's policy-gradient update relative to the raw REINFORCE estimator. Konda and Tsitsiklis provided the rigorous convergence analysis for this coupled iteration, framing it as a two-timescale stochastic approximation process in which the critic updates on a faster timescale than the actor, and establishing convergence to a locally optimal policy under this timescale separation, directly building on Borkar's general two-timescale stochastic approximation theory [9], [12]. Silver et al. subsequently derived the deterministic policy-gradient theorem, showing that for continuous action spaces, a deterministic policy's gradient can be estimated with substantially lower variance than the stochastic policy-gradient theorem's expectation over a full action distribution, since the deterministic gradient integrates only over the state distribution rather than the joint state-action distribution [13] [21].

2.7 Continuous-Time Stochastic Control and the Hamilton–Jacobi–Bellman Equation

The continuous-time analogue of the discrete Bellman equation is the Hamilton–Jacobi–Bellman equation, a nonlinear partial differential equation whose solution characterizes the optimal cost-to-go function for a controlled stochastic differential equation, derived by taking the continuous-time limit of the discrete dynamic programming recursion [14]. Fleming and Soner's viscosity solution theory established that the HJB equation generally admits no classical, continuously differentiable, solution, since the optimal cost-to-go function can develop non-differentiable kinks even for smooth problem data, and that viscosity solutions provide the appropriate weak-solution framework under which existence and uniqueness can be rigorously established [14]. Kalman's linear filtering and, by extension, the linear-quadratic-Gaussian (LQG) control framework represent the principal special case in which the HJB equation reduces to a solvable algebraic Riccati equation, providing a closed-form bridge between continuous-time stochastic

control theory and the value-function approximation architectures used throughout discrete-time reinforcement learning [17] [22].

2.8 Function Approximation, Deep Reinforcement Learning, and the Departure from Classical Convergence Theory

Mnih et al.'s deep Q-network (DQN) demonstrated that a nonlinear function approximator, a deep convolutional neural network, could be combined with Q-learning to achieve human-level performance across a broad suite of Atari game-playing tasks, but did so using engineering mechanisms, experience replay and a separately updated target network, specifically designed to counteract the instability that Tsitsiklis and Van Roy's linear-function-approximation analysis had already identified as a risk under off-policy bootstrapping [4], [6], [15]. Schulman et al.'s proximal policy optimization (PPO) algorithm addressed a related instability in policy-gradient methods, constraining each policy update to remain within a trust region of the previous policy via a clipped surrogate objective, a heuristic solution to the variance and step-size sensitivity that the classical policy-gradient theorem does not itself resolve [11], [16].

2.9 Research Gap

While stochastic approximation theory rigorously establishes convergence for tabular and linear-function-approximation reinforcement learning, and viscosity solution theory rigorously establishes well-posedness for the continuous-time HJB equation, comparatively little theory extends these guarantees to the nonlinear, deep-network function approximators that dominate contemporary practice, and the engineering mechanisms, experience replay, target networks, and trust-region constraints, that stabilize deep RL algorithms in practice are not derived from, and do not obviously correspond to, convergence conditions established by classical stochastic approximation theory [6], [9], [15], [16]. This review addresses this gap by organizing the field's mathematical foundations around the specific point at which each theoretical guarantee, contraction mapping, stochastic approximation, or viscosity solution, is established.

3. Methodology

This study adopts a qualitative, descriptive, and comparative research design synthesizing foundational and contemporary mathematical results in dynamic programming, stochastic approximation, reinforcement learning, and continuous-time stochastic control. Reviewed contributions were compared along three axes: guarantee type (contraction mapping, stochastic approximation, or PDE well-posedness); representation assumption (tabular, linear function approximation, or nonlinear function approximation); and whether the guarantee has been formally extended to, or is instead only empirically approximated by, contemporary deep reinforcement learning practice.

4. Results and Discussion

4.1 Core Mathematical Structures and Their Guarantee Type

Table 1 maps the core mathematical structures reviewed in Section 2 onto their representation assumption and the type of mathematical guarantee each provides, a structure emphasizing what is actually proven rather than a chronological summary of algorithms.

Table 1. Core Mathematical Structures, Representation Assumptions, and Guarantee Types

Mathematical Structure	Representation Assumption	Guarantee Type	Key Reference
Bellman equation / value iteration	Tabular (exact model known)	Contraction-mapping fixed point; geometric convergence	Bellman [1]; Puterman [2]
Approximate dynamic programming	Function-approximated value	Convergence under specified	Bertsekas & Tsitsiklis [3]

		architecture/sampling conditions	
Stochastic approximation (Robbins-Monro)	N/A (general iterative root-finding)	Almost-sure convergence under step-size conditions	Robbins & Monro [8]
Q-learning	Tabular	Almost-sure convergence to optimal action-value function	Watkins & Dayan [4]; Jaakkola, Jordan & Singh [7]
TD learning with linear function approximation	Linear	Convergence on-policy; documented divergence off-policy	Tsitsiklis & Van Roy [6]
Policy gradient theorem	Tabular or compatible function approximation	Unbiased gradient estimator (given compatibility conditions)	Sutton, McAllester, Singh & Mansour [11]
Actor-critic (two-timescale)	Linear (critic)	Convergence to local optimum under timescale separation	Konda & Tsitsiklis [12]; Borkar [9]
HJB equation (continuous-time)	Continuous state/action (PDE)	Viscosity-solution existence and uniqueness	Fleming & Soner [14]
Deep RL (DQN, PPO)	Nonlinear (deep network)	No general convergence guarantee; empirical stabilization only	Mnih et al. [15]; Schulman et al. [16]

4.2 Convergence Rigor Across Representation Regimes

Table 2 contrasts the three representation regimes, tabular, linear, and nonlinear function approximation, directly against one another across the specific dimensions of convergence rigor the reviewed literature addresses, a structure distinct from Table 1's structure-by-structure mapping.

Table 2. Convergence Rigor Across Tabular, Linear, and Nonlinear Representation Regimes

Dimension of Rigor	**Tabular Representation**	**Linear Function Approximation**	**Nonlinear (Deep Network) Representation**
Existence/uniqueness of fixed point	Guaranteed via Bellman operator contraction	Guaranteed under on-policy sampling only	Not generally established
Almost-sure convergence of learning algorithm	Established (Q-learning, TD)	Established on-policy; documented divergence off-policy	No general proof; stability is empirical
Off-policy behavior	Convergent (Q-learning is off-policy stable in tabular case)	Can diverge (Tsitsiklis & Van Roy counterexamples)	Divergence risk managed via replay buffers/target networks
Variance of policy-gradient estimator	N/A (value-based)	Reduced via compatible function approximation	Managed via trust-region/clipped objectives (PPO)
Representative source	Jaakkola, Jordan & Singh [7]	Tsitsiklis & Van Roy [6]	Mnih et al. [15]; Schulman et al. [16]

4.3 Discrete-Time RL and Continuous-Time Control: Shared and Divergent Machinery

Table 3 sets the discrete-time RL literature reviewed in Sections 2.1 through 2.6 directly against the continuous-time stochastic control literature reviewed in Section 2.7, identifying where the two traditions share mathematical machinery and where they diverge.

Table 3. Discrete-Time RL Versus Continuous-Time Stochastic Control: Shared and Divergent Machinery

Aspect	Discrete-Time RL	Continuous-Time Stochastic Control	Shared or Divergent
Governing equation	Bellman equation (functional recursion)	Hamilton–Jacobi–Bellman equation (PDE)	Shared optimality principle; divergent mathematical form
Well-posedness mechanism	Contraction mapping (Banach fixed-point)	Viscosity solution theory	Divergent (discrete algebra vs. PDE analysis)
Model-free extension	Central concern (Q-learning, TD, policy gradient)	Largely unresolved outside LQG special case	Divergent; discrete-time theory considerably more mature
Closed-form solvable special case	Linear-quadratic value function (tabular/linear MDPs)	Linear-Quadratic-Gaussian (LQG) via Riccati equation	Shared special case bridging both traditions
Key reference	Bellman [1]; Puterman [2]	Fleming & Soner [14]; Kalman [17]	—

4.4 Discussion

Taken together, the evidence in Tables 1 through 3 supports a representation-dependent, rather than uniform, account of convergence rigor across reinforcement learning and stochastic control. Table 1 establishes that each mathematical structure reviewed in this paper is associated with a specific, provable guarantee, contraction mapping, stochastic approximation, or PDE well-posedness, and Table 2 shows that this guarantee degrades in a predictable order as the underlying representation moves from tabular to linear to nonlinear: tabular Q-learning and value iteration inherit the strongest guarantees directly from the Bellman operator's contraction property combined with Robbins-Monro step-size conditions [1], [2], [7], [8]; linear function approximation retains convergence under on-policy sampling but Tsitsiklis and Van Roy's counterexamples demonstrate that even this weaker guarantee fails under off-policy sampling [6]; and nonlinear, deep-network function approximation currently has no comparable convergence theory, with the stabilization mechanisms used in practice functioning as empirically validated engineering solutions rather than results derived from an extension of stochastic approximation theory to the nonlinear setting [15], [16]. Table 3's discrete-continuous comparison indicates that this same asymmetry recurs at the boundary between discrete-time RL and continuous-time control: the LQG special case provides a rare, fully closed-form bridge between the two traditions, but outside this special case, the viscosity-solution well-posedness theory Fleming and Soner established for the general HJB equation does not directly supply the sample-based convergence guarantees a model-free, continuous-time reinforcement learning algorithm would require [14], [17]. This asymmetry between mature stochastic approximation and viscosity solution theory on one hand, and immature deep-network and continuous-time model-free convergence theory on the other, is arguably the central open mathematical problem connecting reinforcement learning and stochastic control, rather than a limitation specific to any single algorithm.

5. Conclusion

This paper has reviewed the mathematical foundations connecting reinforcement learning and stochastic control across five core structures: the discrete Bellman equation and dynamic programming, stochastic approximation theory, temporal-difference and Q-learning convergence, policy-gradient and actor-critic methods, and the continuous-time Hamilton–Jacobi–Bellman equation. The evidence demonstrates that no single mathematical framework currently spans the full range of representations used in practice; instead, tabular and linear methods retain the strongest

convergence guarantees, inherited directly from contraction-mapping and stochastic approximation theory, while nonlinear deep reinforcement learning achieves substantially greater empirical capability at the cost of a convergence theory that remains largely unestablished. The continuous-time HJB equation and its viscosity-solution well-posedness theory provide the field's most mathematically complete existence result, but this rigor has not yet been extended into a practical, sample-based, model-free algorithmic framework outside the linear-quadratic-Gaussian special case. Across all the structures reviewed, the absence of a convergence theory for nonlinear function approximation comparable to the classical contraction-mapping and stochastic approximation results remains the central unresolved mathematical problem.

6. Future Work

Future research should prioritize extending stochastic approximation convergence theory to nonlinear function approximation architectures, building directly on the two-timescale and ODE-tracking frameworks Borkar established for the linear and tabular cases. Further work is needed to formally characterize the relationship between deep reinforcement learning's practical stabilization mechanisms, experience replay, target networks, and trust-region policy constraints, and the step-size and sampling conditions that classical stochastic approximation theory identifies as necessary for convergence. Continued research connecting discrete-time reinforcement learning with continuous-time stochastic control should extend viscosity-solution well-posedness results toward sample-based, model-free algorithms operating directly on continuous-time or continuous-state systems, beyond the linear-quadratic-Gaussian special case where closed-form solutions currently exist. Research should also examine variance-reduction techniques for policy-gradient estimation with theoretical guarantees comparable to the compatible-function-approximation conditions established for actor-critic methods. Finally, future work should pursue unified convergence benchmarks that make explicit which classical stochastic approximation or viscosity-solution guarantee, if any, a given deep reinforcement learning algorithm's empirical stability corresponds to, so that algorithm selection in safety-critical stochastic control applications can rest on a more principled mathematical foundation.

REFERENCES

1. R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1957.
2. M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 1994.
3. D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.
4. C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, 1992.
5. R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.
6. J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.
7. T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Computation*, vol. 6, no. 6, pp. 1185–1201, 1994.
8. H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.
9. V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge, U.K.: Cambridge University Press, 2008.
10. R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3–4, pp. 229–256, 1992.
11. R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 1999.
12. V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.
13. D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Machine Learning (ICML)*, 2014.
14. W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, 2nd ed. New York, NY, USA: Springer, 2006.
15. V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

16. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
17. R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
18. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
19. ShirishP Kulkarni(2021). Fixed Point Theorem Apply to Modified Lipschitz condition to Solve Differential Equations, Turkish Online Journal of Qualitative Inquiry (TOJQI), Volume12, Issue 5, June 2021: 5169-5175
20. Shirish P Kulkarni (2021). The role fixed point theorem in differential equations, Research & Reviews: Discrete Mathematical Structures ISSN: 2394-1979 Volume 8, Issue 2,2021,DOI Journal: 10.37591/RRDMS
21. Shirish P Kulkarni (2022) The Role of Fixed-Point Theorem In Fractional Differential Equations Neuroquantology | October 2022 | Volume 20 | Issue 12 | Page 3432-3438| DOI: 10.14704/NQ.2022.20.12.NQ77353
22. Shirish P Kulkarni (2025),Enhanced Mathematical Framework for Non-Integer Order Functional Integro-Differential Equations through Advanced Contractivity Conditions and Dhage's Sophisticated Fixed Point Theorem, Eksplorium, Volume 46 No. 2, June 2025: 25–38, p-ISSN 0854-1418, e-ISSN 2503-42