Review 1. Thanks for your comment. In fact, average reward RL is in many applications more relevant than discounted reward and is typically more challenging to study. See the detailed reasons below.

- Application wise, in many real-world networked systems, there is no naturally defined starting state or discounting factor, and the performance is measured under the stationary distribution. For example, in the wireless communication, the long standing performance metric is the throughput, which measures the average number of packets sent under the stationary distribution.
- In RL, it is widely known that studying the average reward is a more challenging topic. For example, the Bellman operator is no longer a contraction in general (because there is no discounting factor), and the set of fixed points is no longer unique, but a subspace instead. From a historic context, it was long known the average reward required very different algorithms (e.g. the Q function becomes the "differential" Q-function) and analysis techniques, and there had been a series of work discussing the average reward case, e.g. [Tsitsiklis and Van Roy 1999] cited in the paper, and "John Tsitsiklis and Benjamin Van Roy. On average versus discounted reward temporal-difference learning. Machine Learning 49.2-3 (2002): 179-191".
- On top of the above, when coupled with the multi-agent setting, the average reward case brings additional challenges. Specifically, as shown in Appendix A.2 in the paper, our average reward problem captures certain NP-hard instances. Similar complexity results can be found in [Blondel and Tsitsiklis 2000].

Review 2. Thanks for your comment and please find the response to your questions below.

- **(Q-function)** In the definition for the Q-function, the cost $J(\theta)$ is subtracted. This is the standard definition for Q-function for the average reward case, and is sometimes called differential Q-function. The relevant reference can be found in standard textbooks like [Bertsekas 2007] and we will add that to the final version.
- **(Communication)** Yes, access to neighbor's information is needed and we will clarify that in the final paper.
- **(Comparison)** We are happy to add comparison to other learning algorithms. In fact, Appendix E in our supplementary material includes a run of our algorithm with $\kappa = 0$, which is essentially the independent learner approach in the literature. If the reviewer has suggestions on specific algorithms to compare, we are happy to test those.

Review 3. Thanks for your comment and please find the response to your questions below.

- **(Thm 2 vs [Qu et al 2019])** Average reward RL is in general more challenging than the discounted case, see our response to Review 1. Regarding Thm 2 specifically, it actually relies on very different techniques than [Qu et al 2019]. In [Qu et al 2019], at each time, there is a (long) inner loop of critic (TD-learning) steps that estimates the Q-function to a good accuracy during which the policy is *fixed*. This makes the analysis of the critic "decoupled" from the policy updates (actor). In the current paper, there is no inner loop, and critic (TD-learning) and actor (policy update) steps are performed *simultaneously*. This creates many challenges in the analysis, as the critic no longer operates under a fixed policy, and the analysis of it cannot be decoupled from the actor.
- **(Relationship between Thm 1 and Thm 2.)** The result of Thm 1 is stated as Assumption 2 for Theorem 2, and as such Thm 2 will surely rely on Thm 1. In terms of proof, the result of Thm 1 is used in Lemma 10, which shows the fixed point of the critic is a good approximation of the full Q-function. We make such a separation because (a) Theorem 1 can be useful for many types of RL methods, of which Algo 1/Thm 2 is just one particular actor-critic method; (b) The condition in Thm 1 might be conservative and the exponential decay will hold broadly beyond the condition in Thm 1 (cf. Appendix B.4). So writing the exponential decay as an assumption will broaden the applicability of Thm 2.
- **(Other)** Thanks for the suggestion on neural network implementation and we are happy to try that. We will also incorporate your other suggestions to improve clarity.

Review 4. Thanks for your comment and please find our response below.

- **(Comparison with mean-field)** From a theoretic view point, our model and framework is very different from the mean-field approach. While both approaches attempt to address the curse of dimensionality, our framework allows heterogeneous agents, while mean-field approaches typically assume homogeneous agents, whose affect can be well approximated by their mean. Nevertheless, we agree that mean-field approaches could be quite good for grid settings. We are happy to add the above discussions in the revised paper. Also, we are happy to compare our approach with mean-field approach in large scale networks, for both homogeneous agents and heterogeneous agents.
- **(Simulations)** In our supplementary material (Appendix E), we also ran our algorithm with $\kappa = 0$ (each agent runs a single agent actor critic method), which is similar in spirit to the independent learner approach that the reviewer suggested for comparison. That being said, we are happy to incorporate more simulations, including larger number of agents (500*500), different graphs (less "regular" than grid), tasks like Battle games, and comparison with other algorithms (mean field).