We thank the reviewers for their time and comments. We would like to first reiterate our message, which is expertly summarized by Reviewer 4, whom we would like to thank: “This paper proposes an improved robust RL algorithm that finds mixed NE via LD. The paper first constructs a toy example where analytically non-sampling based methods fail to find the NE if initialized poorly and the proposed algorithm consistently finds the NE. Then, it shows that even adaptive non-sampling method cannot find the NE. Finally, the paper presents empirical results where the proposed method outperforms prior baseline for robust RL consistently." We also strongly believe (again to quote Reviewer 4) “the contribution of the paper is important for robust RL" and that we contend that paper has a strong impact “... since robust decision making is essential for high-stake automated decision making." Below, we address the misunderstandings as well as the concerns raised by the reviewers, tagging relevant reviewers with R#.

On the theoretical front:

1. [R1] We systematically study the **algorithmic** advantages of MixedNE-LD, whereas [12] focuses on interpreting mixed NE as a better solution concept; see our discussion in Remark 1.
2. [R1 + R3] We contend that our new non-convex-concave discussion is important and adds theoretical value beyond [12] via a simple counter-example that others can use in their work: In particular, Theorems 1 and 2 shows that state-of-the-arts fail on extremely simple objectives (cf. lines 151-155) so we cannot expect them to perform well on complicated real objectives such as the problem of training robust RL policies.
3. [R3] We clarify that our formulation follows [13, eq (4)], which adds Gaussian noise regardless of the convexity.

On the empirical front:

1. [R1 + R4] Our GAD baseline already covers RARL; in the RARL work they used TRPO, we have used both DDPG and TD3, which are both known to be superior to TRPO.
2. [R4] The reviewer might have missed that we have already provided results for friction changes in the appendix.
3. [R1 + R2] Our theory suggests that, when there are many suboptimal stationary points, MixedNE-LD outperforms existing methods (GAD/EG). In the absence of this property, we expect MixedNE-LD to perform slightly worse since we can simply focus on converging to stationary points without adding the explorative noise in MixedNE-LD. As a side note, we also provide the first numerical results with EG in the RL setting.

   In the experiments, MixedNE-LD outperformed GAD/EG quite consistently. The situations when it outperforms the baselines depend on: (i) nature of the environment (number of suboptimal stationary points), and (ii) the design choices for the practical relaxation of MixedNE-LD such as inner loop iterations count.
4. [R3] In our work, we only consider the deterministic reward setting. Thus the comparison in the experiments is fair.
5. [R1] The reviewer might have missed that we have already extended MixedNE-LD for TD3 and vanilla policy gradient in the appendix.
6. [R1] The shadowed area represents 1 SD from mean cumulative reward.
7. [R1] We have run the experiments for 5 seeds; and performed 100 testing simulations per seed. Our codes are shared so running multiple training simulations per seed would be possible.
8. [R1] We have explicitly stated this at the beginning of Section 5.1: In the experiments, we consider the NR-MDP setting [4]. It can cover only the changes in the transition dynamics that can be simulated via the changes in the action. Nevertheless, MixedNE-LD from [6] applies to general two-player Markov Games as well. We have already spent significant computational effort, and we hope that our results will inspire others to follow up since it is not possible to be thoroughly exhaustive (cf. our changes report).
9. [R2] The reviewer might have missed that we already have a discussion with experimental evidence on improving the computation time of our algorithm in Appendix B. The ONE-PLAYER DDPG with SGLD is a significant computational relaxation of DDPG with MixedNE-LD without compromising the empirical performance.

On the presentation:

1. [R1] Even though we do not have a separate related work section, in the introduction, we have clearly positioned our work in the literature. We are open to suggestions from the reviewer if a different style is more appropriate.
2. [R1 + R2] We will consider including more discussion on Algorithm 1 from [12] and related work.
3. [R2] We will add a table of average improvement over baselines across domains to clearly demonstrate the strong margin of improvement.
4. [R2] We started the TD3 experiments only after the ICML decision as requested in the post rebuttal comments, thus we were not able to cover all the MuJoCo environments considered in the baseline papers, within time. Now, we do have the results for all the environments (observed qualitatively similar improvements over the baselines); if necessary we can present TD3 results in the main paper, still we prefer to keep DDPG results to minimize the changes and match with prior work. Please confer our changes report.