Finite Time Bounds for Temporal Difference Learning with Function Approximation: Problems with some "state-of-the-art" results

Chandrashekar Lakshmi Narayanan Csaba Szepesvári

Abstract

In all branches of mathematics, including learning theory, results build on previous results. Thus, it is important to keep the literature free of erroneous claims. This short report lists some problems with the proofs and claims in the recent papers by Prashanth et al. [2014], Korda and Prashanth [2015], whose longer version containing the proofs are available on arxiv [Prashanth et al., 2013, Korda and Prashanth, 2014]. In particular, it follows that the results in these papers, if true, would need completely new proofs, and thus should not be used in the form stated by the authors.

1 Introduction

This short report lists some problems with the proofs of the claims in two recent papers by Prashanth et al. [2014], Korda and Prashanth [2015]. As the problems seem serious, our conclusion is that to obtain results similar to those claimed in these two works, the assumptions of the stated claims need to be considerably strengthened and the form of the results will also need to be adjusted in significant ways.

The report is not self contained and its scope is limited to an audience who are interested in RL and TD(0) and specifically the above-mentioned papers. Furthermore, since Prashanth et al. [2014], Korda and Prashanth [2015] omitted the proofs, we will instead discuss their longer version in what follows, which have identical statements but include the proofs. The longer version of the paper by Prashanth et al. [2014] is [Prashanth et al., 2013], while the longer version of Korda and Prashanth [2015] is [Korda and Prashanth, 2014]. We will borrow the notation directly from these works.

2 Expected Error Bound

2.1 Bugs in the paper by Prashanth et al. [2013]

One of the main results of Prashanth et al. [2013] is Theorem 1, which states a bound on the expected error. The proof of this theorem can be found in section A.2, starting on page 16. The proof up to Eq. (26) is correct (in the definition of M_{n+1}, F_n should be F). However, after this display, we are told that by (A4), which ensures that $\frac{1}{T}\sum_{i=1}^{T}\phi(s_i)\phi_t(s_i)^\top \succ \mu I$ for some $\mu > 0$, $\bar{A}_n = \frac{1}{n}\sum_{t=1}^{n}\phi_t(\phi_t - \beta\phi'_t)^\top$ is such that $\bar{A}_n - (1-\beta)\mu I$ is positive definite. Here, $n > 0$ is an arbitrary index and for $t \geq 1$ we use the abbreviation $\phi_t = \phi(s_{i_t})$ and $\phi'_t = \phi(s'_{i_t})$, where $i_t \in \{1, \ldots, T\}$ is a random index.

In general, (A4) does *not* imply that $\bar{A}_n - (1-\beta)\mu I$ is positive definite.

Take for example $n = 1$. We would need that $\phi_1(\phi_1 - \beta\phi_1')^\top - (1-\beta)\mu I$ is positive definite. It is easy to construct examples where this is not true: Nothing prevents, for example, $\phi_1 = \beta\phi_1'$, in which case $\bar{A}_1 - (1-\beta)\mu I = -(1-\beta)\mu I$ is *negative* definite. (Note that the matrices involved are *not* symmetric. Unfortunately, none of the two papers defines what is meant by positive definite in this case. We assume that the definition used is that a square, real-valued matrix A is called positive definite if $x^\top Ax \geq 0$ for any x real-valued vector of appropriate dimension.) In fact, we don't see why the claimed relationship would hold even when $\bar{A}_n$ is replaced by $\hat{A}_T \doteq \frac{1}{T}\sum_{i=1}^T \phi(s_i)(\phi(s_i) - \beta\phi(s_i'))^\top$, and we in fact suspect that this claim is false in full generality. But at minimum, a proof would be required and the whole subsequent argument will need to be changed.

2.2 Bugs in the paper by Korda and Prashanth [2014]

In page 14 the expression below "A is a possibly random matrix..." is not justified (personal communication with one of the authors confirmed this). In particular, the claim here is that if A is a random matrix with $\|A\|_2 \leq C$ with C a deterministic constant then for any θ deterministic vector,

$$\mathbf{E}\left[\theta^\top A^\top \mathbf{E}\left[\epsilon_n|\mathcal{F}_n\right] A\theta \,|\, s_0\right] \leq C^2\theta^\top \mathbf{E}\left[\epsilon_n|s_0\right]\theta\,.$$

Recall that here ϵ_n is a matrix of appropriate dimensions, the "mixing-error term" and is actually $\mathcal{F}_n$ measurable ($\epsilon_n = \mathbf{E}\left[a_n|\mathcal{F}_n\right] - \mathbb{E}_\Psi[a_n]$). When the Markov chain is started from its stationary state (which is not ruled out by the conditions of the theorem under question), $\mathbf{E}\left[\epsilon_n|s_0\right] = 0$. If the above inequality was true, we would get

$$\mathbf{E}\left[A^\top \epsilon_n A \,|\, s_0\right] = 0\,.$$

However, letting $B = \mathbf{E}\left[\epsilon_n|\mathcal{F}_n\right]$ and, for example, $A = CB/\left\|B\right\|_2$, we have

$$A^\top \mathbf{E}\left[\epsilon_n|\mathcal{F}_n\right] A = \frac{C^2}{\|B\|_2^2} B^\top BB$$

and it is easy to construct examples where the expectation of this is nonzero.

Also in page 14, we find the following inequality

$$(d+2)\left(e^{2(1+\beta)\sum_{k=1}^n \gamma_k}\left\|\theta_0\right\|_2^2 + e^{2(1+\beta)\sum_{k=1}^n \gamma_k}\left(\sum_{k=1}^n \gamma_k e^{-(1+\beta)\sum_{j=1}^{k-1}\gamma_j}\right)^2 + \left\|\theta^*\right\|_2^2\right)\left\|\mathbf{E}(\epsilon_n|s_0)\right\|_2$$
$$\leq (d+2)\frac{\left\|\theta_0\right\|_2^2 + 1 + \left\|\theta^*\right\|_2^2}{(1-\beta)^2} e^{2(1+\beta)\sum_{k=1}^n \gamma_k}\left\|\mathbf{E}(\epsilon_n|s_0)\right\|_2\,,$$

where it is not clear as to how the $(1-\beta)^2$ factor appears in the denominator (also confirmed by a personal communication with one of the authors).

3 Problems with the Proof: High Probability Bound

3.1 Bugs in the paper by Prashanth et al. [2013]

The proof of the high probability bound starts on page 14, in section A.1. The first problem happens in the display on the bottom of this page in the proof of Lemma 6. Here, we are told

that for $A = \phi\phi^\top - \beta\phi(\phi')^\top$ (we are dropping indices to remove clutter),

$$A^\top A = \|\phi\|^2 \phi\phi^\top - \beta(2 - \|\phi\|^2 \beta)\phi'(\phi')^\top ,$$

where $\|x\|$ denotes the 2-norm of x. However, using $A = \phi(\phi - \beta\phi')^\top$, a direct calculation gives:

$$\begin{aligned} A^\top A &= (\phi - \beta\phi')\phi^\top\phi(\phi - \beta\phi')^\top \\ &= \|\phi\|^2 (\phi - \beta\phi')(\phi - \beta\phi')^\top \\ &= \|\phi\|^2 \left\{\phi\phi^\top - \beta(\phi'\phi^\top + \phi(\phi')^\top) + \beta^2\phi'(\phi')^\top\right\} , \end{aligned}$$

which does not match the previous display. The terms that do not match are the linear-in-β terms. In the first display we have $-2\beta\phi'(\phi')^\top$, while in the bottom we have $-\beta(\phi'\phi^\top + \phi(\phi')^\top)$.

The first equality of their display states (in equivalent form) that

$$A^\top A = \|\phi\|^2 (\phi\phi^\top - 2\beta\phi(\phi')^\top + \beta^2\phi'(\phi')^\top) .$$

We see that while this is closer to the correct result, here the mistake is that $-2\beta\phi(\phi')^\top$ is replacing $-\beta(\phi(\phi')^\top + \phi'\phi^\top)$.

3.2 Bugs in the paper by Korda and Prashanth [2014]

(In this paper, s_t is a sequence of states obtained whole following a fixed policy in an MDP.) In page 10 of Korda and Prashanth [2014] the expression for $\mathbf{E}[a_{j+1}^\top a_{j+1}]$ contains terms that involve the product of P and $P^\top$. This cannot be correct, as here we can take the expectation first over the next state, which will bring in a *single* instance of P. To remove clutter, drop the j subindex, and set $A = \phi(\phi - \beta\phi')^\top$, where $\phi = \phi(s_j)$ and $\phi' = \phi(s_{j+1})$. The incriminated expression from Eq. (15) of the paper is

$$\mathbf{E}\left[A - \tfrac{\gamma}{2}A^\top A\right] = \Phi^\top(I - \beta\Psi P - \tfrac{\gamma}{2}(\Delta - \beta P^\top(2I - \beta\Delta)\Psi P))\Phi . \tag{1}$$

Here, Φ is the $S \times d$ matrix whose sth row is $\phi^\top(s)$ ($s \in \{1, \dots, S\}$), Ψ is the $S \times S$ diagonal matrix whose ith diagonal entry is $\mathbb{P}(s_t = i)$, while Δ is another $S \times S$ diagonal matrix whose sth entry is $\|\phi(s)\|_2^2$. A direct calculation (as before) gives that

$$\begin{aligned} A^\top A &= (\phi - \beta\phi')\phi^\top\phi(\phi - \beta\phi')^\top \\ &= \|\phi\|^2 (\phi - \beta\phi')(\phi - \beta\phi')^\top \\ &= \|\phi\|^2 \left\{\phi\phi^\top - \beta(\phi'\phi^\top + \phi(\phi')^\top) + \beta^2\phi'(\phi')^\top\right\} \\ &= \|\phi\|^2 \phi\phi^\top - \beta\|\phi\|^2 \phi'\phi^\top - \beta\|\phi\|^2 \phi(\phi')^\top + \beta^2\|\phi\|^2 \phi'(\phi')^\top . \end{aligned} \tag{2}$$

The expectation of each terms are as follows:

$$\begin{aligned} \mathbf{E}\left[\|\phi\|^2 \phi\phi^\top\right] &= \sum_s \mathbb{P}(s_t = s)\|\phi(s)\|^2 \phi(s)\phi(s)^\top = \Phi^\top\Delta\Psi\Phi , \\ \beta\mathbf{E}\left[\|\phi\|^2 \phi'\phi^\top\right] &= \beta\sum_s \mathbb{P}(s_t = s)\|\phi(s)\|^2 \left\{\sum_{s'} P(s'|s)\phi(s')\right\}\phi(s)^\top = \beta(P\Phi)^\top\Delta\Psi\Phi , \\ \beta\mathbf{E}\left[\|\phi\|^2 \phi(\phi')^\top\right] &= \beta\mathbf{E}\left[\|\phi\|^2 \phi'\phi^\top\right]^\top = \beta\left\{(P\Phi)^\top\Delta\Psi\Phi\right\}^\top = \beta(\Phi^\top\Delta\Psi P\Phi) , \\ \mathbf{E}\left[\beta^2\|\phi\|^2 \phi'(\phi')^\top\right] &= \beta^2\sum_s P(s_t = s)\|\phi(s)\|^2 \left\{\sum_{s'} P(s'|s)\phi(s')\phi(s')^\top\right\} . \end{aligned}$$

Further,

$$\mathbf{E}\left[A\right] = \Phi^\top \Psi (I - \beta P)\Phi\,.$$

Putting together things we see the mismatch with (1). To see this even more clearly, assume that $\|\phi(s)\|^2 = 1$ for any $s \in \{1, \ldots, S\}$. Then, $\Delta = I$, and by stationarity, $\mathbb{P}\left(s_{t+1} = s\right) = \mathbb{P}\left(s_t = s\right)$, hence,

$$\mathbf{E}\left[\beta^2 \left\|\phi\right\|^2 \phi'(\phi')^\top\right] = \beta^2 \mathbf{E}\left[\phi(\phi)^\top\right] = \beta^2 \Phi^\top \Psi \Phi\,.$$

Thus,

$$\begin{aligned}\mathbf{E}\left[A^\top A\right] &= \Phi^\top \Psi \Phi - \beta \Phi^\top P^\top \Psi \Phi - \beta \Phi^\top \Psi P \Phi + \beta^2 \Phi^\top \Psi \Phi \\ &= \Phi^\top \left(\Psi - \beta(P^\top \Psi + \Psi P) + \beta^2 \Psi\right) \Phi\end{aligned}$$

and hence

$$\begin{aligned}\mathbf{E}\left[A - \frac{\gamma}{2} A^\top A\right] &= \Phi^\top \Psi (I - \beta P)\Phi - \tfrac{\gamma}{2} \Phi^\top \left(\Psi - \beta(P^\top \Psi + \Psi P) + \beta^2 \Psi\right) \Phi \\ &= \Phi^\top \left\{\Psi(I - \beta P) - \tfrac{\gamma}{2}\left(\Psi - \beta(P^\top \Psi + \Psi P) + \beta^2 \Psi\right)\right\} \Phi\,.\end{aligned}$$

while (1) gives

$$\begin{aligned}\mathbf{E}\left[A - \tfrac{\gamma}{2} A^\top A\right] &= \Phi^\top (I - \beta \Psi P - \tfrac{\gamma}{2}(I - \beta P^\top (2I - \beta I)\Psi P))\Phi \\ &= \Phi^\top \left\{I - \beta \Psi P - \tfrac{\gamma}{2}(I - 2\beta P^\top \Psi P + \beta^2 P^\top \Psi P)\right\} \Phi\,.\end{aligned}$$

Choosing $\Phi = I$, we find that the two expressions are equal if and only if

$$\Psi(I - \beta P) - \tfrac{\gamma}{2}\left(\Psi - \beta(P^\top \Psi + \Psi P) + \beta^2 \Psi\right) = I - \beta \Psi P - \tfrac{\gamma}{2}(I - 2\beta P^\top \Psi P + \beta^2 P^\top \Psi P)\,,$$

which implies, e.g., that $\Psi = I$ (by choosing $\gamma = 0$), which is not possible since the diagonal elements of Ψ must sum to one. Even if we correct the first identity to Ψ, we see that we must have

$$\Psi - \beta(P^\top \Psi + \Psi P) + \beta^2 \Psi = I - 2\beta P^\top \Psi P + \beta^2 P^\top \Psi P\,,$$

which again, means that $\Psi = I$, and also that $P^\top \Psi + \Psi P = P^\top \Psi P$ and that $\Psi = P^\top \Psi P$. The first equality is always false, and the others are false except (perhaps) in some very special cases.

4 Issues with the Setup

4.1 Boundedness of iterates: Prashanth et al. [2013]

Prashanth et al. [2013] assume that the parameter vector stays such that the value function $\Phi\theta$ will be bounded in L^∞-norm (see assumption (A3) of Prashanth et al. [2014] and Prashanth et al. [2013]). This assumption is critical in establishing Lemma 7 (see pages 15 and 16, Prashanth et al. [2013]), in an argument that is similar to the proof of McDiarmid's inequality. We suspect the following shortcomings with assumption (A3):

- The assumption is stated in a somewhat sloppy fashion. We take the authors meant to say that $\sup_n \|\Phi\theta_n\|_\infty < +\infty$ holds almost surely. This seems like a strong assumption: ensuring this will most likely further restrict the step-size sequences that can be used. The step-size sequences that give the best rate under (A3) may include step-size sequences which in fact lead to $\mathbb{P}\left(\limsup_{n\to\infty} \|\theta_n\| = \infty\right) > 0$. Without proving that this is not the case, the results of the paper have limited utility.
- One possibility would be to modify the algorithm by adding a projection step to guarantee boundedness. It is still unclear whether this alone would ensure convergence of the error to zero. In any case, the expected error bound analysis is invalidated if a projection step is present (basically, the algebraic identities will all fail to hold) and a new proof will be required.

4.2 Boundedness of iterates: Korda and Prashanth [2014]

Dalal et al. [2017] mention that (citing a personal communication with Korda and Prashanth [2014]) that Korda and Prashanth [2014] assume implicitly a projection step in all the high probability bounds. While this implicit projection in itself does not affect the high probability bound proofs directly, the algebraic steps are invalidated. Furthermore, the set that the iterates are projected to should contain the TD(0) solution. How to ensure this (without knowing A, b) remains to be seen.

4.3 Relation between Covariance Matrix and $\bar{A}_T$ matrix

Prashanth et al. [2013] assume positive definiteness of $(A4)$ covariance matrix $\frac{1}{T}\Phi_T^\top\Phi_T$. However, unlike regression problems, in reinforcement learning problems what appears in the recursion (see Equation (6)) is not the covariance matrix, but a different matrix $\bar{A}_T = \frac{1}{T}\sum_{i=1}^{T}\phi(s_i)(\phi(s_i) - \beta\phi(s_i'))^\top$ defined in pages 2, 4 (below Equation (5)), 8 and 16 of Prashanth et al. [2013]. Usually, without a sampling assumption known as the 'on-policy' case (see Sutton et al. [2009] for a discussion on 'on-policy' vs 'off-policy') the eigenvalues of $\bar{A}_T$ cannot be guaranteed to have all positive real parts. While Prashanth et al. [2013] mention the 'on-policy' sampling in the introduction, there is no explicit sampling assumption in the list of assumption. In fact, we doubt that the proposed algorithm will converge without extra assumption (as discussed above).

4.4 Blow Up of the Bound

We would like to note that the rate expression in Corollary 4 of Prashanth et al. [2013] (or Corollary 2 of Prashanth et al. [2014]) contains a constant C. The authors do mention that the sampling error (a.k.a. variance) blows up as $\alpha \to 1$. However, it also looks like that even the constant $C = \sum_{n=1}^{\infty}\exp(-\mu c n^{1-\alpha})$ (appearing in the bound of the bias) will blow up as $\alpha \to 1$, in which case it seems that the claim that the $1/\sqrt{n}$ rate can be achieved in the limit will not hold.

4.5 Doubt about the Effectiveness of the Algorithms

In Corollary 4 of Prashanth et al. [2013] (or Corollary 2 of Prashanth et al. [2014]), we learn that the value of c governing the stepsize of the primary update must be in a small range (it must be between 1.33 and 2). This means, that effectively, the stepsize γ_n behaves as $1/n^\alpha$ (c has very little effect). At least when $\alpha = 1$, we know that stepsizes like this make the bias decrease slowly and averaging remains ineffective. This seems to be at odds with the suggestion after this result

that $\alpha \to 1$ is a desirable choice. In fact, we would be inclined to choose $\alpha = 1/2$, i.e., its lowest value. This is because then the bias is decreased relatively quickly, while the variance will be controlled by the additional averaging. However, given all the problems with this, it remains to be seen whether this is indeed a reasonable choice and under exactly what conditions this is reasonable.

References

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analysis for TD(0) with linear function approximation. *CoRR*, abs/1704.01161, 2017.

Nathaniel Korda and LA Prashanth. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. *CoRR*, abs/1411.3224, 2014.

Nathaniel Korda and LA Prashanth. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *ICML*, pages 626–634, 2015.

LA Prashanth, Nathaniel Korda, and Rémi Munos. Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control. *arXiv preprint arXiv:1306.2557*, 2013.

LA Prashanth, Nathaniel Korda, and Rémi Munos. Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 66–81. Springer, 2014.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000. ACM, 2009.