Scope: We thank all reviewers for their useful comments. Our goal was to theoretically investigate how side observations, formalized by feedback graphs and available in several RL applications, can be used to learn faster in MDPs. For a first study of this problem, a tabular assumption is a natural starting point that already posed several new challenges compared to the bandit setting, which we addressed in a comprehensive way. While direct applications of any tabular approach are limited, we are excited about extensions to non-tabular settings that would capture many real-world applications and believe that the insights in our paper provide the basis for such extensions. A careful empirical evaluation would also be a nice complement indeed, but our theoretical study has already led to a rich and dense content. Thus, we have chosen to devote a separate study to the applications of our theory and experiments in the future.

Reviewer 1:

- **Tabular MDP, accurate side info:** We cover biased (inaccurate) side observations in the appendix. Regarding the tabular assumption, please see comments about the scope above. We will clarify that Lines 45-54 are motivation for feedback graphs in MDPs generally and not necessarily for the tabular case. We believe that our assumptions and claims are carefully stated but we will be happy to rectify any part of the paper that the reviewer views as "careless".
- **More intuition around feedback graph:** There appears to be a misunderstanding about the role of a feedback graph. An edge in the feedback graph does *not* indicate a valid transition. An edge between (s, a) and $(\bar{s}, \bar{a})$ indicates that when the agent is in state s and takes action a it also observes some transition observation $(\bar{s}, \bar{a}, \bar{r}, \bar{s}')$ from $(\bar{s}, \bar{a})$. However, the edge does not indicate anything about the successor states from (s, a) or $(\bar{s}, \bar{a})$. We will provide more intuition about the graph properties and move examples from Appendix A to the main body (see item 2 of R3).
- **Correctness of implicit self loops:** The analysis is correct and consistent with our definitions. We defined feedback graphs to not contain any self-loops and to only stipulate what observations are available *in addition* to the transition performed by the agent ($\mathcal{O}_h(G)$ on pg 3). This is in line with prior work in bandits [24]. If we wanted to make self-loops explicit, then the definitions of the graph properties would need to be changed to ignore such self loops.

Reviewer 2:

- **Replay buffer and off-policy RL:** These are interesting directions and we believe that future work with regret bounds for model-free algorithms is a good next step (as replay buffers and explicit off-policy RL matters most here).
- **Challenges for model-free RL:** Model-free methods use the Q-value of the successor state to update the current state. In the feedback graph setting, one might use side observations to update a state s many times using Q-values of a successor state that has never been observed. Then the Q-value estimate of s is still bad, even though the number of observations is large. Without feedback graphs, where all observations come in full trajectories, this cannot happen.
- **Multi-task settings:** It is correct that our assumptions only hold in certain multi-task settings. The first phase of Alg 3 is one example. Other examples are multiple-destination shortest path problems where the agent needs to learn how to reach different goal states in the same environment. Each goal corresponds to one reward function that can often be assumed to be known.

Reviewer 3:

- **Example applications:** Besides *recommender systems* and *image augmentation* discussed in the introduction, we expect that such structured side observations to be available in *certain robotics applications* where partial knowledge of the environment is often available. Other examples include problems in *personalized tutoring systems*, *autonomous driving* and *personalized medicine*. We will include a list of tasks with more details in the paper.
- **Feedback graph examples (App A):** Thank you for highlighting their helpfulness. We will move some to Sec 2.
- **Gap between upper- and lower-bound:** We will provide a discussion in the appendix showing that non-randomized UCB algorithms like Alg. 1 cannot avoid an mas-number dependency in a lower-order term. But whether any algorithm can achieve scaling with independence number in the main-order $\sqrt{T}$ term is an interesting open question.
- **Feedback graphs vs. other approaches to structured MDPs:** Good suggestion, We will add a brief discussion.

Reviewer 4:

- **Alg 3 assumes dominating set known:** The assumption of a known dominating set was made to simplify the presentation. As we briefly allude to in Appendix F.1, one can apply a sightly modified version of Alg. 3 to problems where a dominating set is unknown. One then has S tasks in the first phase (one to reach each state) and move on to the second phase as soon as a suitable dominating set was discovered (which we can test at run-time). The sample-complexity of this modified algorithm is identical up to log-terms. We will expand on this in Appendix F.1.
- **Benefit of multi-task learning process:** If Algorithm 3 did not use the multi-task learning process, this would yield a sample-complexity bound of order $\frac{\gamma\mu\widehat{S}H^2}{p_0}$ when the dominating set is known and $\frac{S\mu\widehat{S}H^2}{p_0}$ when it is unknown (since we then pay an additional linear factor in the number of states we want to learn to reach). This is substantially worse than the bound enabled by our analysis. We will expand on the sketch of why this is true in the paper.