We thank the reviewers for the positive feedback and their interest in our work! Below we address some questions.

Response to Reviewer 2: Empirical evaluation: Interestingly, we actually did an empirical evaluation in the earlier phase of the project. We found that UCBZero can actually learn the navigation policies to all states in a 20 x 20 random 2D maze with the same number of samples required for UCB-H to learn the navigation policy to a particular target state. Both algorithms are well-tuned for hyperparameters. We didn't include it in the submission because after all the algorithm is only for tabular RL and is not of much empirical value to real complex problems. But we indeed hope that our work provides some theoretical justification to the multi-task RL problem showing that efficient multi-task learning is at least possible in the tabular setting. **Variables used before defined**: Thank you very much for pointing that out. We will make sure to define them earlier in the paper in the revision.

Response to Reviewer 3: Thanks for the many technical questions! We are happy to clarify them. **Additional related works**: [Hazan et. al. 2019] is an interesting work that we are not aware of, which explicitly aims at visiting each state as uniformly as possible. In that sense, it is more similar to the reward-free paper of [Jin et. al. 2020], which explicitly aims at visiting each state enough time as a subgoal of the exploration phase. Our result, however, shows that the approach taken by Jin et. al. is too pessimistic (suffering from an additional factor of S) if the eventual goal is still to perform tasks. The other two papers [Even Dar et. al. 2009, Shani et. al. 2020] studies the problem of regret minimization under adversarial rewards, which is very different from our setting, where the rewards are still stochastic but only unseen during exploration, and the goal is still to perform best policy identification, rather than regret minimization because the regret is not even defined in our setting without a reward function. Overall, we think it's not very precise to define UCBZero as an algorithm that merely tries to visit all the state-action pairs in an MDP. Arguably, all provably efficient RL algorithms must visit all essential states enough times in order to guarantee either small regret or small sample complexity. The goal of UCBZero is to perform multi-task learning even when the reward is absent during exploration. We did show, however, that as a bi-product, UCBZero visits each state enough times to allow other forms of downstream learning tasks. **UCRL2 with reward zero and nonoptimal dependency on H**: These are great questions! And they are connected. Since UCBZero is essentially a zero-reward version of UCB-H, our dependence on H essentially inherits from UCB-H, which is known to be suboptimal because of its model-free nature. At a higher-level, our result is an example that the zero-reward version of a standard provably efficient RL algorithm can be adapted to the task-agnostic RL setting and achieve good sample complexity. It is therefore very reasonable to conjecture that a more efficient RL algorithm, e.g. UCRL2, if adapted in a similar way, can achieve a better dependency on H. This is a great direction for future work. **Necessity of** $\log \mathbf{N}$: Again, great question! The key observation is that we assume the rewards are stochastic, so even if you have lots of data on each state-action pair, given infinitely many stochastic reward functions, there will be one reward function whose instantiation deviates from its mean, with a large probability. As a very simple sample, consider that you want to estimate the mean of N Bernoulli random variables simultaneously. The number of times you need to sample each Bernoulli random variable will scale as $\log N$ due to union bound. In fact, this result is profound. It shows that the reward-free RL task defined in [Jin et. al. 2020] is impossible when the reward functions are stochastic. The bound must scale with $\log N$. All the questions are very good, and we could see why readers may have these confusions. We will make sure to discuss all of the above points in the revision. Thanks again for the very valuable feedback!

Response to Reviewer 4: Empirical works in task-agnostic RL: Yes, this is actually one of the big motivations behind this paper. We've seen a lot of empirical works on this topic (e.g. an ICLR workshop last year), but very little theoretical discussion, and our work aims to close the gap. We have one paragraph in the related work sections dedicated to the empirical work and mentioned algorithms such as Go-Explore and Hindsight Experience Replay (HER), but we are happy to add the missing ones you mentioned to the paragraph. **Dependency on H and N**: The dependency on H can potentially be improved upto H^2, matching the lower bound. See also response 2 to reviewer 3. The dependency on N, however, is shown to be unavoidable by our lower-bound. This is due to our more realistic assumption that only instantiations of reward are available rather than full reward function assumed in [Jin et. al. 2020]. In most empirical problems, however, N is small. For example, if one wants to learn to navigate to all states, $N = S$, and $\log S$ is negligible compared to the other S term. **Notation in Alg 1**: Sorry for the confusion. So both b_t and α_t means to be functions of t. $t = ++N(s, a)$ is used immediately in the next lines as inputs to b_t and α_t. We will make it clearer and add the suggested comment lines in the revision! **N dependent v.s. N independent**: Sorry again for the confusion. All tasks are always independent of each other. Here "dependent" means whether the sample complexity depends on the number of tasks N, and "N independent" means that the bound doesn't scale with N and therefore still holds finite even when the number of tasks goes to infinity. Thank you for pointing out the many typos and for the polishing suggestions! We will make sure to follow them in the revision.

Response to Reviewer 5: Thank you for the very positive feedback on our work! We agree that the algorithm is mainly of theoretical interest and the current gap between the empirical success of deep RL and the theoretical understanding is still large. We hope that our paper can provide some intuition to the (fast-growing) empirical task-agnostic RL community!