Response to all reviewers:

We thank the reviewers for carefully reviewing our paper and providing constructive feedback. To the best of our knowledge, we are the first paper formulating policy poisoning in batch RL and control, and there are fruitful new directions to explore, such as extending our attack to more complex RL learners. We believe our paper will trigger a line of new research on data poisoning attack on RL/Control.

Response to reviewer 1: We thank reviewer 1 for positive comments.

We will move some of the analysis and proofs back to the main text for better readability.

Response to reviewer 3: We thank the reviewer for valuable comments.

(1) Is the attack trivial? Our study has an emphasis on finding the *optimal* way of changing the rewards to achieve a target policy. The optimality guarantee is not trivial - one needs to solve a (bi-level) optimization to achieve it. Furthermore, we provide theoretical analysis to bound the optimal change of rewards measured by ℓ_p norm for arbitrary p, which is important to understanding the minimal effort an attacker has to spend to achieve successful attack.

(2) Definition of poisoning ratio. We apologize for the confusing definition in the paper. In our experiment, we actually computed poisoning ratio as $\|\mathbf{r} - \mathbf{r}^0\|_2/\|\mathbf{r}_0\|_2$ instead of $\|\mathbf{r} - \mathbf{r}^0\|_2/\|\mathbf{r}\|_2$, so this is just a typo of writing. We agree with the reviewer that if all rewards are shifted by a constant, the policy does not change, but the poisoning ratio does. However, the poisoning ratio is a metric that tries to capture the notion of *attack cost*, which not only depends on the policy change, but also ties to the magnitude of the clean rewards. Conceptually, the same "absolute" change on larger clean rewards should mean smaller attack cost, thus should have smaller poisoning ratio. This is exactly the case if we divide by $\|r_0\|$. Besides, from the optimization perspective, our attack optimizes $\|\mathbf{r} - \mathbf{r}_0\|$, which is equivalent to optimizing $\|\mathbf{r} - \mathbf{r}_0\|/\|\mathbf{r}_0\|$ or $\|\mathbf{r} - \mathbf{r}_0\|/\|\mathbf{r}_0 - \text{mean}(\mathbf{r}_0)\|$. But we do think the reviewer's suggestion is also reasonable, so we computed the poisoning ratio using the suggested metric $\|\mathbf{r} - \mathbf{r}_0\|/\|\mathbf{r}_0 - \text{mean}(\mathbf{r}_0)\|$, and the results for experiment 2,3 and 4 are 14.66%, 8.64%, and 0.77% respectively, which is just slightly higher than those reported in the paper. We will include the metric suggested by the reviewer in the revised version.

Regarding large attack on a single transition, it is true that small poisoning ratio does not fully capture the magnitude of change on each individual reward, if the metric is ℓ_2 norm (as we did in our experiments). However, as our formulation uses general ℓ_p norm, one can consider using ℓ_∞ norm instead, which will avoid the example the reviewer mentioned.

Response to reviewer 4: We thank the reviewer for constructive comments.

(1) Bi-level optimization formulation. We follow the notation in state-of-the-art literatures on data poisoning attacks (e.g, [1]). However, we can map the notation to Bard et al. 2000 paper as follows.

The leader optimization is (25), where the decision variables is the poisoned rewards $\mathbf{r}$. We have only one follower problem (30), together with LQR solution constraint (26) to (29). The decision variables for the follower problem is on the LHS of (30). The (19) is not a follower problem because the estimates $\hat{A}$ and $\hat{B}$ are attack-independent, thus can be computed beforehand based on the clean data. The estimates $\hat{A}$ and $\hat{B}$ then appear as parameters of our main optimization in (26) to (29), which are constraints to ensure that the LQR solution is equal to the target policy (our attack goal). Constraint (31) is simple. We point out that when converting the follower problem (30) into its equivalent KKT condition, one should incorporate its decision variables into the leader-level optimization since there will be no follower problem anymore, which is why we write decision variables for both leader and follower problem on the LHS of (25). We will make it more clear in the revised version.

(2) scalability issue. We agree that the scalability issue is a weakness of our paper. However, since we are aiming for *global optimality* in our attack, we require the RL learner to take the form of some convex optimization (e.g., LP in discrete MDP case), which can be replaced with its KKT equivalent. For more complex learners such as deep RL learners, as they rely on highly non-convex neural networks, the KKT condition approach may not apply. In that case, solving the global optimality using convex optimization might not be feasible, but approximate methods such as gradient descent can be used to find locally optimal solutions. We agree scalability is important, and we will discuss it in the revised paper. We leave that as future work, as our main goals of the paper is to first formally define the problem of adversarial poisoning in RL/control and study if *globally optimal attack* is achievable.

We actually also experimented on a continuous domain: Linear Quadratic Regulator, which is a classic continuous control problem and arguably represents many real world control problems.

References

[1] Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.