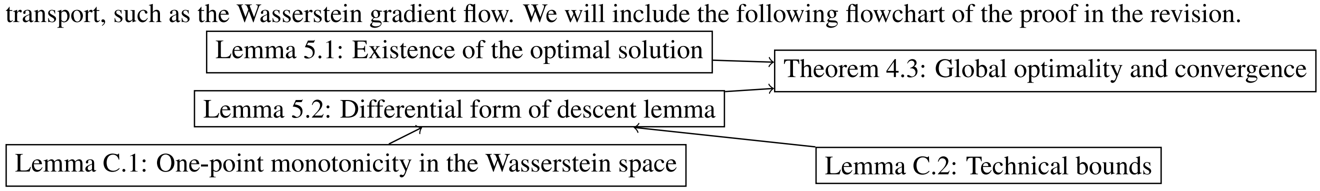
We appreciate the valuable comments from the reviewers. We will revise accordingly.

Reviewer #2. (Motivation of mean-field regime.) As discussed in Lines 27-31, 47-50, and 75-76, the empirical success of deep RL is empowered by its ability to learn data-dependent feature representation. However, the NTK-based analysis of TD [21] requires an implicit local linearization with respect to the initial feature representation, which is not data-dependent, and thus, fails to explain how the feature representation evolves. In contrast, the mean-field regime allows the feature representation evolving. Specifically, for the induced kernel $\mathbb{K}(\cdot, \cdot; \rho_t)$ defined in (3.7), our mean-field regime allows ρ_t to go beyond ρ_0, while the NTK regime requires $\rho_t = \rho_0$ (Lines 214-227). (Comparison to the NTK-based analysis.) The NTK-based analysis requires a proper scaling of the neural network to allow the implicit local linearization (Lines 43-45), while our analysis does not require linearization. Moreover, the analysis in [21] is based on the one-point monotonicity in the Euclidean space, while we generalize such a notion to the Wasserstein space (Lines 63-66). (Representation learning.) We discuss the representation learning in Lines 27-31, 47-50, 75-76, 176-184, and 214-227. We study the evolution of the induced kernel $\mathbb{K}(\cdot, \cdot; \rho_t)$ defined in (3.7), which is fully characterized by ρ_t. We show the global convergence of ρ_t in Theorem 4.3, which implies that the induced kernel also converges to the globally optimal one. (Discretization.) We study the discretization of the trajectory of PDE in Proposition 3.1 and Appendix D, based on which we establish a discrete-time convergence rate in Corollary 4.4 by aggregating the the discretization error. (Missing reference.) We will cite the paper in our revision. Thank you for pointing out.

Reviewer #3. (Assumptions for Q-learning.) As discussed in Lines 477-478, similar assumptions are employed in the analysis of Q-learning in simpler settings (linear or NTK). On the other hand, we do understand that Assumptions B.1 and B.3 are strong by themselves. Thus, we put Q-learning in the appendix as an extension of our main results for TD. In the revision, we will not claim the convergence of Q-learning as our contribution and emphasize the restrictiveness of such assumptions. (Target network.) When a target network is employed, TD becomes a bilevel optimization problem, in which case the convergence can be proved by similar technical tools. (Wasserstein 2 distance.) Similar to the ℓ_2 distance in $\mathbb{R}^d$, the Wasserstein 2 distance induces an "inner product" (more precisely, a weak Riemannian metric) on the space of probability measures and is well studied in the literature of optimal transport, which forms the basis of our analysis. It may be possible to generalize our results to the Wasserstein p distance by exploiting the duality of the p and q norms, where $1/p + 1/q = 1$. (Assumption 4.1.) As discussed in Lines 189-191, similar assumptions are commonly used in the mean-field analysis of neural networks and can be ensured through normalizing the state-action space. Moreover, our analysis can be straightforwardly generalized to the setting where $\|x\| \leq C$ for an absolute constant C. Such a setting covers a majority of RL problems, but yes, we do agree that Assumption 4.1 doesn't always hold, especially when the state or action space is unbounded. (MSPBE and universal function approximation.) As discussed in Lines 198-199, our function class defined in (4.3) captures a rich class of functions because of the universal approximation theorem (UAT). It is worth noting that UAT requires additional conditions on the target function, e.g., an upper bounded first moment of the Fourier coefficients [11]. As UAT doesn't ensure the approximation of *any* target function, we use MSPBE rather than MSBE. (The example in Tsitsiklis and Van Roy.) Yes, we square the counterexamples of Tsitsiklis and Van Roy (1997) and Baird (1995) via overparameterization. The divergence in their examples comes from the nonconvexity and the bias of the semigradient. In contrast, we show in Lemma C.1 that, coupled with an infinitely wide neural network, TD becomes a weakly convex problem (in the sense of one-point monotonicity) with respect to the distribution of the parameter in the Wasserstein space. We will add the numerical example in the revision.

Reviewer #4. (A flowchart of the proof.) The proof is technical and requires certain preliminary knowledge on optimal transport, such as the Wasserstein gradient flow. We will include the following flowchart of the proof in the revision.



(The definition of div.) The operator div is the divergence operator from vector calculus. We will specify the meaning of div in our revision. Thank you for pointing out. (Activation function in the second layer.) As discussed in Lines 194-203, we apply the activation function only to the first layer of our neural network, which is commonly used in the mean-field analysis of neural networks. With an activation function applied to the second layer, our analysis still carries over but becomes more involved to present. (Relation of d and D.) Yes, the dimensions D and d are closely related, which are used to cover the common cases in the study of neural networks. (Relation of (3.3) and (3.4).) As discussed in Lines 163-175, (3.4) can be viewed as the continuous-time and infinite-width limit of (3.3), while (3.3) can be viewed as the discretization of (3.4). In Proposition 3.1, we show that (3.3) approximates (3.4) in the limit, whose detailed proof is included in Appendix D. (Dirac delta.) Yes, the notation δ_{θ_i} in Line 150 is the Dirac delta. (Definition of $\bar{\rho}$.) We will define $\bar{\rho}$ explicitly with a standalone line in the revision. Thank you for pointing out.