We thank all the reviewers for their comments and constructive feedback.

Response to Reviewer 1: *Theoretical results:* The reviewer is correct in that some assumptions are needed to derive the results. However, we believe that our results are interesting for two reasons: (i) our paper is the first to present a theoretical analysis of Double Q-learning versus Q-learning that goes beyond asymptotic convergence. (ii) the experimental results (including additional ones we have conducted) support our theory. In particular, as noted in the paper, some of our experiments (e.g., Gridworld) are for models which do not satisfy the assumptions needed for the theory.

Experimental results: The suggestion to add experiments for the Sutton-Barto example is very interesting. We have indeed performed those simulations now, see Figures (a) and (b). To be consistent with Sutton and Barto, for this example, we have plotted the probability of going left as a function of the number of episodes. Double Q-learning with twice the step-size and averaging does indeed perform better as the theory suggests. We will include the results for these experiments along with the implementation details in the final version of the paper.

Response to Reviewer 2: *Theoretical results:* Even though the theory does assume a unique optimal policy, our experimental results indicate that our conclusions hold more broadly. More importantly, we would like to point out that our paper is the first to show that Q-learning will perform strictly better than Double Q-learning (see Theorem 4 in the supplementary material which further strengthens the result in Theorem 2) at least under some nontrivial conditions. Our primary goal is to obtain a theoretical understanding of when and why Double Q-learning performs well. We hope that our idea of analyzing the asymptotic covariance will stimulate further research leading to a deeper understanding of Double Q-learning and its variants.

Experiments beyond linear/tabular settings: We have now conducted experiments where we have used a neural network to estimate the Q-function. One such result is shown in Figure (b) for the same example suggested by Reviewer 1 since it has estimation bias. And again Double Q-learning with twice the step size and averaging performs the best, especially in the initial episodes. We will include the results for these experiments along with the implementation details in the final version of the paper. The comments about Adam/other step-size selection techniques and initialization are interesting, we will explore them in the final version (if the paper is accepted) and/or a longer version of the paper.

References: We will add and discuss the suggested references in the paper.

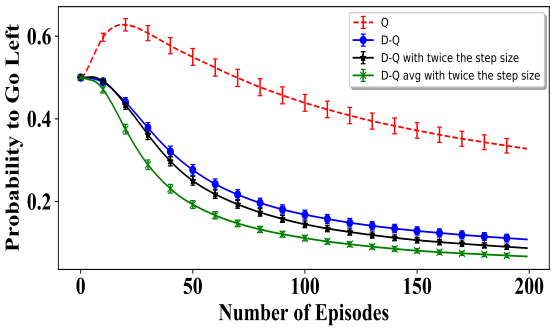
Response to Reviewer 3:

Experimental validation of the theory: As mentioned, Figures (a) and (b) present experimental results on an example with estimation bias, where we have also used a neural network for Q-function approximation (Figure (b)). The experimental results in the paper also include examples with non-unique optimal policies. We will include more experimental results in the paper to show robustness with respect to choice of hyper-parameters and also conduct additional experiments in complex RL environments. In the cart pole experiment, we have used the number of episodes to reach a certain reward as our metric and we have the distribution of this quantity. This appears to us to be stronger than knowing just the errors of the means.

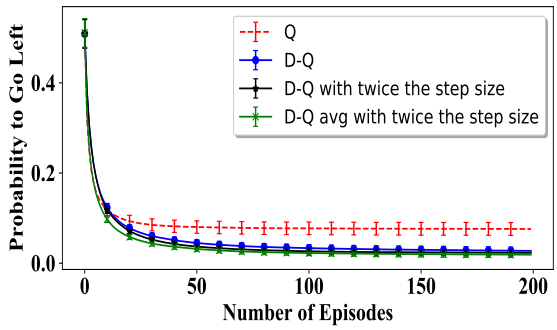
Clarity: Thanks for the suggestion to include more explanations, we will do so in the final/longer version of the paper.

Response to Reviewer 4:

Experiments: Thanks for suggestion about not using code in multiple languages. We will migrate them over to Python before the conference if the paper is accepted. We have now conducted experiments for cases where a neural network is used to approximate the Q-function. Studying such a model theoretically is currently a significant challenge in the field.



(a) Sutton-Barto Tabular



(b) Sutton-Barto Neural Network