We thank all reviewers for their positive comments. Below we first address common concerns among the reviewers, and then respond to questions raised by individual reviewers.

1. Response to common concerns

-*"Knowledge of upper bounds of P_T and D_T"*: We remark that this type of assumptions is common and standard in literature on dynamic regret analysis of RL algorithms; see e.g. [22, 27, 39]. And even with access to upper bounds of P_T and D_T, it was unclear how to achieve dynamic regret bounds for policy optimization as our paper does. We do agree that it will be interesting to investigate the setting without these assumptions; we will pursue this direction by using the techniques developed in [12].

- *"Full-information reward feedback"*: Such assumption is standard in literature on RL problems with non-stationary rewards; see e.g. Even-Dar et al, "Online Markov Decision Processes" (2008). Extension of our results to the case of bandit feedback is reasonably straightforward by augmenting our algorithms with a reward estimator similar to [18]. We will explore this direction in future work.

- *"Efficiency compared to previous algorithms"*: Previous algorithms with dynamic regret guarantees are UCRL-based and need to solve large linear programs in each step of each episode. This makes such algorithms prohibitively expensive in computation and memory on practical problems. On the other hand, our algorithms do not require solving linear programs and all of their steps can be computed efficiently. We will add this discussion in our final paper.

- *"Numerical experiments"*: Our paper focuses on theoretical aspects of non-stationary RL. It is an excellent suggestion to conduct numerical experiments to support our theoretical results. We will follow up on this.

2. Response to individual reviewers

Review #2

- *"W1, algorithmic novelty"*: In addition to the restart mechanism, our Alg 2 features OMD steps for active prediction, which helps it achieve a better dynamic regret bound than our Alg 1; see Sec 3.2, as well as Thm 2 and the remarks beneath it for details. To the best of our knowledge, this is the first time that OMD steps are used in RL algorithms for tackling non-stationary environments.

- *"W2, full-information reward feedback"*: Please see our responses in the previous section.

- *"W3, fixed length of execution"*: To obtain guarantees for varying execution duration, one may augment our algorithms with a "doubling" trick commonly used in literature. We will explore this extension in future work.

- *"W4, tightness of analysis"*: When the magnitude of non-stationarity is moderate or large and P_T is on the same order of change in rewards, the results in our Thm 1 and 2 (setting $D_T = KH^3$) match those of [6] wrt the order of T under the multi-arm bandit setting, which is a special case of our episodic RL setting.

- *"C1, non-stationary environments"*: We agree that allowing varying transitions would give a more complete picture of non-stationary environments. On the other hand, we do believe that our setting, in spite of fixed transitions, is by itself an interesting and practical instance of non-stationary environments, as illustrated in Sec 1 and 2.3 of our paper.

- *"C2, decaying bonus over time"*: An excellent point. The purpose of bonus is to stabilize the algorithms under unknown transitions. Since we assume fixed transitions, there is no need for re-exploration.

- *"C3, not including reward in the LS objective"*: The two ways of including rewards are equivalent. We choose the current way as in our paper to streamline our proofs.

- *"C4, restart mechanism"*: When the level of non-stationarity is moderate or high, restarting is necessary to ensure the learning process is not adversely affected by the irrelevant historical reward information. Another approach that serves the same purpose is sliding window [12, 22]. Note that the master algorithm in [12] also employs a restart mechanism.

- *"R1, more efficient"*: Please see our responses in the previous section.

Review #3

- (1)–(3): Please see our responses in the previous section.

- *"Other comments"*: Thanks for pointing out the additional references. We will add them in our final paper.

Review #4

- *"Not practical, knowledge of upper bounds"*: Please see our responses in the previous section.

We appreciate the minor issues pointed out by the reviewers, and we will fix them in our final version.