All Reviewers: Thank you for your effort and the insightful comments! We will revise our paper accordingly.

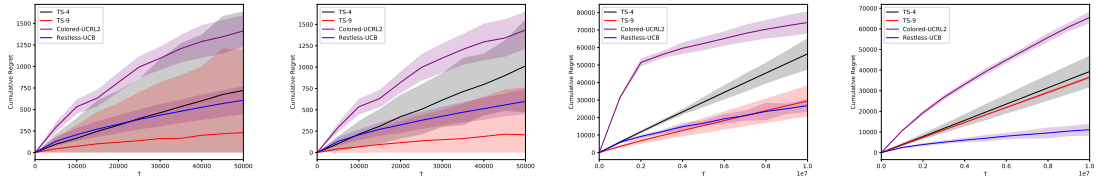
Common comment on the setting: (i) First, we would like to emphasize again that the birth-death Markov setting, though more restricted, has many applications in areas such as networks, communications and recommendations. (ii) Second, even under the birth-death setting, our algorithm is the only one with sublinear regret upper bound and polynomial time complexity. The time complexity of the Colored-UCRL2 in [Ortner et. al., 2012] ([28]) remains exponential with N under our setting. This is so because Colored-UCRL2 cannot explore the structure of the Markov chain, and only regards the restless bandit problem as an MDP problem, whose states are the belief states $z = \{(s_i, \tau_i)\}_{i=1}^N$ (as mentioned in lines 191-198 of our paper). Since the number of states is exponential in N, it means that Colored-UCRL2 need an exponential time complexity to find out the best policy for the MDP, even under our birth-death setting. This is also demonstrated by our experiment. In Table 1 (page 8 in our paper), we consider birth-death Markov chains, and the time cost of Colored-UCRL2 grows exponentially as N increases. (iii) Thirdly, our algorithm and analysis are not limited in the birth-death setting. In particular, Lemma 1 (which ensures low complexity) holds so long as we have the following property: $\forall i, k$, $P_i(k) \gtrsim P_i(k+1)$, where $P_i(k)$ represents the transition vector of arm i under state k, and $v \gtrsim v'$ is defined as $\forall k, \sum_{j=1}^{k} v_j \geq \sum_{j=1}^{k} v'_j$. Lemma 2 (which reduces the constant factor) also only requires that the Markov chains are ergodic. This means that it can be applied to reduce the constant factors in regret upper bounds in general restless bandit problems, such as [Ortner et. al., 2012] ([28]) and [Jung et. al., 2019] ([17]).

In particular, compared with [Ortner et. al., 2012] ([28]), there are two points we want to highlight: (i) even under our setting, our algorithm is the only one that achieves sublinear regret and polynomial complexity; (ii) our analysis helps to reduce the constant factor in the regret bounds of [Ortner et. al., 2012] ([28]). We will improve our explanation of the comparison to make the claims clear and avoid misunderstanding.

Reviewer 1

Q1: Adapting our results on stochastic processes that have infinite states or do hot have a birth-death structure. We can adapt our algorithm and analysis as long as the Markov chains M_i's are ergodic and have the property that $\forall i, k$, $P_i(k) \gtrsim P_i(k+1)$. For example, consider a discrete queueing system. In each time slot, there is at most one arrival (with probability λ) and similarly at most one departure (with probability μ), so that the corresponding Markov chain M has a birth-death structure. We also assume that the buffer size is infinity, so that M has infinite number of states. For this system, we can adapt our algorithm and analysis as long as $1 > \mu > \lambda > 0$. This is because that i) M is positive recurrent and aperiodic when $1 > \mu > \lambda > 0$; ii) in Markov chain M, $P(k, k+1) + P(k+1, k) = \lambda(1-\mu) + (1-\lambda)\mu < 1$, which implies that $\forall k, P(k) \gtrsim P(k+1)$.

Q2: About the error bar. Below are some experiments (Figures 1.(a), 1.(b), 1.(c) and 1.(d) in our paper) with error bar. We can see that Restless-UCB does not lead to a large variance. On the other hand, the TS policy suffers from a large variance when T is small. This is due to the high degree of randomness on the samples it draws when T is small.



(a) Figure 1.(a) (b) Figure 1.(b) (c) Figure 1.(c) (d) Figure 1.(d)

Q3: Is the M^3 factor tight enough? We conjecture that a lower bound will be $\Omega(M^2)$. Formally establishing this result is an interesting future research topic.

Reviewers 2 and 4

Q4 - Relation to prior work. Following your helpful suggestions on the weak regret, non-stationary bandits and assumptions in existing restless bandit solutions, we will including more related works and discuss the relation with them in details.

Q5 - About the oracle. We use the offline policy proposed in [Liu and Zhao, 2010] as the oracle in our experiments, and we will clarify this in the paper. In particular, [Liu and Zhao, 2010] proposes an offline policy for the case where all Markov chains only have two states. Note that in addition to the exact oracle, Restless-UCB policy can also be combined with approximate oracles to maintain low time-complexity, such as [Guha et. al., 2010] and [Liu and Zhao, 2009], while achieving performance guarantees. This is a unique feature not possessed by other existing algorithms.