We thank the reviewers for the detailed feedback. Before addressing the comments, we acknowledge the typographic / clarification errors, the scope for improvement in the constants, and in the subsequent version of the paper will include a comparison with existing work (such as [SVBB19]) in Table 1. Below we address each reviewer's comments.

R1: We analyze the standard behavior cloning (BC) approach to give a worst-case rate at which its error goes to 0 as a function of N (size of expert dataset), as $\lesssim |\mathcal{S}|H^2/N$. Although this rate largely follows from existing work, we only state this as an achievability counterpart for our more important contribution here: to establish a universal lower bound of $\gtrsim |\mathcal{S}|H^2/N$. We show that for *any algorithm* (even if it can actively query the expert) there exists an instance on which large error ($\gtrsim |\mathcal{S}|H^2/N$) must be incurred. This compounding error lower bound does not have anything in particular to do with BC, and uniformly applies for any learner algorithm. In contrast, the lower bound example of [RB10] (**mentioned by R4**) applies only for supervised learning. They construct a particular MDP and show that a particular learner strategy which plays an action different than the expert with probability ϵ has error $\gtrsim H^2\epsilon$. It turns out, in their instance, the error incurred by BC is *exactly* 0 *given just a single expert trajectory*. Thus it does not imply a uniform lower bound on the error of all learner algorithms as a function of N: even BC performs well on their example.

Comparison with FAIL *[SVBB19]:* In the det. expert setting, without additional assumptions, the worst case error guarantee of BC is superior to FAIL. In [Theorem 3.3, SVBB19] choosing Π to be the set of all deterministic policies (of size $|\mathcal{A}|^{|\mathcal{S}|}$) shows that FAIL achieves error $\sqrt{|\mathcal{S}||\mathcal{A}|H^5/N}$ (ignoring log-factors). In contrast, we show that behavior cloning incurs error $|\mathcal{S}|H^2/N$ [1]. This is always better: not only is it independent of $|\mathcal{A}|$, but has optimal dependence on H and N. However, we clarify that FAIL also applies in the ILFO setting where the expert actions are not observed. Furthermore, if the expert is non-deterministic, we show that MIMIC-EMP has expected error $|\mathcal{S}|H^2/N$ (ignoring log-factors). This again significantly improves on FAIL and surprisingly is independent of $|\mathcal{A}|$. We emphasize that the proof of this result is quite involved and uses a particular coupling based argument - as discussed in the paper, applying simple reduction based analyses is loose, failing to avoid dependence on $|\mathcal{A}|$ and converging slowly at a $1/\sqrt{N}$ rate.

Although some of our results apply only in the det. expert setting, we appeal that in single agent RL every optimal policy is deterministic. Thus, studying the case where the expert policy is deterministic is not too restrictive as it includes the best possible expert policy. Another critique is that formulating IL as a minimax problem and studying worst-case guarantees might be too pessimistic. We argue that studying the minimax approach is a basic formulation for such statistical problems, and provides a benchmark for further improvements.

R2: in this context, by studying in the minimax framework, a conclusion of our work is that additional assumptions on the MDP / reward structure are necessary for an active query algorithm such as DAgger to outperform BC, and explain its superior empirical performance. We believe that the assumptions imposed are critical: weak assumptions may not be able to separate the sample complexity in the active and no-interaction settings, while strong assumptions may compromise the practical relevance of obtained results. Also, to clarify, the BC error upper bound in Theorem 1 translates to the active setting by a modification to the algorithm: by playing the expert's action at visited states when interacting with the MDP, the learner can generate N expert trajectories; then the learner performs BC on this dataset.

Next, we give an intuition for MIMIC-MD in the known transition setting: MIMIC-MD copies the expert action on states seen in the dataset, so the learner incurs error only upon visiting new states (i.e. not seen in the expert dataset). So the appropriate quantity to match is the probabilities induced over states and actions by the expert, when at some point in the episode a new state is visited. Data splitting can in fact be used to estimate these probabilities which immediately leads to the form of MIMIC-MD. That said, it is unclear how to efficiently carry out simulations to identify whether MIMIC-MD admits better error guarantees (**as suggested by R4**). The error incurred by MIMIC-MD (or any policy for that matter) requires evaluating against its corresponding worst MDP instance, one among exponentially many instances. We believe this is an important open problem to resolve in the context of the known-transition setting.

The basis for conjecturing that existing distribution matching approaches are unlikely to be optimal is because they do not take into account that the expert's actions are known at all states seen in the dataset. These policies may choose to play a different action at a state, even if the expert's action is observed in the dataset. In contrast, MIMIC-MD returns a policy that is constrained to mimic the expert at states visited in the expert dataset, and avoids this problem.

R4: MIMIC-MD is not a polynomial-time algorithm as the optimization in Eq. (8) is over multivariate degree-H polynomials. An important future work is to translate this intuition to a polynomial-time algorithm - perhaps one which returns a policy which approximately solves Eq. (8). We also add that the lower bound instances in the known-transition setting showing that the error of any policy is $\gtrsim |\mathcal{S}|H/N$ are indeed homogeneous MDPs. In fact, the time-variance of the reward function can also be lifted if the action space is large enough (say $\gtrsim H$). On the other hand, in the no-interaction setting our lower bounds are inhomogeneous. Here, it is an interesting question to design an algorithm which leverages time-invariant transitions to improve on the $\lesssim |\mathcal{S}|H^2/N$ error rate of BC.

[1] We plan to include an improvement to Theorem 1: we improve the $1/\sqrt{N}$ dependence in the high probability term in Eq. (4) to $\sqrt{|\mathcal{S}|}/N$. Thus BC enjoys a $1/N$ worst-case error rate, both in expectation as well as with high probability.