We thank all three reviewers for their careful reading and constructive suggestions. We will revise the paper thoroughly, incorporating all the comments.

[reviewer 1] We will provide precise references from the classical literature on the hardness of N-player games, including [PR05]. We will add the definition of the ℓ_1-Wasserstein distance to make the paper self-contained. In addition, we will correct s^t (of line 67) to s_t, and rewrite the formula $\mu_t(s)$ as $\mu_t(s) = \frac{\sum_{j=1, j\neq i}^{N} 1_s(s_t^j)}{N}$, where the indicator function $1_s(s_t^j) = 1$ if $s_t^j = s$ and 0 otherwise. Finally, we will add a section to clearly define all notations.

[reviewer 2] (1) We will revise the presentation carefully, as suggested. (2) Thank you for asking the clarification between the stationary versus non-stationary MFGs. Stationary solutions are commonly adopted for MFGs with an infinite-time horizon, see [12,15]. Non-stationary solutions are mostly used for MFGs with a finite-time horizon, see [Bar12]. Our work shows the existence and the uniqueness theorems for both the stationary (Appendix B) and the non-stationary MFGs (Theorem 1). The algorithms in Section 4 are focused on Q-learning algorithms, which are primarily designed for stationary MDPs and hence appropriate for stationary MFGs.

(3) On the contribution: The GMFG framework (Section 3) incorporated *both* the state distribution and the action distribution. With the additional action distribution, Γ_2 was different from the one defined in [12] and the proof for the uniqueness and the existence of the solution needed further modifications. To clarify the difference with [25]: [25] showed the convergence to a *local* (Nash equilibrium) solution, and the uniqueness of the local solution given the presence of a unique *global* solution. However, [25] did not analyze the existence of a unique *global* solution. We established the existence, the uniqueness, and the convergence to a *global* solution. We will add this discussion in the revision.

(4) On the related works of MFG: Apologies for missing some references, which we will add with careful discussions of their contributions and relationship to our work ([HM17, MJdC18]).

(5) On the literature related to soft Q-learning: We will include additional references. Thank you for pointing out the potential divergence of SARSA using the Boltzmann operator [AL17]. Indeed, it is now more interesting to see the guaranteed convergence with Q-learning using the Boltzmann policies. We will add this comparison and discussion in the revision. Indeed, we think it is worth testing the performance using the Mellowmax exploration, in addition to the Boltzmann exploration.

(6) On the definition of NE: MFG is a game with an infinite number of identical players. The NE solution is therefore the same for each individual by symmetry. If each individual in the population follows the conditional optimal solution (from the single player side), the consistency means that no player in the population has the incentive to deviate (from the solution of the single player side). This is consistent with the NE definition for N-player games.

(7) For the Ad auction example, apologies for the confusion. M is only one of several model parameters and indicates the competition intensity. The game interaction is more extensive than M alone: for each agent, all of her reward, her winning probability and her budget dynamics, depend on the strategies from other opponents.

[reviewer 3] We will rewrite the repeated auction example in Section 2.3, in order to be consistent with the general model setting in Section 2.2. We will also clearly define and explain the quantities in Theorem 2.

(3) For the stationary setting in Section 4, the corresponding uniqueness and existence theorems for the time-independent MFG solution (i.e., Theorem 4) are given in Appendix B (see Line 174) under slightly different conditions from the non-stationary setting. Note that due to the introduction of the mean information process in the MFG, an infinite-time horizon MFG is generally associated with a parabolic type PDE, hence the Nash equilibrium could still be time dependent. This is fundamentally different from the theory of single-agent MDP where the optimal control, if exists uniquely, would be time independent in an infinite-time horizon setting.

(4) For Assumption 1 and inequality (5), we can impose Γ_1 to be single-valued by using *e.g.*, **argmax-e**. Moreover, in the linear-quadratic continuous state-action setting, the assumption can be translated into constraints on model parameters. We will add this in the revision.

(5) In practice, a uniform grid for the epsilon-net would suffice, as shown in our experiments. That is to replace the projection of $\tilde{\mathcal{L}}_k$ onto the epsilon-net by truncating the resulting $\tilde{\mathcal{L}}_k$, up to a certain number of digits. For example, 4 was used in the experiment. The choice of c is fairly simple, as the experiments are robust with respect to different values of c, ranging from 1 to 100. For instance, we chose $c = 5$.

(6) For the iteration complexity in Theorem 2, indeed, one could simplify the order of iteration complexity, by simply taking h to be 3/4 and η to be 1. We will clarify this.

(7) Thank you for the reference [SBR18] for discrete-time MFGs, which will be added accordingly.

References

[AL17] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. 2017.

[Bar12] M. Bardi. Explicit solutions of some linear-quadratic mean field games. 2012.

[HM17] M. Huang and Y. Ma. Mean field stochastic games with binary actions: Stationary threshold policies. 2017.

[MJdC18] D. Mguni, J. Jennings, and E. M. de Cote. Decentralised learning in systems with many, many strategic agents. 2018.

[PR05] C. H. Papadimitriou and T. Roughgarden. Computing equilibria in multi-player games. 2005.

[SBR18] N. Saldi, T. Basar, and M. Raginsky. Markov-Nash equilibria in mean-field games with discounted cost. 2018.