We would like to thank all the reviewers for the detailed and insightful comments.

Reviewer 1

Regarding applying our technique to other methods: The clipping trick and much of the subsequent analysis can be used to analyze other methods, including UCB-VI, UBEV, and other model-based methods. In fact, Simchowitz and Jamieson (2019) use the clipping trick to analyze model-based algorithms in the tabular setting, not the model-free one we analyze here. We do not include this in our paper because the model-based algorithms incur an S^2 dependence in the "lower order" K-independent term. This degrades the rate substantially in the nonparametric case, since effectively we set S in terms of the number of episodes K and the zooming dimension. So this term is not actually "lower order" in the nonparametric case, which leads to a worse final bound.

Regarding future work: Based on the above remarks, an intriguing question is whether model-based methods can achieve similar "zooming" guarantees via adaptive partitioning. A very recent paper of Sinclair et al. (`https://arxiv.org/abs/2007.00717`) make some progress in this direction, but does not obtain zooming-dimension dependence. We also mention two other questions: Can our techniques and adaptive guarantees extend to the infinite horizon discounted setting? Can adaptive discretization be combined with function approximation, analogous to the policy zooming approach of Krishnamurthy et al. (2019) for contextual bandits? This would allow us to use function approximation to deal with large state spaces and adaptive discretization to deal with large action spaces, which may be quite effective in practice. We can mention these directions in the final version.

Reviewer 2. Thanks for the kind words!

Reviewer 3. Thanks for catching the typos. We will fix them in the final version.

Regarding linear methods: This is a great question, but quite subtle! Remember, our method applies when $Q^\star$ is *Lipschitz* continuous. This is of course implied by $Q^\star$ being linear in some known features (assuming boundedness), but linearity is a much stronger assumption. A helpful analogy to think about is regression: if one knew the true regression function were linear, one should not use a nonparametric method. But nonparametric methods are applicable much more generally. It's the standard estimation error/approximation error tradeoff.

However, the RL setting is more subtle. If $Q^\star$ is Lipschitz, then work from Lipschitz bandits suggests that our method is optimal. If $Q^\star$ is linear (and that is all we assume), then our method acheives the $K^{\frac{d+1}{d+2}}$ regret rate, but actually we do not know of any better guarantees for this setting. In particular, the recent $\text{poly}(d)\sqrt{K}$ results for linear function approximation require much stronger assumptions, such as linear MDP or low inherent Bellman error. $Q^\star$ linear is a much weaker assumption than e.g., linear MDP, and it is not clear whether $\text{poly}(d)\sqrt{K}$ rates are achievable here at all.

Reviewer 4 Thanks for catching the typos. We will fix them in the final version.

Regarding zooming dimension vs covering dimension: We encourage the reviewer to examine the experiments of Sinclair et al., (2019), which shows (1) that adaptive partitioning performs much better than uniform discretization and (2) provides concrete examples and visualizations of this benefit. The zooming dimension gives a tighter bound for Sinclair et al.,'s adaptive partitioning scheme and captures the improvements observed in the experiments. Note that we analyze exactly their algorithm but their bound does not capture these improvements.

In addition, (1) the example we give in Figure 1 of our paper is actually quite general, as any problem where the near-optimal actions concentrate onto a low dimensional manifold has smaller zooming dimension than covering dimension; (2) Theorem 1 is even more refined than the zooming dimension bound; (3) zooming dimension and related quantities like "near optimality dimension" are widely studied in the bandit literature (20+ papers in the top ML conferences) and it is natural to extend these notions to RL.

Regarding the clipping trick: Please see Lemma 11 and 13 in the supplement for a mathematical statement. In words, instead of trying to bound the number of balls (which would be quite large), we notice that the regret incurred is determined by the sum of *clipped* surpluses for this ball, where we clip at level $\text{gap}/(H+1)$. This is because we only play a bad ball when there is a large error at a later time step, so we can credit this mistake to the later error and clip the current surplus. Using this decomposition, clipping will quickly take effect for "bad" balls, which allows us to bound the regret incurred by this ball.