An MRP Formulation for Supervised Learning: Generalized Temporal Difference Learning Models (Abstract Reprint)

Yangchen Pan[1], **Junfeng Wen**[2], **Chenjun Xiao**[3], **Philip Torr**[1]

[1]University of Oxford
[2]Carleton University
[3]The Chinese University of Hong Kong (Shenzhen)

Abstract Reprint. This is an abstract reprint of the journal article by Pan, Wen, Xiao, and Torr (2025).

Abstract

Background: Traditional supervised learning (SL) assumes data points are independently and identically distributed (i.i.d.), which overlooks dependencies in real-world data. Reinforcement learning (RL), in contrast, models dependencies through state transitions. Objectives: This study aims to bridge SL and RL by reformulating SL problems as RL tasks, enabling the application of RL techniques to a wider range of SL scenarios. We aim to model SL data as interconnected, and develop novel temporal difference (TD) algorithms that can accommodate diverse data types. Our objectives are to (1) establish conditions where TD outperforms ordinary least squares (OLS), (2) provide convergence guarantees for the generalized TD algorithm, and (3) validate the approach empirically using synthetic and real-world datasets. Methods: We reformulate traditional SL as a RL problem by modeling data points as a Markov Reward Process (MRP). We then introduce a concept analogous to the inverse link function in generalized linear models, allowing our TD algorithm to handle various data types. Our analysis, grounded in variance estimation, identifies conditions where TD outperforms OLS. We establish a convergence guarantee by conceptualizing the TD update rule as a generalized Bellman operator. Empirical validation begins with synthetic data progressively matching theoretical assumptions to verify our analysis, followed by evaluations on real-world datasets to demonstrate practical utility. Results: Our theoretical analysis shows that TD can outperform OLS in estimation accuracy when data noise is correlated. Our approach generalizes across various loss functions and SL datasets. We prove that the Bellman operator in our TD framework is a contraction, ensuring convergence for both expected and stochastic TD updates. Empirically, TD outperforms SL baselines when data aligns with its assumptions, remains competitive across diverse datasets, and is robust to hyperparameter choices. Conclusions: This study demonstrates that SL can be reformulated as a problem of interconnected data modeled by an MRP, effectively solved using TD learning. Our generalized TD is theoretically sound, with convergence guarantees, and practically effective. It generalizes OLS, offering superior performance on correlated data. This work enables RL techniques to benefit SL tasks, offering a pathway for future advancements.

References

Pan, Y.; Wen, J.; Xiao, C.; and Torr, P. 2025. An MRP Formulation for Supervised Learning: Generalized Temporal Difference Learning Models. *Journal of Artificial Intelligence Research*, 83: 33.