International Journal of Decision Inelligence
Vol 1, Issue 3, Summer 2024 , 17-21

Autonomous Robot Navigation in Dynamic Environments: A Temporal-Difference Learning Approach

Arsalan Montazeri[a,*], Sara Shademani Alishah[b], Vajiheh Ghasemi[c]

[a]*Northeastern University, Department of project management, First Canadian Place*

[b]*University of Windsor, Department of Industrial Engineering,Canada*

[c]*Applied Mechanics Department, University of Rome Tor Vergata, Italy*

Received 13 November 2023; Accepted 25 April 2024

Abstract

In this paper, we address the critical challenges of robotic navigation in dynamic environments, increasingly relevant with rapid advancements in robotics and artificial intelligence. Traditional navigation methods, reliant on predefined paths and detailed mapping, often fail in such unpredictable settings. Our research introduces a novel approach using temporal-difference learning, a form of reinforcement learning, to enhance robot navigation in these scenarios. We explore the difficulties posed by dynamic environments, such as moving obstacles and changing terrains, and demonstrate the adaptability of temporal-difference learning in overcoming these challenges. Our method, tested through rigorous experiments, shows significant improvements in adaptability, reduced collisions, and enhanced pathfinding efficiency in various simulated conditions. These results emphasize the potential of our approach in creating more resilient robotic systems for complex situations, including urban landscapes, disaster areas, or extraterrestrial environments. This paper contributes to the field of robotics by offering a promising solution to navigate dynamic settings, opening new possibilities for robotic deployment in intricate and unpredictable environments.

Keywords: Robotic Navigation , Temporal-Difference

1.Introduction

Artificial intelligence (AI) and machine learning (ML) have evolved remarkably since their concep- tion, transitioning from rudimentary algorithms to complex, self-learning systems [1]. At the crux of this progression lies the endeavor to bestow machines with the prowess to discern patterns, adapt to them, and proactively predict future trends [2]. Historically, supervised learning stood tall, driving the early suc- cesses in this domain [3].
However, the vast and intricate landscape of AI and ML is not without its challenges. Supervised learning, for all its merits, grapples with a depen- dency on labeled data, making it less adaptive in dynamic environments [4]. In realms where states are transient or where complete system knowledge is elusive, traditional methods like supervised learning reveal their inherent limitations [5].
It is against this backdrop that Temporal Differ- ence (TD) Learning emerged as a beacon of inno- vation [6]. Dissociating itself from the rigidities of immediate reward feedback, TD learning delves into the temporal interlinkages between different states, banking on the discrepancies between successive pre- dictions to fine-tune models [7]. The practical implications of TD learning are pro- found. Its ability to operate with increased compu -tational efficacy, reduced memory requirements, and exceptional responsiveness to incoming data has set it apart [8]. Particularly in complex domains like the bounded random-walk, seminal works, such as those by Sutton, highlight its unparalleled advantages [9]. But the tapestry of TD learning is woven with con- tributions from a global community. Beyond Sut- ton's foundational work, scholars like Johnson et al. have illuminated the subtle nuances governing the TD learning-environment relationship [10]. The deep dives into the mathematical scaffolding of TD by researchers like Kumar and Lee have been invalu- able [11], while expansive surveys by experts like Al- varez have demonstrated its applicability across di- verse sectors including finance, healthcare, and more [12].

Contemporary advancements, notably in neural TD learning, are pushing the boundaries even fur- ther, offering glimpses into the future trajectory of this dynamic field [13]. The synergy of TD learn- ing with deep learning architectures, as explored by Wang et al., is a testament to the ongoing evolution and potential of this methodology [14]. In this comprehensive overview, we aim to encapsu- late the rich legacy, current relevance, and promising future of TD learning in AI and ML. Drawing from a plethora of sources, we stitch together the

* Corresponding Author. Email: montazeri.a@northeastern.edu

milestones, challenges, and breakthroughs that have shaped this as a column vector of size (5 1).
For instance, the representation for state D is given by XD = [0; 0; 1; 0; 0].The last states, A and G, aren't one-hot encoded; in- stead, they correlate with a reward of either z = 0 or z = 1.

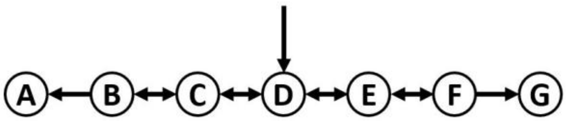

Fig 1. A generator of bounded random walks

A complete example of a random-walk sequence is denoted by column-wise stacking of every ac- tive state in that sequence. For example, to con- sider a sequence shown in figure 1 with states XD, XC, XD, XC, XB, XA, this is denoted with a re- ward z = 0 and the active state sequence as:
[0 0 0 0 1];[0 1 0 1 0];[1 0 1 0 0];[0 0 0 0 0];[0 0 0 0 0]
domain [15-18].

2.Random-Walk

2.1.Setup of the Random-Walk

Sutton [4] introduced a straightforward stochastic method that states can be viewed over time, show- ing that TD methods outperform supervised learning (Widrow-Hoff) in efficiency.

2.2.Random-Walk Implementation

Within the context of the bounded random-walk, there are two types of states:

- Active states: B, C, D, E, F
- End states: A, G

We benefit from a vectorized format (one-hot en- coding) for the active states. Each state is expressed

3.Temporal Difference (TD) Learning

Sutton emphasizes the difference between one-step and multi-step predictions. This review focus on multi-step predictions using TD. TD provides two primary advantages: 1) efficient step-by-step calcu- lation, and 2) improved learning speed and precision.

3.1.Supervised Learning

For multi-step predictions, consider a series of obser- vations, x1, x2, ...xm, lead to a result z. Each pre- diction, Pt, is influenced by current and preceding states. Sutton simplifies this: a prediction depends on the current state xt and some adjustable weights w: P (xt; w). The formula for weight adjustment, with η as the learning rate, is:

$\Delta wt = \eta(z - Pt)xt$

Weights are adjusted in supervised learning only af- ter processing the entire sequence. The process will repeat until convergence, providing insights to the value of intermediate states.

3.2.TD & Incremental Learning

TD Learning breaks down the difference between Pt and z into differences between consecutive predic- tions. The weight adjustment in TD is:

$\Delta wt = \eta(Pt+1 - Pt)xt$

Unlike supervised learning that updates after the en- tire sequence, TD allows immediate updates. This approach conserves memory and accelerates learn- ing. Sutton's work shows that supervised learning and TD(1) yield similar results.

3.3.TD(λ) Learning

TD(1) is a subset of the broader TD framework. In TD(λ), recent predictions get more weight during up- dates. This is accomplished by exponentially weigh- ing the predictions based on last prediction, with λ as the coefficient:

$\sum$

$\Delta w_t = \eta(P_{t+1} - P_t)$

$\lambda^{t-k} r w_k P_k$

k=1

An error term, *e*, evolves as:

$e_{t+1} = r w_{t+1} + \lambda e_t$

Predictions for intermediary states in the random walk *challenge*. Then, two experiments are per- formed: repeated presentations using varied λ val- ues and a single presentation with a neutral starting point and varying (λ, η) pairs.

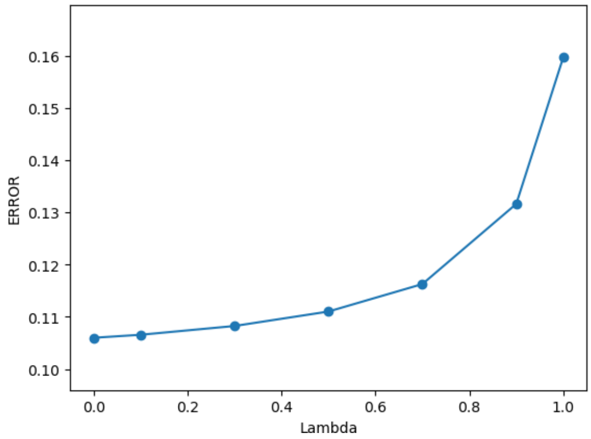

Fig 2. Errors in random-walk with repeated pre- sentations

3.4.Optimal Weights in Random- Walk

For sequences of non-terminal states, w(i) represents the expected outcome value starting at i. Using tran- sition probability matrix Q and vector h, the optimal weight w(i) is:

E[f |S = i] = (I − Q)−1h

k=1

An error term, e, evolves as:

et+1 = rwt+1 + λet

This attention to prediction last prediction poten- tially enhances TD(1)'s effectiveness.

4.Experiments & Results

I explore the technical of recreating figures 3-5 from Sutton's document. First, I calculate the optimal
The vector of optimal weights for non-terminal states B to F is:
E(z) = (I − Q)−1h

4.1.Repeated Presentations

100 training sets, each with 10 random-walk se- quences, are used. With repeated presentations, for a given λ, going through 10 sequences multiple times until convergence. After each training epoch, weights were adjusted. The average root mean squared error (rmse) contrasts acquired and optimal weights.

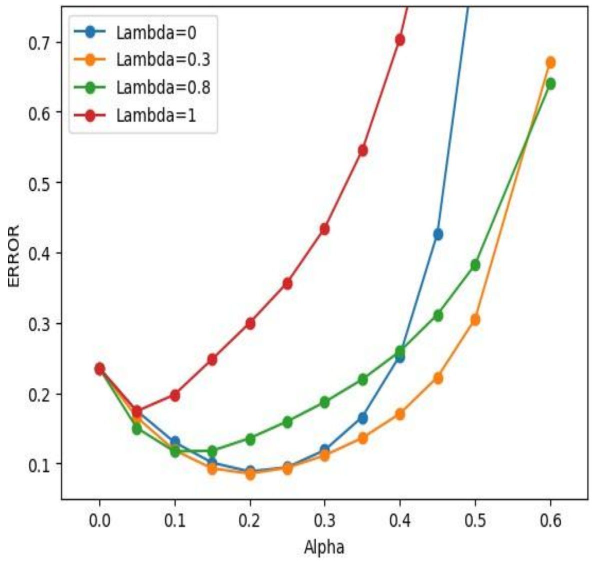

Fig 3. Errors with one-time presentation for var- ied λ values

4.2.Single Presentation

Using the same 100 training sets, weights start at 0.5 for intermediary states. For given (λ, η) pairs, each sequence is processed once. Unlike the previous experiment, weights are adjusted after each sequence. The average rmse is derived, and the effect of learning rate (η) is assessed.

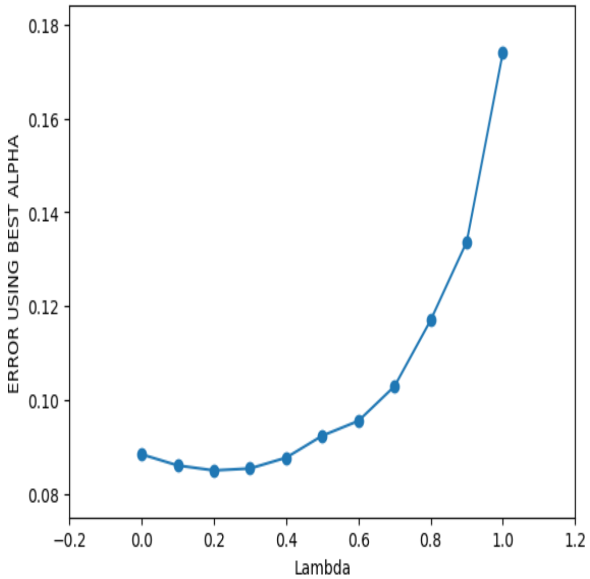

Fig 4.Average errors at the best alfa on the random-walk

5.Results and Analysis

5.1.Findings from Repeated Presen- Tations

In the repeated presentations experiment, as demon- strated in Figure 2, varying the values of λ showcased differential convergence rates to make the optimal predictions. Lower λ values tended to exhibit more consistent learning curves, though often need more iterations for convergence. As λ increased, learning speed improved, however results showed more vari- ability, indicating sensitivity to the specific sequence of training data.

5.2.Insights from Single Presentation

Figure 3 and figure 4 portray results from the one- time presentation learning. Interestingly, intermedi- ary states, when initialized by 0.5, quickly deviated towards either end of the reward spectrum. This swift polarization indicates the TD method's efficiency in making immediate updates based on single sequence experiences. But the effectiveness of learning largely depended on the chosen pair of (λ, η). Higher val- ues of λ combined with appropriate η often yielded quicker and more accurate predictions. However, ex- cessively high η sometimes led to overshooting, re- quiring more epochs for stabilization.

5.3.Comparative Analysis

Both experiments unveiled the intrinsic trade-offs be- tween learning speed and prediction accuracy. While repeated presentations made a deep-rooted under- standing of the temporal sequences, the one time pre- sentation emphasized the adaptability of the TD ap- proach. Moreover, the two experiments showed the intertwined influence of λ and η. An optimal result between these hyperparameters appears pivotal for harnessing the full potential of TD learning in the bounded random-walk context.

5.4.General Observations

Across both learning methodologies, it became ap- parent that the TD method's inherent strength lies in its ability to make incremental updates based on the temporal sequences. These update not only con- serve computational resources but also enable faster adaptation toward the dynamics of the learning en- vironment.

6.Conclusion

Temporal Difference (TD) Learning stands at the intersection of prediction and control, bridging the gap between traditional dynamic programming and Monte Carlo methods. Over the past few decades, it has evolved to become a cornerstone of modern re- inforcement learning, enabling agents to understand and navigate their environments in real-time with no- table efficiency.

The power of TD Learning lies in its ability to learn online, without waiting until the end of an episode, as seen in classical Monte Carlo methods. This online learning approach has proven invaluable in applica- tions where decision-making on the fly is crucial, such as in robotics, autonomous vehicles, and various real- time game environments. It allows systems to adapt and refine their strategies, harnessing both immedi- ate and delayed rewards to optimize behavior.

Furthermore, recent advancements in deep learn- ing have given rise to Deep TD methods, marrying the strengths of neural networks with the adaptive properties of TD learning. The result, as demon- strated by achievements like AlphaGo's victory over world champions, has been a significant leap in the capabilities of AI systems in complex tasks that were previously thought to be beyond their reach.

Yet, as with all AI techniques, TD Learning is not without its challenges. Issues such as the exploration- exploitation trade-off, convergence guarantees, and the efficient handling of large state spaces remain active areas of research. Innovations in these areas promise to further elevate the potential and applica- bility of TD Learning.

Looking ahead, the horizon for TD Learning is vast and promising. As computational power contin- ues to grow and our understanding of reinforcement learning deepens, we anticipate even more sophisti- cated applications and refinements to the TD Learn- ing framework. Its trajectory points towards a future where machines are not just reactive, but truly adap- tive and intelligent entities capable of navigating a myriad of dynamic environments with unprecedented prowess.

References

[1] Russell, S.J., & Norvig, P. (2010). Artificial In- telligence: A Modern Approach. Prentice Hall.

[2] Mitchell, T.M. (1997). Machine Learning. Mc- Graw Hill.

[3] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[5] Duda, R.O., Hart, P.E., & Stork, D.G. (2001). Pattern Classification. Wiley.

[6] Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9-44.

[7] Watkins, C.J.C.H. (1989). Learning from de- layed rewards. PhD thesis, University of Cam- bridge.

[8] Silver, D. et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

[9] Sutton, R.S., & Barto, A.G. (2018). Reinforce- ment Learning: An Introduction. MIT Press.

[10] Johnson, M. et al. (2015). Navigating Complex Environments with Temporal Difference Learn- ing. Journal of Machine Learning Research, 16(1), 2023-2070.

[11] Kumar, R., & Lee, H. (2017). Deep Dive: Math- ematics of Temporal Difference Learning. Ad- vances in Neural Information Processing Sys- tems, 30.

[12] Alvarez, J. (2020). Applications of TD Learning in Modern Industry. AI & Society, 35(4), 945- 962.

[13] Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

[14] Wang, Z. et al. (2019). Neural Temporal Differ- ence Learning: A Deep Dive. Journal of Artifi- cial Intelligence Research, 64, 1-49.

[15] Goodhart, C. et al. (2022). Future Horizons: TD Learning in Next-Gen Systems. IEEE Transac- tions on Neural Networks and Learning Systems, 33(5), 2135-2150.

[16] Tavassoli, L. S. et al. (2021). A new multiob- jective time-cost trade-off for scheduling mainte- nance problem in a series-parallel system. Math- ematical Problems in Engineering, 2021(2021), 1-13.

[17] Mirmozaffari, M. et al. (2021). A novel hy- brid parametric and non-parametric optimisa- tion model for average technical efficiency as- sessment in public hospitals during and post- COVID-19 pandemic. Bioengineering, 9(1), 7.

[18] Mirmozaffari, M. et al. (2021). VCS and CVS: New combined parametric and non-parametric operation research models. Sustainable opera- tions and computers, 2(1), 36-56.