We appreciate the valuable comments from the reviewers. We will revise accordingly.

Reviewer #1. (**Novelty in techniques.**) Our proof of the convergence of PPO/TRPO has several building blocks, which, to the best of our knowledge, is not covered by existing work.

- **TD and SGD:** Our unified analysis covers both TD and SGD, where **TD is not covered by any existing supervised learning analysis**. In particular, the semigradient used in the TD update is **not an unbiased stochastic gradient of any supervised learning objective**. To see this, one may take the derivative of the semigradient, which gives an asymmetric matrix. In contrast, in supervised learning, such a matrix is the Hessian, which must be symmetric.

A "straightforward adaptation" of existing supervised learning analysis does not yield the global convergence of TD. In fact, most algorithms in supervised learning are known to at least converge (although they may converge to undesired stationary points). In contrast, TD with nonlinear function approximator **is known to generally diverge** [1], which eludes existing supervised learning analysis. Our unified analysis of both TD and SGD shows how overparametrization allows bypassing the divergence of TD, which has not been observed in the context of supervised learning before.

- **Nonconvex mirror descent:** Most existing analysis of mirror descent's convergence to a **global optimum** builds on the critical assumption that the objective is convex, which is not the case for the objective $J(\pi)$ in RL. Our **global convergence** analysis of PPO/TRPO only builds on **one-point monotonicity** (Lemma 5.1), instead of convexity. In fact, we are not aware of any existing analysis of **infinite-dimensional** mirror descent under one-point monotonicity.

- **Error propagation:** RL is divided into **policy-based** and **value-based** approaches. **PPO/TRPO falls into the former**, while **Q-learning falls into the latter**. More specifically, PPO/TRPO explicitly tracks a policy, while Q-learning does not. In particular, the Q-function tracked in Q-learning is not the action-value function of any policy. As a result, the error propagation in Q-learning relies on the **γ-contraction of the Bellman optimality operator**, which operates on the **Q-function**. In contrast, the error propagation in PPO/TRPO in our analysis relies on the **convergence of mirror descent**, which updates the **policy**. Such two mechanisms are not comparable.

(**Expectation over initialization.**) Yes, the expectation is also taken over the initialization of the neural networks. Our notation indeed lacks clarification. Thanks for pointing this out. We will revise accordingly.

(**Sufficiently large function class.**) Yes, the divergence is largely due to the limited representation power of **finite-dimensional** linear function approximators. In contrast, overparametrized neural networks provide sufficient representation power as **infinite-dimensional** nonlinear function approximators, which enables the global convergence.

Reviewer #2. (**Shallow architecture.**) Thanks for pointing this out. The name "two-layer neural network" is standard in recent work (see, e.g., [2]), counting the output layer as one of the layers. We will emphasize more on the shallow architecture considered here.

(**Bounded reward.**) The state-value function and action-value function defined in Section 2 are normalized by a factor of $1-\gamma$. Hence, they have the same upper bound $R_{\max}$ as the reward function.

(**Simulation.**) Thanks for the suggestion. We will add illustrative simulation examples to the appendix.

Reviewer #3. (**Conclusion section.**) We will add a conclusion section to briefly summarize this paper as well as potential future work.

(**Nonconvexity, infinite-dimensionality, and technical challenges.**) The nonconvexity arises from two aspects: (i) The neural network parametrization of the energy-based policy and the action-value function makes the subproblems of policy improvement and policy evaluation nonconvex. (ii) The RL objective $J(\pi)$ is also nonconvex in π. Meanwhile, the continuous state space introduces infinite-dimensionality. The nonconvexity and infinite-dimensionality combined makes the global convergence analysis challenging.

(**Algorithm 1.**) Yes, in Algorithm 1, lines 5 and 6 are just one projected TD step. Correspondingly, lines 5 and 6 in Algorithm 2 are one projected SGD step.

(**Overparametrization.**) Our analysis does require the width of the neural network to be large enough. In practice, in order to ensure the desired representation power of the neural network, a larger number of parameters is required.

[1] TSITSIKLIS, J. N. and VAN ROY, B. (1997). Analysis of temporal-diffference learning with function approximation. In Advances in Neural Information Processing Systems.

[2] MUNOS, R. and SZEPESVARI, C. (2008). Finite-time bounds for fitted value iteration. Journal of Machine Learning Research.

[3] ARORA, S., DU, S. S., HU, W., LI, Z. and WANG, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv:1901.08584.