# A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation

**Jalaj Bhandari** JB3618@COLUMBIA.EDU
*Industrial Engineering and Operations Research, Columbia University*

**Daniel Russo** DAN.JOSEPH.RUSSO@GMAIL.COM
*Decision Risk and Operations, Columbia Business School*

**Raghav Singal** RS3566@COLUMBIA.EDU
*Industrial Engineering and Operations Research, Columbia University*

**Editors:** Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

## [1] Abstract

Temporal difference learning (TD) is a simple iterative algorithm used to estimate the value function corresponding to a given policy in a Markov decision process. Although TD is one of the most widely used algorithms in reinforcement learning, its theoretical analysis has proved challenging and few guarantees on its statistical efficiency are available. In this work, we provide a *simple and explicit finite time analysis* of temporal difference learning with linear function approximation. Except for a few key insights, our analysis mirrors standard techniques for analyzing stochastic gradient descent algorithms, and therefore inherits the simplicity and elegance of that literature. A final section of the paper shows that all of our main results extend to the study of a variant of Q-learning applied to optimal stopping problems.

**Keywords:** Reinforcement learning, temporal difference learning, finite sample bounds, stochastic gradient descent.

## 1. Introduction

Reinforcement learning (RL) offers a general paradigm for learning effective policies for stochastic control problems. At the core of RL is the task of value prediction: the problem of learning to predict cumulative discounted future reward as a function of the current state of the system. Usually, this is framed formally as the problem of estimating the value function corresponding to a given policy in a Markov decision process (MDP). Temporal difference learning (TD), first introduced by Sutton (1988), is the most widely used algorithm for this task. The method approximates the value function by a member of some parametric class of functions. The parameters of this approximation are then updated online in a simple iterative fashion as data is gathered.

While easy to implement, theoretical analysis of TD is quite subtle. A central challenge is that TDs incremental updates, which are cosmetically similar to stochastic gradient updates, are not true gradient steps with respect to any fixed loss function. This makes it difficult to show that the algorithm makes consistent, quantifiable, progress toward any particular goal. Reinforcement learning researchers in the 1990s gathered both limited convergence guarantees for TD and examples of divergence. Many issues were clarified in the work of Tsitsiklis and Van Roy (1997), who established

[1] Extended abstract. Full version available on arXiv with the same title.

© 2018 J. Bhandari, D. Russo & R. Singal.

precise conditions for the asymptotic convergence of TD with linear function approximation, and provided counterexamples when these conditions are violated. With guarantees of asymptotic convergence in place, a natural next step is to understand the algorithm's statistical efficiency. How much data is required to reach a given level of accuracy? Can one give uniform bounds on this, or could data-requirements explode depending on the problem instance? Twenty years after the work of Tsitsiklis and Van Roy (1997), such questions remain largely unsettled.

In this work, we take a step toward correcting this by providing *a simple and explicit finite time analysis of temporal difference learning.* We draw inspiration from the analysis of projected stochastic gradient descent. These analyses are simple–enough so that they are frequently taught in machine learning courses–and the explicit bounds they produce provide clear assurance of the robustness of SGD. Unfortunately, there are critical differences between TD and SGD, and as such these simple analyses do not apply to TD. Instead, past work on TD has needed to invoke powerful results from the theory of stochastic approximation. In this work, we uncover an approach to analyzing TD which, except for a few crucial steps, leverages the standard tools for finite time analysis of SGD. In addition to the several novel guarantees we derive in the paper, we feel the analysis offers insight into the dynamics of TD, and we hope our approach helps future researchers derive stronger bounds and principled improvements to the algorithm.

## References

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 42(5), 1997.