We would like to thank the reviewers for their insightful comments.

Reviewer 1. We use the same assumption of knowing an upper bound to the optimal bias span as in (Bartlett and Tewari), but the algorithmic approach is very different. First, Fruit et al. ICML 2018 showed that the planning problem solved by REGAL.C is likely to be ill-posed and it is not clear how to derive a tractable algorithm to solve it (such an algorithm would likely be computationally inefficient). Furthermore, unlike (Fruit et al. ICML 2018), we introduce an exploration bonus formulation to the problem. This different algorithmic approach (the regret guarantees are indeed very similar) allows us to easily extend the algorithm to continuous MDPs and to achieve a "tighter" optimism, which results in a better empirical performance when Γ is big enough.

Reviewer 2. The best known minimax regret lower bound for discrete communicating MDPs is $\sqrt{DSAT}$ and the proof relies on the construction of a family of adversarial MDPs for which $\Gamma = 2$ is fixed (unlike D, S and A). It is still an open question whether a higher lower-bound can be shown (using a different family of adversarial MDPs for example) e.g., a lower-bound of order $\sqrt{DS\Gamma AT}$. If such a lower bound can be proved, then it is impossible to remove the dependence on Γ in the upper-bound (but maybe such a lower-bound does not exist, this is the reason why we mention that it is not clear if it is possible to remove Γ in the upper-bound or not). The comparison with the Politex paper is not simple due to the suboptimal dependence in T. The absence of explicit dependence in Γ is due to the assumption of uniformly fast mixing MDP. As shown in Thm 5 there is a term of order $T^{1/2}k$ where k is the mixing coefficient. This term hides the structural properties of the MDP. Moreover, the comparison is even more difficult due to the very suboptimal dependence on T (ie $cT^{3/4} + \epsilon T$). The main contribution in the continuous case is a practical algorithm with theoretical guarantees. UCCRL cannot be implemented as it requires solving the same planning problem as in REGAL.C (optimistic span-regularized optimization). As argued in (Fruit et al. ICML 2018) this problem may be ill-posed and there is currently no known algorithmic solution. The exploration bonus approach makes it easy to manage the continuous case and its analysis follows by combining the discrete case with ideas from UCCRL original analysis.

Reviewer 4. We show empirically that SCCAL+ is a very stable algorithm: in the experiments we did not optimize any parameter, we simply harmonized the confidence intervals. In fact, the low variance in the performance obtained across different runs shows that the performance of SCCAL+ is very consistent (i.e., high initialization is sometimes much better than low and viceversa, depending on the problem). On the other hand, RVI Q-learning is not very stable and the final performance heavily depends on the initialization of the Q-function. While there are results about the asymptotic behaviour of RVIQ, at the best of our knowledge, no finite-time guarantee and/or guidance on how to tune parameters is available. Concerning RVIQ-UCB, this is a first attempt to design a model-free algorithm for the average reward. This algorithm is inspired by the recent results in finite-horizon, but further work is needed to study its regret. In the final version, we will provide more interpretation of the experiments (in particular in the continuous case).

Presentation: thanks for the comments. We will provide clearer guidance to the reader through notation and tools/results from past references.