We would like to thank the reviewers for their time. We humbly request for more of your time to reevaluate scores, given our responses below. We propose a general procedure to modify bandit algorithms for stochastic contextual bandit problems to be compatible with multiple adversarial bandit algorithms which allows us to obtain previously unattainable model selection regret guarantees that can be applied to a wide variety of problems in the setting of i.i.d. contexts. These assumptions (stochastic environments and i.i.d. contexts) are fairly standard in the literature. Because of its portability, this allows us to easily plug in multiple existing bandit algorithms and obtain regret guarantees for many problems for which no model selection guarantees were known before. Notably, we are able to obtain meaningful model selection guarantees even when the best base algorithm's regret is not fully known. We summarize our contributions:

(Section 4.1) Mis-specified contextual bandit with unknown error: We provide the first solution for the case of changing action sets. For fixed action sets, we improve upon the best existing result (Lattimore et al, 2020), and match the lower bound. To understand this result, consider k arms of dimension $d < \sqrt{k}$. If the arms are linear, then UCB has regret $\tilde{O}(\sqrt{kT})$ and LinUCB has $\tilde{O}(d\sqrt{T})$, so the best base's regret is $\tilde{O}(d\sqrt{T})$. If the arms are not linear, then UCB has regret $\tilde{O}(\sqrt{kT})$ and LinUCB has linear regret, so the best base's regret is $\tilde{O}(\sqrt{kT})$. It is impossible to know the best base's regret without knowing the environment, but our method achieves a regret matching the lower bound.
(Section 4.2) Linear contextual bandit with unknown dimension: For finite action sets, we provide a regret bound that does not depend on the number of actions, in contrast to the best existing result (Foster et al, 2019). We provide the first solution for infinite action sets. We provide the first solution when both the dimension and the mis-specified error are unknown. This is also the first result in literature that can combine multiple types of model selection.
(Section 4.2) Non-parametric contextual bandit with unknown dimension: we provide the first solution.
(Section 4.3) Tuning the exploration rate of ϵ-greedy: we provide the first solution for this problem.
(Section 4.4) Choosing between multiple feature maps for RL: we provide the first solution.
(Appendix A1) Generalized linear bandits with unknown link function: we provide this problem's first solution.
(Appendix A2) Bandit with heavy tail: we provide the first solution for this problem.

Reviewer 1: "The selection of the range": The regret is multiplied by at most a factor of the number of bases M, which is $\log(B)$ if the upper bound of the parameter is B. Therefore B can be chosen quite loosely as long as $\log(B)$ is not too large. In the paper we choose the largest ϵ to be $100,000$ but in practice such a large ϵ is unreasonable. Still, even with such a loosely chosen B, the performance is reasonably well.

We would like to emphasize that the exponential grid division of the parameter space is not central to our key contributions. How to define the best range and search efficiently in the parameter space is an interesting but separate research topic not in the scope of this paper.

Reviewer 2: Thank you for your comments. The citation to the non-stationary bandit paper was left from an earlier version of this submission, and will be removed.

Reviewer 3: <u>*Weaknesses:*</u> *A)* Relation to prior work in Section 4: In Section 4, we show that the strategy can be seamlessly used to solve a number of challenging open problems (reiterated above). We provide the first solution for these problems, and there are no existing solutions to compete with.

B) We would present the full description of the algorithm before Section 4.

C) "The assumption that Base algorithms only have access to rewards of rounds when they are selected...": This is not an assumption but a standard algorithmic design choice (modularity). It is not restrictive because we are able to produce a variety of state-of-the-art results (listed at the beginning). It is not clear if letting the base algorithms share the rewards will increase the performance.

<u>*Correctness:*</u> *A)* "I think it's not trivial to reproduce the results...": We would like more explanations as to why it would be difficult to reproduce the results. We believe we have provided all the details for our experiments. We have listed all configurations, parameter choices and number of repetitions. We also reproduced the master algorithms in Appendix B for convenience.

B) "...the instantaneous regret of Step 2 is 1/s times...": Let the cumulative regret of step 1 at round s be $U(s)$, then the cumulative regret of step 2 at round S is: $\frac{1}{1}U(1) + \frac{1}{2}U(2) + ... + \frac{1}{S}U(S) < (\frac{1}{1} + \frac{1}{2} + ... + \frac{1}{S})U(S) = \log(S)U(S)$. More details are in the proof in Lemma F1, page 24.

<u>*Additional feedback:*</u> *A)* "On line 94: what is i?": i is the index of the base.

B) "...cases where the action set changes...": In Step 2 of the smoothing procedure, we repeat the policy (which is the way to compute the chosen action), not the action. Even if the action set changes, the policy stays the same.

C) "...lines 6 through 11 in the pseudocode...": this is for the analysis of term II, as explained from line 90-98.