We thank all reviewers for their comments. Minor comments will be addressed in the final version.

Reviewer 1

Clear description of the setting We want to emphasize first that our problem setting is a standard restless bandit setting with a few specific choices. P_k^{active} and P_k^{passive} are the transition matrices of the arm k when it is pulled or not, respectively. X_t is a K dimensional binary vector such that the k^{th} component $X_{t,k}$ represents the reward of the arm k. Since the learner only observes the rewards of pulled arms, only the N components X_{t,A_t} will be available to the learner. These notions are defined in lines 49 - 58. We will make our description clearer in the final version.

Messages of the experiments The first experiment empirically checks the Bayesian regret of our algorithm is indeed $\tilde{O}(\sqrt{T})$. The second experiment shows the algorithm still works in the frequentist setting. Figure 3 (left) illustrates how the value function of the policy π_l chosen in an episode converges to the baseline value for a variety of competitor mappings (the best fixed action, the myopic policy, and the Whittle index policy). Figure 3 (right) shows the posterior weights on the true parameters monotonically increase. We will describe the details of our experiments more carefully and make figures more readable.

Reviewer 2

1. Motivating application of the episodic setting Yes, the assumption of periodic restart of the system is somewhat limiting, and the regret analysis in the infinite time horizon is an interesting open question. Analyzing the finite horizon case should be an intermediate step towards solving this question. Moreover, the episodic case itself has a few motivating applications. For example, in the dynamic channel access problem that we consider in our experiment, the channel provider might reset their system every night when network traffic is low for maintenance related reasons. After the reset, every channel should be available for use, which can be thought as the beginning of a new episode.

2. Super-time-instant It is indeed possible to tackle the problem by considering each deterministic policy as an arm. However, this would result in very large (possibly infinite) K, the number of arms, and the existing bounds become vacuous as they depend (polynomially) on K. The bound in Dai et al. [2011] is meaningful since there are only two competing policies. This perspective still conveys interesting points, and we will add the comparison in the final version.

3. More complete picture in intro We totally agree that existing results, including ours, are just limited in different aspects. We will clarify this point more clearly. Nevertheless, we want to point out that this is the *first* paper that analyzes Thompson sampling in *multi-armed* restless bandit problems.

4. The optimal policy depends on L Yes, the optimal policy will depend on the episode length. It will change the baseline value in our regret definition in (2), but the same regret bound will still apply. It is one of our main contributions that the regret bound applies regardless of the choice of the benchmark.

Reviewer 3

Finer analysis within each episode First of all, the point raised by the reviewer is completely true. The episode length L should remain small to make the regret bound meaningful. We mainly considered the case where L is fixed as a constant and the number of resets, m, increases arbitrarily so that the posterior distribution concentrates sufficiently around the truth. A fundamental reason why we did not do the finer analysis within the episode is because our algorithm fixes a policy π_l and runs it throughout the episode l. If we get to do the finer analysis, then that means our algorithm changes the policy more often, which comes with an extra cost. For example, in the regret analysis by Ouyang et al. ["Learning Unknown Markov Decision Processes: A Thompson Sampling Approach," NIPS 2017], who analyze Thompson sampling in non-episodic fully observable MDPs, the bound includes K_T, the number of different policies that Thompson sampling runs up to time T.

Tightness of the regret bound As pointed out in the remark right after Theorem 5, our result reproduces the regret bound of $O(\sqrt{KT\log T})$ in the stationary MAB problem, whose lower bound is shown to be $\Omega(\sqrt{KT})$. This suggests that our bound is optimal in K and T up to a logarithmic factor. When $L = 1$, the problem becomes a combinatorial bandit problem (of choosing a set of N active arms out of a total of K) in which case the best known regret bound is $O(\sqrt{KN^3T\log K})$ (e.g., see "Combinatorial bandits" by Cesa-Bianchi and Lugosi [2012]. Their bound is actually $O(\sqrt{KNT\log K})$, but they normalize the loss to be in $[0, 1]$, whereas our reward is in $[0, N]$). Our bound agrees with their bound up to logarithmic terms. Finally, the optimal dependence on L remains open. We will add the discussion of tight dependence in the final version.