We would like to thank the reviewers for their important thorough reviews. We are happy to make use of their insightful comments in order to make this paper as clear and as complete as possible.

First of all, we would like to point out some important comments made by the reviewers, that will also be added to the final discussion of the paper. This paper presents the first no-regret algorithms for adversarial online MDPs with both an unknown transition function and a bandit feedback. It can and should pave the way for future algorithms to remove the loop-free assumption and to achieve tighter regret bounds. Handling both of these factors makes this problem a very hard one, as can be seen by the fact that it remained completely untouched while all the other variations of the online MDP setting were pretty much solved.

This setting combines two challenges that together become extra difficult, while previous papers only have to face one of them. When the transition function is known but feedback is bandit, this is actually an online linear optimization problem so importance sampling can be implemented without a problem. When the transition function is unknown but feedback is full information, the problem is more difficult but the learner does not need to estimate the loss, so she just needs to follow the regularized leader while estimating the transition function and building confidence sets around it.

When the two challenges are combined, this gives rise to new problems. First, the learner cannot use an unbiased estimator anymore and therefore the bias must be handled meticulously to make sure that it does not sum up to a large drift. Second, the learner cannot just follow the regularized leader because she has no idea what is going on in parts of the MDP that she did not visit enough, and the adversary can take advantage of the fact that the learner must explore more than in other scenarios. This is where the $\beta > 0$ assumption becomes so important, and getting rid of it is the most difficult part. The technique we use (perturbed transition function) and the way to analyze it are entirely novel, and it could pave the way for more advanced techniques that will not suffer the extra $O(T^{1/4})$ regret.

Reviewer $\#1$**:** A discussion with all the comments made above about hardness and comparison to [4] will be added.

Reviewer $\#2$**:** We would like to mention that the problem formulation is fully described in Section 2 and that there are exact explanations of our algorithms. We believe that most of the confusion stems from the lack of pseudo-code (which will be added) and from the partial introduction of the UC-O-REPS algorithm in Section 4. We will make Section 4 a lot more comprehensive and will lay the foundation to make sure that the algorithms that follow it are clear.

In section 4 the steps of the algorithm are presented in the equations after line 114, and the constraints that define the confidence sets are formally defined in the supplementary material. The part that may be incomplete is maintenance of the confidence sets, but we must emphasize that these are standard techniques. Nevertheless, for completeness, we will present the counters $N_t(x,a)$ and $N_t(x,a,x')$ that the UC-O-REPS algorithm uses to count number of visits up to time t. We will explain that the algorithm proceeds in epochs that their purpose is to shrink the confidence sets as much as possible while they still contain $\Delta(M)$ with high probability. The confidence sets are updated in the beginning of every epoch and an epoch ends once the number of visits to some state-action pair is doubled. Then we will give the definitions of the empirical transition function (based on the counters) and of the confidence sets $\Delta(M,t)$. We will explain that $\Delta(M,t)$ is a set of occupancy measures such that their induced transition function (as described in Section 3) has L_1-distance of at most ϵ_t from the empirical transition function, and present $\epsilon_t(x,a)$ as a parameter that controls the size of the confidence set and scales as $\tilde{O}\left(\sqrt{|X|/N_t(x,a)}\right)$. We will mention that $\Delta(M,t)$ is constructed using constraints that are linear with respect to the occupancy measure and refer to the supplementary material where there is already a full description of the constraints and there will be pseudo-code of UC-O-REPS for completeness.

In section 5.1 we explain the changes from UC-O-REPS but we also present the steps of the algorithm in the equations after line 135. The confidence sets are maintained as before (explained in Section 4) and the constraints are found in the supplementary material. We will give the full pseudo-code of the algorithm to avoid any confusion. We will also make it clear that the new constraint of $q(x) \geq 0$ is a linear constraint because $q(x) = \sum_{a,x'} q(x,a,x')$ (as described in Section 3). We will then refer to the supplementary material where a full description of the constraints is already given.

Section 5.2 will become a lot clearer once the UC-O-REPS algorithm is adequately explained in Section 4. A discussion about the motivation for defining the perturbed transition function will also be added. The reason for this perturbation is to enforce that the learner performs sufficient exploration, since now it is not guaranteed by the $\beta > 0$ assumption. The technical reason is that it enables us to use the Bounded Bandit UC-O-REPS algorithm (it requires the β assumption).

We would also like to point out that there is no summation missing in the computation of expectation in equation (1), in contrast to the comment of reviewer $\#2$. The probability that a state-action pair (x,a) was visited in episode t is exactly $q^{P,\pi_t}(x,a)$. Finally, all the technical comments made by reviewer $\#2$ will be fixed, and a discussion about the contribution of epochs and confidence sets to the regret analysis (which remains the same as [4]) will be added.

Reviewer $\#3$**:** Previous techniques can be found in many of the papers presented in the Related Work but the perturbed transition function technique is novel.