Reviewer #1: **Q1.** *The projection step is impractical. In addition, if one notices, later in the proof of Lemma 2 (see eq. (27) in the Appendix), another assumption requires the norm of feature vectors to be less than 1. Making these two assumptions simultaneously, does not necessarily guarantee that the true Q-function can be approximated by $\phi^T\theta$.*
Response: As we clarify in our answer to Q2 below, projection is not needed for practically implementing SARSA, and convergence still holds. Even if it is used, to guarantee approximation of Q-function by $\phi^T\theta^*$, as stated in our Theorem 1, we require that the optimal θ^* is within the projection radius. Such a projection radius can be obtained in practice by using our Lemma 1 and designing an online estimator of w_l.

Q2. *Later, it is claimed that without the projection, the method is still useful, and one just needs to set R to possibly a very large value; however, it can be seen from the right-hand side of equations (3), (4), and (11), that the second and third power of R shows up through G, and thus the bounds will possibly be very loose.*
Response: We clarify that projection is NOT necessary to implement SARSA and does not affect its convergence based on existing literature. (Our statement "set R to possibly a very large value" was not accurate.) [Gorden 2001] has shown that SARSA converges to a bounded region. [Melo et. al. 2008] and [Perkins and Precup 2003] further showed that SARSA with a Lipschitz continuous policy improvement operator converges even without projection. Thus θ_t generated by SARSA is already bounded by itself, and projection is not necessary. Thus, the non-asymptotic bound for SARSA without projection is the same as if we use projection with $R = \max_t \|\theta_t\|_2$ (only for the purpose of analysis). Moreover, R is determined by the nature of algorithm. If it happens to be large, then the error should be large by nature, which doesn't mean the bound is loose.

Q3. *Another concern was also raised by [Chen et. al. 2019]. It is claimed in [Chen et. al. 2019] that Thm 1 of [Melo et. al. 2008], that this work builds on, cannot be verified. This concern further weakens the contributions of this work.*
Response: We clarify that this paper studies SARSA, **not** Q-learning. Our proof for SARSA does not have the issue pointed out in [Chen et. al. 2019] for Q-learning. We are aware of the fact that one step in the proof of Thm 1 in [Melo et. al. 2008] cannot be verified for Q-learning, which might be due to greedy policy and off-policy training taken in Q-learning. Such an issue does not affect SARSA as an on-policy algorithm.

Q4. *Assumption 2 requires C small enough to guarantee a negative eigenvalue. No clear characterization of w_s. In practice, how one can guarantee a "feasible" or "meaningful" policy for which it is possible to have such a small C.*
Response: In practice, one can numerically tune the parameter to find a "feasible" policy improvement operator, e.g., β in softmax (which corresponds to a C). Assumption 2 (for guaranteeing convergence) can be empirically checked to provide guidance for parameter tuning.

Q5. *It is also recommended to provide some numerical tests to verify the theoretical results.*
Response: We have run numerical results and will include them in the revision.

Reviewer #2: **Q6.** *Write the proof for Lemma 5 explicitly.* **Response:** Done.

Q7. *Remind that Eq (50) follows from (49) because of the relationship between α_t and w_s.* **Response:** Done.

Q8. *How realistic is assumption that for any θ the induced Markov chain is ergodic (right before Assumption 1)? Is it guaranteed for any special class of environments or representations? Discussion on the plausibility of assumptions.*
Response: First note that a Markov chain is uniformly ergodic if it is irreducible (i.e., possibly get to any state from any state) and aperiodic [Levin & Peres 2017]. Now consider an environment for which there exists a policy that can map any state to any state with nonzero probability (i.e., irreducibility holds) and can get back to the same state aperiodically. (Note that such environments are commonly encoutered in practice.) Then for any θ, as long as the policy improvement operator explores (i.e., with non-zero probability to take any action at any state), the induced Markov chain remains to be irreducible and aperiodic, and is hence ergodic. Therefore, such an assumption is realistic, and is guaranteed for aforementioned environments. Please see response to Q1 and Q4 for other assumptions.

Q9. *Does this result bear implications for framing a RL problem or designing features? Highlight any insights that follow from these results.*
Response: First, our result characterizes sample complexity of SARSA for both constant and diminishing step sizes, which is useful for choosing learning rate to design fast RL algorithms. Second, our result indicates that the faster the underlying Markov process mixes, the faster SARSA converges. This motivates the design of tricks (e.g., experience replay for TD and Q-learning [Wang et. al. arXiv:1809.08926 2017]) to improve the mixing property of Markov process.

Reviewer #3: **Q10.** *The analysis is limited in linear function approximation case.*
Response: One great advantage of linear function approximation is that it is very easy to implement. Even for this case, there is no existing finite sample analysis in the literature, and this paper accomplishes this step (which is already technically nontrivial). It is of great interest to further explore more advanced function spaces, such as deep neural networks. Looking forward, this work serves as a first step towards understanding the more complicated case.