Temporal Difference Learning as Gradient Splitting

Rui Liu [1] **Alex Olshevsky** [2]

Abstract

Temporal difference learning with linear function approximation is a popular method to obtain a low-dimensional approximation of the value function of a policy in a Markov Decision Process. We provide an interpretation of this method in terms of a splitting of the gradient of an appropriately chosen function. As a consequence of this interpretation, convergence proofs for gradient descent can be applied almost verbatim to temporal difference learning. Beyond giving a fuller explanation of why temporal difference works, this interpretation also yields improved convergence times. We consider the setting with $1/\sqrt{T}$ step-size, where previous comparable finite-time convergence time bounds for temporal difference learning had the multiplicative factor $1/(1-\gamma)$ in front of the bound, with γ being the discount factor. We show that a minor variation on TD learning which estimates the mean of the value function separately has a convergence time where $1/(1-\gamma)$ only multiplies an asymptotically negligible term.

1. Introduction

Reinforcement learning is a basic machine learning paradigm which concerns learning optimal policies in Markov Decision Processes (MDP). It has been applied to many challenging practical problems, such as, autonomous driving (Chen et al., 2015), robotics (Gu et al., 2017), bidding and advertising (Jin et al., 2018), and games (Silver et al., 2016). An important problem in reinforcement learning is to estimate the value function for a given policy, often referred to as the policy evaluation problem. Temporal difference (TD) learning originally proposed by Sutton (1988) is one of the most widely used policy evaluation algorithms. TD uses differences in predictions over successive time steps to drive the learning process, with the prediction at any given time step updated via a carefully chosen step-size to bring it closer to the prediction of the same quantity at the next time step.

Despite its simple implementation, theoretical analysis of TD can be involved. This is particularly true when TD methods are applied to problems with large state-spaces by maintaining an approximation to the value function. Precise conditions for the asymptotic convergence of TD with linear function approximation were established by viewing TD as a stochastic approximation for solving a suitable Bellman equation in Tsitsiklis & Van Roy (1997). Before the last few years, there have been few non-asymptotic analyses of TD methods. The first non-asymptotic bounds for TD(0) with linear function approximation were given by Korda & La (2015), obtaining an exponential convergence rate for the centered variant of TD(0) when the underlying Markov chain mixes fast. However, some issues with the proofs of Korda & La (2015) were listed by the subsequent work of Narayanan & Szepesvári (2017).

In Lakshminarayanan & Szepesvari (2018), it was shown that TD algorithms with a problem independent constant step size and iterate averaging, achieve a problem dependent error that decays as $O(1/t)$ with the number of iterations t. Convergence rates in probability with an $O(1/t)$ step-size were provided by Dalal et al. (2018). Both analyses of Dalal et al. (2018) and Lakshminarayanan & Szepesvari (2018) assume samples used by the algorithm are i.i.d. rather than a trajectory in the underlying Markov chain. For the Markov chain observation model, Bhandari et al. (2018) provide a $O(1/\sqrt{T})$ convergence rate with step-size that scales as $1/\sqrt{T}$ and $O((\log t)/t)$ convergence rate with step size $O(1/t)$ for projected TD algorithm. The constant factors in the latter bounds depend on $1/(1-\gamma)$, where γ is the discount factor; this scaling is one of the things we will be studying in this paper.

A number of papers also work on algorithms related to and inspired by the classic TD algorithm in the setting with Markovian sampling. Srikant & Ying (2019) give finite-time bounds for the TD algorithms with linear function approximation and a constant step-size. The two time-scale TD with gradient correction algorithm under a Markovian sampling

[1] Division of Systems Engineering, Boston University, Boston, MA, USA [2] Department of ECE and Division of Systems Engineering, Boston University, Boston, MA, USA. Correspondence to: Rui Liu <rliu@bu.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

and linear function approximation are discussed by Xu et al. (2019b) and shown to converge as fast as $O((\log t)/t^{2/3})$. A method called TD-AMSGrad under linear function approximation is studied by Xiong et al. (2020); with a constant step size, TD-AMSGrad converges to a neighborhood of the global optimum at a rate of $O(1/t)$, and with a diminishing step size, it converges exactly to the global optimum at a rate of $O((\log t)/t)$. Xu et al. (2019a) present performances of variance reduced TD and give a reduced bias error over the classic TD.

In this paper, we will study the convergence of TD(0) and TD(λ) with linear function approximation under Markovian observations. Our main contribution is to provide an interpretation of temporal difference learning: we show how to view it as a "splitting" (a term we will define later) of an appropriately chosen quadratic form. As a consequence of this interpretation, it is possible to apply convergence proofs for gradient descent almost verbatim to temporal difference learning.

The convergence times bounds we obtain this way improve on existing results. In particular, we study step-sizes of $1/\sqrt{T}$, which are typically recommended because the resulting error bounds do not depend on the inverse eigenvalues of the matrices involved in the linear approximation, which can be quite large; by contrast, methods that achieve faster than $O(1/\sqrt{T})$ decay have performance guarantees that scale with these same eigenvalues. We provide a minor variation on TD(0) for which we obtain a convergence rate that scales as $O\left[(1/(1-\gamma))^2\right]/T+\widetilde{O}(1/\sqrt{T})$, with the constant in the $\widetilde{O}(\cdot)$ term not blowing up as $\gamma \to 1$. We will also explain why a factor of $1/(1-\gamma)^2$ multiplying the asymptotically negligible $O(1/T)$ term as here is unavoidable.

2. Preliminaries

In this section, we describe the basics of MDPs and TD learning methods. While all this material is standard and available in textbooks (e.g., Sutton & Barto (2018)), it is necessary to standardize notation and make our presentation self-contained.

2.1. Markov Decision Processes

We consider a discounted reward MDP described by a 5-tuple $(\mathcal{S},\mathcal{A},\mathcal{P},r,\gamma)$, where $\mathcal{S}=[n]=\{1,2,\cdots,n\}$ is the finite state space, $\mathcal{A}$ is the finite action space, $\mathcal{P}=(\mathcal{P}(s'|s,a))_{s,s'\in\mathcal{S},a\in\mathcal{A}}$ are the transition probabilities, $r=(r(s,s'))_{s,s'\in\mathcal{S}}$ are rewards which are determined deterministically by the state transition pair (s,s') and $\gamma\in(0,1)$ is the discount factor. The (stationary) policy to be evaluated is a mapping μ: $\mathcal{S}\times\mathcal{A}\to[0,1]$, where $\mu(s,a)$ is the probabilities to select action a when in state s and $\sum_{a\in\mathcal{A}}\mu(s,a)=1$ for all states $s\in\mathcal{S}$. We adopt the shorthand s_t for the state at step t, a_t for the action taken at step t, and $r_{t+1}=r(s_t,s_{t+1})$.

The value function of the policy μ, denoted $V^\mu:\mathcal{S}\to\mathbb{R}$ is defined as

$$V^\mu(s)=E_{\mu,s}\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right],$$

where $E_{\mu,s}[\cdot]$ indicates that s is the initial state and the actions are chosen according to μ.

The immediate reward vector $R^\mu:\mathcal{S}\to\mathbb{R}$ is defined as

$$R^\mu(s)=E_{\mu,s}(r_1)=\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu(s,a)\mathcal{P}(s'|s,a)r(s,s').$$

For the remainder of the paper, we will be fixing the policy μ; consequently, we can talk about the probability transition matrix P^μ defined as

$$P^\mu(s,s')=\sum_{a\in\mathcal{A}}\mu(s,a)\mathcal{P}(s'|s,a).$$

In the following, we will treat V^μ and R^μ as vectors in $\mathbb{R}^n$, and treat P^μ as a matrix in $\mathbb{R}^{n\times n}$. It is well-known that V^μ satisfies the Bellman equation (Sutton & Barto, 2018): defining the Bellman operator $T^\mu:\mathbb{R}^n\to\mathbb{R}^n$ as

$$(T^\mu V^\mu)(s)=\sum_{s'=1}^{n}P^\mu(s,s')(r(s,s')+\gamma V^\mu(s'))$$

for $s\in[n]$, we can then write Bellman equation as

$$T^\mu V^\mu=V^\mu.$$

Next, we state some standard assumptions from the literature. The first assumption is on the underlying Markov chain.

Assumption 1. *The Markov chain whose transition matrix is the matrix P^μ is irreducible and aperiodic.*

Following this assumption, the Markov decision process induced by the policy μ is ergodic with a unique stationary distribution $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$, a row vector whose entries are positive and sum to 1. It also holds that $\pi_{s'}=\lim_{t\to\infty}(P^\mu)^t(s,s')$ for any two states $s,s'\in[n]$. *Note that we are using π to denote the stationary distribution of P^μ, and not the policy (which is denoted by μ).*

We will use the notation $r_{\max}$ to denote an upper bound on the rewards; more formally, $r_{\max}$ is a real number such that

$$|r(s,s')|\le r_{\max}\text{ for all }s,s'\in[n],a\in\mathcal{A}.$$

Since the number of states is finite, such an $r_{\max}$ always exists.

We next introduce some notation that will make our analysis more concise. For a symmetric positive definite matrix $A\in$

$\mathbb{R}^{n\times n}$, we define the inner product $\langle x,y\rangle_A = x^T Ay$ and the associated norm $\|x\|_A = \sqrt{x^T Ax}$. Let $D = \text{diag}(\pi_1,\cdots,\pi_n)$ denote the diagonal matrix whose elements are given by the entries of the stationary distribution π. Given value functions V and V' on the state space $\mathcal{S}$, we define

$$\langle V,V'\rangle_D = V^T DV' = \sum_{s\in\mathcal{S}} \pi_s V(s)V'(s),$$

and the associated norm

$$\|V\|_D^2 = V^T DV = \sum_{s\in\mathcal{S}} \pi_s V(s)^2.$$

Finally, we define the Dirichlet seminorm, which is often called the Dirichlet form in the Markov chain literature (Diaconis et al., 1996); we follow here the notation of (Ollivier, 2018). The Dirichlet seminorm depends both on the transition matrix P and the invariant measure π:

$$\|V\|_{\text{Dir}}^2 = \frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi_s P(s,s')(V(s')-V(s))^2.$$

It is easy to see that, as a consequence of Assumption 1, $\|V\|_{\text{Dir}} = 0$ if and only if V is a multiple of the all-ones vector.

Similarly, we introduce the k-step Dirichlet seminorm, defined as

$$\|V\|_{\text{Dir},k}^2 = \frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi_s P^k(s,s')(V(s')-V(s))^2.$$

2.2. Policy Evaluation, Temporal Difference Learning, and Linear Function Approximation

Policy evaluation refers to the problem of estimating the value function V^μ for a given stationary policy μ. If the size of the state space is large, computing $V^\mu(s)$ for all states s may be prohibitively expensive. A standard remedy is to use low dimensional approximation V_θ^μ of V^μ in the classical TD algorithm as in Sutton (1988); Sutton & Barto (2018). For brevity, we omit the superscript μ throughout from now on.

The classical TD(0) algorithm with function approximation V_θ starts with an arbitrary value of the parameters θ_0; upon observing the t^{th} transition $s_t \to s_t'$, it computes the scalar-valued temporal-difference error,

$$\delta_t = r(s_t,s_t') + \gamma V_{\theta_t}(s_t') - V_{\theta_t}(s_t),$$

and updates the parameter vector as

$$\theta_{t+1} = \theta_t + \alpha_t \delta_t \nabla V_{\theta_t}(s_t). \tag{1}$$

Here $\nabla V_{\theta_t}(s_t)$ denotes the gradient of the function $V_\theta(s_t)$ w.r.t to θ evaluated at $\theta = \theta_t$, and α_t is the step size. Intuitively, updating in the direction $\delta_t \nabla V_{\theta_t}(s_t)$ moves $V_{\theta_t}(s_t)$ closer to the bootstrapped value of $r(s_t,s_t') + \gamma V_{\theta_t}(s_t')$.

We will be considering the TD(0) algorithm with a linear function approximation V_θ defined as

$$V_\theta(s) = \sum_{l=1}^{K} \theta_l \phi_l(s) \quad \forall s \in \mathcal{S},$$

for a given set of K feature vectors $\phi_l : \mathcal{S} \to \mathbb{R}$, $l \in [K]$. For each state s, we will define the vector $\phi(s)$ which stacks up the features of s as $\phi(s) = (\phi_1(s),\phi_2(s),\cdots,\phi_K(s))^T \in \mathbb{R}^K$. Finally, $\Phi \in \mathbb{R}^{n\times K}$ is defined to be the matrix $\Phi = [\phi_1,\cdots,\phi_K]$.

We thus have that $V_\theta(s) = \theta^T\phi(s)$ and the approximate TD(0) update becomes

$$\theta_{t+1} = \theta_t + \alpha_t(r(s_t,s_t') + \gamma\theta_t^T\phi(s_t') - \theta_t^T\phi(s_t))\phi(s_t). \tag{2}$$

Next we state a common assumption on the feature vectors, which requires that features used for approximation are linearly independent (Tsitsiklis & Van Roy, 1997; Bhandari et al., 2018).

Assumption 2. *The matrix Φ has full column rank, i.e., the feature vectors $\{\phi_1,\ldots,\phi_K\}$ are linearly independent. Additionally, we also assume that $\|\phi(s)\|_2^2 \le 1$ for $s \in \mathcal{S}$.*

It is always possible to make sure this assumption holds. If the norm bound is unsatisfied, then the standard approach is to normalize the feature vectors so that it is. If the matrix Φ does not have full column rank, one can simply omit enough feature vectors so that it does.

It is well-known that under Assumptions 1-2 as well as an additional assumption on the decay of the step-sizes α_t, temporal difference learning converges almost surely; furthermore, its limit is the fixed point of a certain projected Bellman equation (Tsitsiklis & Van Roy, 1997). Henceforth we will use θ^* to denote this fixed point.

It is convenient to introduce the notation

$$g_t(\theta) = \left(r(s_t,s_t') + \gamma\phi(s_t')^T\theta - \phi(s_t)^T\theta\right)\phi(s_t)$$

for the direction taken by TD(0) at time t. Note that $g_t(\theta)$ is a scalar multiple of $\phi(s_t)$, the feature vector of the state encountered at time t.

Furthermore, $\bar{g}(\theta)$ will denote the average of $g_t(\theta)$ when the state is sampled according to the stationary distribution:

$$\bar{g}(\theta) = \sum_{s,s'\in\mathcal{S}} \pi(s)P(s,s')\left(r(s,s') + \gamma\phi(s')^T\theta - \phi(s)^T\theta\right)\phi(s).$$

Naturally it can be seen (see Tsitsiklis & Van Roy (1997)) that

$$\bar{g}(\theta^*) = 0. \tag{3}$$

2.3. Eligibility Traces

We will also study a larger class of algorithms, denoted by TD(λ) and parameterized by $\lambda \in [0,1]$, that contains as a special case the TD(0) algorithm discussed above. While TD(0) makes parameter updates in the direction of the (scaled) last feature vector $g_t(\theta_t)$, the TD(λ) algorithm maintains the "eligibility trace" :

$$z_t = \sum_{k=-\infty}^{t} (\gamma\lambda)^k \phi(s_{t-k}),$$

which is a geometric weighted average of the feature vectors at all previously visited states, and takes a step in the direction of z_t.

In practice, the sum will start at $k = 0$ (or some other finite time); however, parts of the analysis are done with the sum starting at negative infinity because many of the results are much simpler in this setting, and doing so introduces only an exponentially decaying error term.

It is shown in Tsitsiklis & Van Roy (1997) that, subject to Assumptions 1-2 and appropriate decay of step-sizes, TD(λ) converges with probability one, and its limit is a fixed point of a certain projected & averaged Bellman equation. We will denote this limit by θ^*_λ.

2.4. Markov Chain Observation Model

In this paper, we are interested in TD in the setting where the data is collected from a single sample path of a Markov chain. Our final assumption is that the Markov chain mixes at a uniform geometric rate.

Assumption 3. *There are constants $m > 0$ and $\rho \in (0,1)$ such that*

$$\sup_{s \in \mathcal{S}} d_{\mathrm{TV}}(P^t(s,\cdot),\pi) \le m\rho^t \quad t \in \mathbb{N}_0,$$

where $d_{\mathrm{TV}}(P,Q)$ denotes the total-variation distance between probability measures P and Q. In addition, the initial distribution of s_0 is the steady-state distribution π, so that $(s_0, s_1, \cdots)$ is a stationary sequence.

Under Assumption 1, i.e., for irreducible and aperiodic Markov chains, the uniform mixing assumption always holds (Levin & Peres, 2017). It is worth noting that the assumption that s_0 is the the stationary distribution is primarily done to make the analysis and results tidier: given the uniform mixing assumption, one can apply analysis after the Markov chain is close to its steady-state.

3. Temporal Difference Learning as Gradient Splitting

All existing analyses temporal difference learning proceed by comparing it, either explicitly or implicitly, to the expected update, usually referred to as the mean-path update; for TD(0), this is

$$\theta_{t+1} = \theta_t + \alpha_t \bar{g}(\theta_t).$$

Stochastic approximation (Robbins & Monro, 1951) is a common tool to make this comparison. Generally, one wants to argue that the mean-path TD update brings θ_t closer to its final value θ^*.

The first theoretical analysis of TD(0) in Tsitsiklis & Van Roy (1997) proceeded based on the observation that $\bar{g}(\theta)$ forms a positive angle with $\theta^* - \theta$, that is

$$\bar{g}(\theta)^T(\theta^* - \theta) > 0. \tag{4}$$

An explicit version of this inequality was used in (Bhandari et al., 2018) where it is stated as Lemma 3:

$$\bar{g}(\theta)^T(\theta^* - \theta) \ge (1-\gamma)\|V_{\theta^*} - V_\theta\|_D^2. \tag{5}$$

Our main result is an interpretation of the quantity $\bar{g}(\theta)$ which explains why such an inequality holds, as well as allows us to derive stronger results. To do this, we first introduce the concept of a "gradient splitting."

Definition 1. *Let A be a symmetric positive semi-definite matrix. A linear function $h(\theta) = B(\theta - a)$ is called a gradient splitting of the quadratic $f(\theta) = (\theta - a)^T A(\theta - a)$ if*

$$B + B^T = 2A.$$

Note that whenever we state that $h(\theta)$ is a splitting of the gradient $f(\theta)$, this presumes that $h(\theta)$ is a linear function of θ while $f(\theta)$ is a quadratic.

To the best of our knowledge, the concept is introduced here for the first time. We next explain why it is useful.

3.1. Gradient Splitting and Gradient Descent

Observe first that, as one should expect from the name, $(1/2)\nabla f(\theta)$ is a splitting of the gradient of f since

$$\frac{1}{2}\nabla f(\theta) = A(\theta - a).$$

Of course, it is far from the only splitting, since there are many B that satisfy $B + B^T = 2A$. In particular, B may be non-symmetric. For example, one can take B to be equal to the upper triangular part of $2A$ plus the diagonal of A.

The key property of splittings that make them useful is the following.

Proposition 1. *Suppose $h(\theta)$ is a splitting of the gradient of $f(\theta)$. Then*

$$(\theta_1 - \theta_2)^T(h(\theta_1) - h(\theta_2)) = \frac{1}{2}(\theta_1 - \theta_2)^T(\nabla f(\theta_1) - \nabla f(\theta_2)).$$

Proof. Indeed,

$$\begin{aligned}&(\theta_1-\theta_2)^T\left(h(\theta_1)-h(\theta_2)\right)=(\theta_1-\theta_2)^TB(\theta_1-\theta_2)\\ =&\frac{1}{2}(\theta_1-\theta_2)^TB(\theta_1-\theta_2)+\frac{1}{2}(\theta_1-\theta_2)^TB^T(\theta_1-\theta_2)\\ =&(\theta_1-\theta_2)^TA(\theta_1-\theta_2)\\ =&\frac{1}{2}(\theta_1-\theta_2)^T(\nabla f(\theta_1)-\nabla f(\theta_2)).\end{aligned}$$

■

Thus, while $h(\theta)$ may be quite different from $\nabla f(\theta)$, the difference disappears once one looks at the inner products considered in Proposition 1.

A particular consequence of Proposition 1 can be obtained by plugging in $\theta_1=a$, the global minimizer of $f(\theta)$. In that case, $\nabla f(a)=0$ and $h(a)=0$ as well, and we obtain that for all θ,

$$(a-\theta)^Th(\theta)=\frac{1}{2}(a-\theta)^T\nabla f(\theta). \qquad (6)$$

Thus *the splitting $h(\theta)$ has the exact same angle with the "direction to the optimal solution" $a-\theta$ as the true gradient.*

Most analysis of gradient descent on convex functions are ultimately based on the observation that gradient descent "makes progress" towards the optimal solution because it has a positive inner product with the direction to optimality. As a consequence of this discussion, the same argument can be applied to gradient splittings.

3.2. Our Main Contribution

We now come back to temporal difference learning. To analyze TD learning, it is tempting to see if we can write the TD(0) and TD(λ) updates as gradient descent on some appropriately chosen function. Unfortunately, it is well-known (and easy to see) that this cannot work. Indeed, in the TD(0) case, it is possible to express the average direction $\bar{g}(\theta)$ as $\bar{g}(\theta)=B(\theta-\theta^*)$ and in some cases the matrix B is not symmetric; this linear map cannot be the gradient of anything since the non-symmetry of B would contradict equality of partial derivatives (see (Maei, 2011)).

Our main results show that the temporal difference direction can, however, be viewed as a splitting of the gradient of an appropriately chosen function.

Theorem 1. *Suppose Assumptions 1-2 hold. Then in the TD(0) update, $-\bar{g}(\theta)$ is a splitting of the gradient of the quadratic*

$$f(\theta)=(1-\gamma)\|V_\theta-V_{\theta^*}\|_D^2+\gamma\|V_\theta-V_{\theta^*}\|_{\text{Dir}}^2.$$

Theorem 2. *Suppose Assumptions 1-2 hold. Then, in the TD(λ) update, the negative of the expected update $-E[\delta_t z_t]$ is a splitting of the gradient of the quadratic*

$$\begin{aligned}f^{(\lambda)}(\theta)=&(1-\gamma\kappa)\|V_\theta-V_{\theta^*_\lambda}\|_D^2\\ &+(1-\lambda)\sum_{m=0}^{+\infty}\lambda^m\gamma^{m+1}\|V_\theta-V_{\theta^*_\lambda}\|^2_{\text{Dir},m+1},\end{aligned}$$

where $\kappa=(1-\lambda)/(1-\gamma\lambda)$.

The proof of these theorems can be found in the supplementary information. Assumptions 1 and 2 are not particularly crucial: they are used only to be able to define the stationary distribution π and the unique fixed point θ^*.

These results provide some insights into why temporal difference learning works. Indeed, there is no immediate reason why the bootstrapped update of Eq. (2) should produce a reasonable answer, and it is well known that the version with nonlinear approximation in Eq. (1) can diverge (see Tsitsiklis & Van Roy (1997)). Convergence analyses of TD learning rely on Eq. (4), but the proof of this equation from (Tsitsiklis & Van Roy, 1997) does not yield a conceptual reason for why it should hold.

The previous two theorems provide such a conceptual reason. It turns out that TD(0) and TD(λ) are, on average, attempting to minimize the functions $f(\theta)$ and $f^{(\lambda)}(\theta)$ defined in those theorems, by moving in direction of a gradient splitting. Moreover, the functions $f(\theta)$ and $f^{(\lambda)}(\theta)$ are plainly convex (they are positive linear combinations of convex quadratics), so that Equation (6) immediately explains why Eq. (4) holds.

These theorems are inspired by the recent preprint (Ollivier, 2018). It is shown there that, if P is reversible, then $-\bar{g}(\theta)$ is exactly the gradient of the function $f(\theta)$, even in the case when the function approximation is nonlinear. Theorem 1 may be viewed as a way to generalize this observation to the non-reversible case for the case of linear function approximation.

4. Consequences

We now discuss several consequences. These will all be along the lines of improved convergence guarantees. Indeed, as we mentioned in the previous section, viewing TD learning as gradient splitting allows us to take existing results for gradient descent and "port" them almost verbatim to the temporal difference setting.

In the main body of the paper, we focus on TD(0); the case of TD(λ) is discussed in the supplementary information. As mentioned earlier, existing analyses of TD(0) rely on Eq. (4) as well as its refinement Eq. (5). However, as a consequence of Proposition 1, we can actually write out explicitly the inner product between the mean TD(0) direction $\bar{g}(\theta)$ and the direction to optimality $\theta^*-\theta$.

Corollary 1. *For any $\theta \in \mathbb{R}^K$,*

$$(\theta^* - \theta)^T \bar{g}(\theta) = (1-\gamma)\|V_{\theta^*} - V_\theta\|_D^2 + \gamma\|V_{\theta^*} - V_\theta\|_{\text{Dir}}^2.$$

Proof. Indeed, we can use Eq. (3) to argue that

$$\begin{aligned}(\theta^* - \theta)^T \bar{g}(\theta) &= (\theta^* - \theta)^T \left(-\bar{g}(\theta^*) - (-\bar{g}(\theta))\right) \\ &= \frac{1}{2}(\theta^* - \theta)^T (\nabla f(\theta^*) - \nabla f(\theta)), \quad (7)\end{aligned}$$

where the last step follows by Theorem 1 and Proposition 1; here $f(\theta)$ is the function from Theorem 1.

However, for any quadratic function $q(\theta) = (\theta - a)^T P(\theta - a)$ where P is a symmetric matrix, we have that

$$(a - \theta)^T (\nabla q(a) - \nabla q(\theta)) = 2q(\theta).$$

Applying this to the function $f(\theta)$ in Eq. (7), we complete the proof. ■

This corollary should be contrasted to Eq. (4) and Eq. (5). It is clearly a strengthening of those equations. More importantly, this is an equality, whereas Eq. (4) and Eq. (5) are inequalities. We thus see that the average TD(0) direction makes more progress in the direction of the optimal solution compared to the previously available bounds.

In the remainder of the paper, we will use Corollary 1 to obtain improved convergence times for TD(0); also, a natural generalization of that Corollary which appeals to Theorem 2 instead of Theorem 1 results in improved convergence times for TD(λ), as explained in the supplementary information. We focus on a particular property which is natural in this context: scaling with the discount factor γ.

Indeed, as we discussed in the introduction, an undesirable feature of some of the existing analyses of temporal difference learning is that they scale multiplicatively with $1/(1-\gamma)$. It is easy to see why this should be so: it is natural to rely on Eq. (5) or its variations, and as $\gamma \to 1$, that equation guarantees smaller and smaller progress towards the limit. Unfortunately, it is natural to set the discount factor close to 1 in order to avoid focusing on short-term behavior of the policy.

But now we can instead rely on Corollary 1 and this corollary suggests that as $\gamma \to 1$, the inner product between the expected TD(0) direction $\bar{g}(\theta)$ and the direction to the optimal solution $\theta^* - \theta$ will be lower bounded by $\gamma||V_\theta - V_{\theta^*}||_{\text{Dir}}^2$. A difficulty, however, is that the Dirichlet seminorm can be zero even when applied to a nonzero vector. We next discuss the results we are able to derive with this approach.

4.1. Improved Error Bounds

As mentioned earlier, a nice consequence of the gradient splitting interpretation is that we can apply the existing proof for gradient descent almost verbatim to gradient splittings. In particular, temporal difference learning when the states are sampled i.i.d. could be analyzed by simply following existing analyses of noisy gradient descent. However, under our Markov observation model, it is not true that the samples are i.i.d; rather, we proceed by modifying the analysis of so-called *Markov Chain Gradient Descent*, analyzed in the papers (Sun et al., 2018; Johansson et al., 2010), where, to minimize the function $F(x) = \sum_i f_i(x)$, we have access to samples $\nabla f_{i_1}(\cdot), \nabla f_{i_2}(\cdot), \ldots$, with the sequence i_k following a Markov chain.

One issue is the choice of step-size. The existing literature on temporal difference learning contains a range of possible step-sizes from $O(1/t)$ to $O(1/\sqrt{T})$ (see (Bhandari et al., 2018; Dalal et al., 2018; Lakshminarayanan & Szepesvari, 2018)). A step-size that scales as $O(1/\sqrt{T})$ is often preferred because, for faster decaying step-sizes, performance will scale with the inverse of the smallest eigenvalue of $\Phi^T D\Phi$ or related quantity, and these can be quite small[1]. This is not the case, however, for a step-size that decays like $O(1/\sqrt{T})$.

We will be using the standard notation

$$\bar{\theta}_T = \frac{1}{T}\sum_{t=0}^{T-1} \theta_t$$

to denote the running average of the iterates θ_t.

We will be considering the projected TD(0)[2] update

$$\theta_{t+1} = \text{Proj}_\Theta(\theta_t + \alpha_t g_t(\theta_t)), \quad (8)$$

where Θ is a convex set containing the optimal solution θ^*. Moreover, we will assume that the norm of every element in Θ is at most R_Θ. Setting $G = r_{\max} + 2R_\Theta$, we have the following error bound.

Corollary 2. *Suppose Assumptions 1-3 hold. Suppose further that $(\theta_t)_{t\geq 0}$ is generated by the Projected TD algorithm of Eq. (8) with $\theta^* \in \Theta$ and $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Then*

$$\begin{aligned}&E\left[(1-\gamma)\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 + \gamma\|V_{\theta^*} - V_{\bar{\theta}_T}\|_{\text{Dir}}^2\right] \\ &\quad\leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2\sqrt{T}}, \quad (9)\end{aligned}$$

where τ^{mix} is standard notation for the mixing time of the Markov chain:

$$\tau^{\text{mix}}(\varepsilon) = \min\left\{t \in \mathbb{N}, t \geq 1 | m\rho^t \leq \varepsilon\right\}.$$

[1] For example, λ in Theorems of Dalal et al. (2018), ρ_d in Theorems of Lakshminarayanan & Szepesvari (2018), and ω in the Theorem 3 of Bhandari et al. (2018)

[2] The challenges of analyzing Markovian samples and the reason of using a projected step has been discussed in Section 8 of Bhandari et al. (2018)

The proof is available in the supplementary information. The bound is very similar to the standard bounds for SGD, with the exception of the τ^{mix} term. That term arises in the analysis of Markov gradient descent (Sun et al., 2018; Johansson et al., 2010). Informally, in Markov gradient descent, one has to wait τ^{mix} iterations to obtain a new independent sample of the gradient, which is why τ^{mix} enters the bound multiplicatively.

We next compare this bound to the existing literature. The closest comparison is Theorem 3(a) in (Bhandari et al., 2018) which shows that

$$E\left[||V_{\theta^*} - V_{\bar{\theta}_T}||_D^2\right] \leq \frac{||\theta^* - \theta_0||_2^2}{2(1-\gamma)\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2(1-\gamma)\sqrt{T}}. \quad (10)$$

Corollary 2 is stronger than this, because this bound can be derived from Corollary 2 by ignoring the second term on the left hand side of Eq. (9). Moreover, we next argue that Corollary 2 is stronger an interesting way, in that it offers a new insight on the behavior of temporal difference learning.

Observe that the upper bound of Eq. (10) blows up as $\gamma \to 1$. On the other hand, we can simply ignore the first term on the left-hand side of Corollary 2 to obtain

$$E\left[||V_{\theta^*} - V_{\bar{\theta}_T}||_{\text{Dir}}^2\right] \leq \frac{||\theta^* - \theta_0||_2^2}{2\gamma\sqrt{T}} + \frac{G^2\left[9 + 12\tau^{\text{mix}}\left(1/\sqrt{T}\right)\right]}{2\gamma\sqrt{T}}. \quad (11)$$

In particular, we see that $E\left[||V_{\theta^*} - V_{\bar{\theta}_T}||_{\text{Dir}}^2\right]$ does not blow up as $\gamma \to 1$. To understand this, recall that the Dirichlet seminorm is equal to zero if and only if applied to a multiple of the all-ones vector. Consequently, $||V||_{\text{Dir}}$ is properly thought of as a way to measure norm of the projection of V onto $\mathbf{1}^{\perp}$. *We therefore obtain the punchline of this section: the error of (averaged & projected) temporal difference learning projected on $\mathbf{1}^{\perp}$ does not blow up as the discount factor approaches 1.*

There are scenarios where this is already interesting. For example, if TD(0) is a subroutine of policy evaluation, it will be used for a policy improvement step, which is clearly unaffected by adding a multiple of the all-ones vector to the value function. Similarly, Proposition 4 of (Ollivier, 2018) shows that the bias in the policy gradient computed from an approximation $\hat{V}$ to the true value function V can be bounded solely in terms of $||V - \hat{V}||_{\text{Dir}}^2$ (multiplied by a factor that depends on how the policies are parameterized).

It is natural to wonder whether the dependence on $1/(1-\gamma)$ can be removed completely from bounds on the performance of temporal difference learning (not just in terms of projection on $\mathbf{1}^{\perp}$). We address this next.

4.2. Mean-adjusted Temporal Difference Learning

Unfortunately, it is easy to see that the dependence on $1/(1-\gamma)$ in error bounds for temporal difference learning cannot be entirely removed. We next give an informal sketch explaining why this is so. We consider the case where samples s, s' are i.i.d. with probability $\pi_s P(s,s')$ rather than coming from the Markov chain model, since this only makes the estimation problem easier.

Estimating the mean of the value function. Let us denote by V the true value function; because it is the fixed point of the Bellman operator, $V = R + \gamma P V$, we have that

$$V = (I - \gamma P)^{-1} R = \left(\sum_{m=0}^{\infty} \gamma^n P^m\right) R.$$

Define $\bar{V} = \pi^T V$; then

$$\bar{V} = \pi^T V = \pi^T \left(\sum_{m=0}^{\infty} \gamma^n P^m\right) R = \frac{\pi^T R}{1-\gamma}. \quad (12)$$

Under i.i.d. sampling, what we have are samples from a random variable $\tilde{R}$ which takes the value $r(s,s')$ with probability $\pi_s P(s,s')$. From Eq. (12), we have that

$$(1-\gamma)\bar{V} = E[\tilde{R}].$$

From T samples of the scalar random variable $\tilde{R}$, the best estimate $\hat{R}_T$ will satisfy

$$E[(\hat{R}_T - E[\tilde{R}])^2] = \Omega(1/T)$$

in the worst-case. This implies that the best estimator $\hat{V}$ of $\bar{V}$ will satisfy

$$E[(\hat{V} - \bar{V})^2] = \Omega\left((1/(1-\gamma)^2)/T\right).$$

To summarize, the squared error in estimating just the mean of the value function will already scale with $1/(1-\gamma)$. If we consider e.g., Φ to be the identity matrix, in which case V_{θ^*} is just equal to the true value function, it can easily be seen that it is not possible to estimate V_{θ^*} with error that does not scale with $1/(1-\gamma)$.

A better scaling with the discount factor. Note, however, that the previous discussion implied that a term like $(1/(1-\gamma)^2)/T$ in a bound on the squared error is unavoidable. But with $1/\sqrt{T}$ step-size, the error will in general decay more slowly as $1/\sqrt{T}$ as in Corollary 2. Is it possible to derive a bound where the only scaling with $1/(1-\gamma)$ is in the asymptotically negligible $O(1/T)$ term?

As we show next, this is indeed possible. The algorithm is very natural given the discussion in the previous subsection: we run projected and averaged TD(0) and estimate the mean of the value function separately, adjusting the outcome of TD(0) to have the right mean in the end. Building on Corollary 2, the idea is that the mean will have expected square error that scales with $(1/(1-\gamma)^2)/T$ while the temporal difference method will estimate the projection onto $\mathbf{1}^\perp$ without blowing up as $\gamma \to 1$.

The pseudocode of the algorithm is given next as Algorithm 1 and Corollary 3 bounds its performance. Note that, in contrast to the bounds in the last subsection, Corollary 3 bounds the error to the true value function V directly. This is a more natural bound for this algorithm which tries to directly match the mean of the true value function.

Algorithm 1 Mean-adjusted TD(0)

```
Initialize $\bar{A}_0 = 0$, $s_0 \sim \pi$, and some initial condition $\theta_0$.
for $t = 0$ to $T-1$ do
    Projected TD(0) update:
    $\theta_{t+1} = \mathrm{Proj}_\Theta\left(\theta_t + \alpha_t g_t(\theta_t)\right)$
    Keep track of the average reward: $\bar{A}_{t+1} = \frac{t\bar{A}_t + r_{t+1}}{t+1}$
end for
Set $\hat{V}_T = \frac{\bar{A}_T}{1-\gamma}$
Output $V'_T = V_{\bar{\theta}_T} + \left(\hat{V}_T - \pi^T V_{\bar{\theta}_T}\right)\mathbf{1}$
```

Corollary 3. *Suppose that $(\theta_t)_{t\geq 0}$ and V'_T are generated by Algorithm 1 with step-sizes $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$. Suppose further that Θ is a convex set that contains θ^*. Let t_0 be the largest integer which satisfies $t_0 \leq 2\tau^{\mathrm{mix}}\left(\frac{1}{2(t_0+1)}\right)$. Then as long as $T \geq t_0$, we will have*

$$E\left[\|V'_T - V\|_D^2\right] \leq O\left(E\left[\|V_{\theta^*} - V\|_D^2\right] + \frac{r_{\max}^2 \tau^{\mathrm{mix}}\left(\frac{1}{2(T+1)}\right)}{(1-\gamma)^2 T} + \frac{\|\theta^* - \theta_0\|_2^2 + G^2\left[1 + \tau^{\mathrm{mix}}(1/\sqrt{T})\right]}{\sqrt{T}} \min\left\{\frac{r(P)}{\gamma}, \frac{1}{1-\gamma}\right\}\right).$$

Here $r(P)$ is the inverse spectral gap of the additive reversibilization of the transition matrix P, defined as follows:

$$r(P) = \frac{1}{1-\lambda_2(Q)},$$

where $\lambda_2(Q)$ is the second-largest eigenvalue of the matrix Q in turn defined as

$$Q = \frac{P + P^*}{2},$$

where, finally, P^* denotes the transition matrix of the reversed Markov chain, i.e.,

$$P^*(j|i) = \frac{\pi(j)}{\pi(i)} P(i|j).$$

Let us parse the bound of Corollary 3. The bound has three terms. The first term is just the difference between the limit of TD(0) and the true value function; such a term is inevitable in any TD(0)-based method that compares its performance to the true value function (rather than to V_{θ^*}). The second term comes from the error in mean estimation; as described earlier, scaling with $(1/(1-\gamma))^2/T$ is inevitable here. The multiplicative factor of τ^{mix} is present because, due to the Markovian sampling, it can take τ^{mix} steps to obtain an independent sample of the mean.

Finally, the last term of the bound of Corollary 3 is the one that scales as $\widetilde{O}\left(1/\sqrt{T}\right)$; compared to it, the second term is negligible for large T. Crucially, this term does not blow up as $\gamma \to 1$, so that the only blowup occurs in the asymptotically negligible second term. Indeed, observe that while $1/(1-\gamma)$ appears in the third term, it appears as a minimum of $1/(1-\gamma)$ and a quantity that depends on the matrix P, so that it does not blow up as $\gamma \to 1$.

Note that, unlike the previous bounds discussed in this paper, this bound does depend on an eigenvalue gap associated with (a function of) the matrix P. However, this dependence is in such a way that it only helps: when $1/(1-\gamma)$ is small, there is no dependence on the eigenvalue gap, and it is only when $\gamma \to 1$ that performance "saturates" at something depending on P.

5. Conclusion

We have provided an interpretation of temporal difference learning in terms of a splitting of gradient descent. As a consequence of this interpretation, analyses of gradient descent can apply to temporal difference learning almost verbatim.

We have exploited this interpretation to observe that temporal difference methods learn the projection of the value function onto $\mathbf{1}^\perp$ without any blowup as $\gamma \to 1$; by contrast, previous work tended to have error bounds that scaled with $1/(1-\gamma)$. While, as we explain, it is not possible to remove the dependence on $O(1/(1-\gamma))$ in general, we provide an error bound for a simple modification to TD(0) where the only dependence on $1/(1-\gamma)$ is in the asymptotically negligible term.

An open problem might be to improve the scaling of the bounds we have obtained in this paper with P. Our focus has been on scaling with $1/(1-\gamma)$ but one could further ask what dependence on the transition matrix P is optimal. It is natural to wonder, for example, whether the $r(P)$ factor in the last term of Corollary 3 measuring how the performance "saturates" as $\gamma \to 1$, could be improved. Typically, error bounds in this setting scale with the spectral gap of P, which can be much smaller than $r(P)$. We thus do not believe our bound is tight.

More broadly, if temporal difference learning is a splitting of gradient descent, this opens up several possibilities for future work. For example, one might wonder whether there are other, more attractive, ways to split gradient descent amenable to a bootstrapped interpretation. Alternatively, instead of splitting gradient descent, one might attempt to split mirror descent whenever there are further constraints on the parameter θ.

References

Aldous, D. and Fill, J. Reversible markov chains and random walks on graphs, 1995.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pp. 1691–1692, 2018.

Chen, C., Seff, A., Kornhauser, A., and Xiao, J. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2722–2730, 2015.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analysis for td (0) with linear function approximation. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Diaconis, P., Saloff-Coste, L., et al. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2193–2201, 2018.

Johansson, B., Rabi, M., and Johansson, M. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2010.

Korda, N. and La, P. On td (0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*, pp. 626–634, 2015.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355, 2018.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Maei, H. R. *Gradient Temporal-Difference Learning Algorithms*. PhD thesis, University of Alberta, 2011.

Narayanan, C. and Szepesvári, C. Finite time bounds for temporal difference learning with function approximation: Problems with some “state-of-the-art” results. Technical report, Technical Report, 2017.

Ollivier, Y. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on Learning Theory*, pp. 2803–2830, 2019.

Sun, T., Sun, Y., and Yin, W. On markov chain gradient descent. In *Advances in Neural Information Processing Systems*, pp. 9896–9905, 2018.

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

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 2018.

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Xiong, H., Xu, T., Liang, Y., and Zhang, W. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. *arXiv preprint arXiv:2002.06286*, 2020.

Xu, T., Wang, Z., Zhou, Y., and Liang, Y. Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*, 2019a.

Xu, T., Zou, S., and Liang, Y. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems*, pp. 10633–10643, 2019b.