We thank all the reviewers for their encouraging comments.

Response to Reviewer 1:

1. We can give bounds for the case of $k \leq \tau$ but such bounds are not very insightful since they scale linearly with τ (see Lemma 3 in our supplementary material and Theorem 3 in Bhandari et al., 2018 - https://bit.ly/2YpL5Bm). Moreover, in prior work (e.g., Bhandari et al, 2018), when there are bounds for $k \geq 0$, the underlying assumption is either that the noise is i.i.d. or, if it is Markovian, it is in steady state starting at $k = 0$. In both these cases, τ is effectively zero.

2. The adaptive learning rate rule does reduce the step-size, but the decay is not periodic. The rule first tests whether the algorithm is in steady-state, and if so, then the step-size is reduced. The number of iterations to reach steady-state will be different for different step-sizes and hence the decay will not be periodic. We do expect the rule to converge to the optimal parameter vector, although a proof of this result is a direction for further research.

3. The Konda-Tsitsiklis paper suggested by the reviewer seems to assume i.i.d noise, but it is an important reference and we will cite it. We would like to mention that we had implemented the simulations using $\alpha = 1$ and $\beta = \frac{2}{3}$, but we reported the results for $\alpha = 0.99$ and $\beta = 0.66$ since Dalal et al., 2017 shows that $\alpha \to 1$ and $\beta \to \frac{2}{3}$ are optimal for the polynomial decay step-size rule. Moreover, $\alpha = 1, \beta = \frac{2}{3}$ had almost the same performance as $\alpha = 0.99, \beta = 0.66$. Nevertheless, for the sake of completeness, in the final version of the paper, we will include experiments in which we use $\alpha = 1$ and choose the best β by conducting an extensive grid search.

4. The complexity of scheduling the step-sizes in Algorithm 1 depends on the implementation of linear regression to compute the slope. In practice, instead of using consecutive points to compute the slope, one can use every m^{th} point to reduce the amount of computation and storage. Further, Recursive Least Squares can be potentially used to further reduce the computational complexity.

Response to Reviewer 2:

1 and 2. We thank the reviewer for bringing the two papers (Liu et al., JAIR 2019 and Sutton et al., JMLR 2016) to our attention. Liu et al. shows how GTD-class algorithms can be formally derived using a primal-dual saddle point objective function. They use insights from this derivation to provide finite-time guarantees for GTD-class algorithms with linear function approximation and design a more efficient algorithm called GTD2-MP. However, the authors restrict their analysis and results to single time-scale GTD algorithms (see Algorithms 1 and 2). In fact, in Section 5.4, they discuss how their analysis cannot be extended to TDC because TDC is a two time-scale algorithm (no single time-scale version). In addition to analyzing only single time-scale algorithms, the paper also assumes an additional projection step and i.i.d. sampling of data points from the steady-state distribution.

Sutton et al. presents a (single time-scale) variant of linear TD learning, which they call emphatic TD and show that it is stable under off-policy training unlike standard linear TD. As mentioned in our submission (Section 4, lines 209-226), although we present our analysis for two time-scale algorithms, it is more general and when specialized to the single-time scale, it will recover the finite-time guarantees in Srikant-Ying (COLT, 2019). These finite-time guarantees will apply to both emphatic TD as well as GTD2-MP, since they are both single time-scale, and further improve upon prior work since we do not assume an additional projection step or i.i.d. noise.

3. Maei et al., 2009 (link - https://bit.ly/32P3hrw) proposes generalizations of GTD2 and TDC to nonlinear arbitrary smooth function approximation. They also provide an asymptotic convergence analysis to the set of local optima. We are currently trying to extend our analysis to these generalizations in order to gain insight into using these algorithms with deep neural networks. Once the analysis is complete, we will use it to further refine our adaptive learning rate rule and apply it in experiments using nonlinear deep neural networks.

4. We will add concise and accessible explanations of the main theoretical results in the final version of the paper.

Response to Reviewer 3:

1. If the paper is accepted, we will work further on improving the clarity of the work. Specifically, we will try to make the theoretical results more accessible to readers who do not have strong mathematical familiarity with RL. In particular, we will present more intuition regarding the stability theory for singularly perturbed ODEs and how it relates to the model and proofs in our paper.

2. The main theoretical result in the paper provides an insight into the role of different parameters (ϵ, α, β and τ) in determining the rate of convergence of a two time-scale algorithm. For a practitioner, this result can be of immense utility in designing sample-efficient two time-scale RL algorithms, since the rate of convergence can be better optimized with the knowledge of the impact of different parameters. The adaptive learning rate rule that we have designed is one such illustration of using the result to make it directly useful to an experimentalist. Further practically useful improvements to the convergence rate may be possible using the theory in this paper, and we are currently investigating such ideas.