We thank all reviewers for their valuable comments! Below we address the issues raised by each reviewer.

R2: "most of the proofs read more like sketches". We thank the reviewer for pointing out the two places where more derivations are helpful and will revise accordingly. However, we respectfully disagree with the claim that *most proofs are like sketches* — indeed, our proof is 21 pages long, and R7 agrees that we have "given very detailed proofs for the regret derivation".

"The core ideas of this work are not novel ... the algorithm is essentially FTRL with the regularizer in Zimmert et al. 2019..." In fact, as we mentioned in the last paragraph of Sec 2, the regularizer $-\sum_a(\sqrt{q(a)} + \sqrt{1-q(a)})$ for multi-armed bandits is not even explicitly mentioned or analyzed in Zimmert et al. 2019, let alone the extra modification we need to do for MDPs (replacing 1 by $q(s) = \sum_a q(s,a)$ for state s). Also note that a direct extension of Zimmert and Seldin, 2019 does not work as we argue in the second paragraph of Sec 3, implying that careful thinking is needed to extend the ideas to MDPs. R6 also finds our "regularizer design very interesting". Algorithmic novelty aside, our analysis also contains many novel ideas, as the reviewer agrees (in the "Strengths" section).

"Δ_{MIN} factor in the bound comes exactly from the use of log-barrier ... not at all clear if this factor is necessary." To clarify, the Δ_{MIN} factor *does not* come from the log-barrier; instead, it comes from the second term is the penalty bound stated in Lemma 6, which is only about the Tsallis entropy part. It is indeed unclear if this factor is necessary, but to our knowledge, no existing algorithms, optimistic or not, enjoy a logarithmic regret bound without this dependence.

"how to implement the optimization problem in the FTRL step." Since this is a convex problem with $\mathcal{O}(L+|S||A|)$ linear constraints, one can apply any standard convex solver to implement the algorithm (accuracy $1/T$ is enough clearly).

R3: There seem to be quite some misunderstandings in the "Weaknesses" section, and we are not sure we fully understand all comments. We try our best to clarify below and sincerely hope that the paper can be re-evaluated.

"the algorithm does NOT attain optimal regret in the adversarial setting ... when $|S||A|L$ is a large number, the second term in the upper bound proposed by this paper will dominate." Our bound $\widetilde{\mathcal{O}}(\sqrt{L|S||A|T} + L|S||A|)$ *is* order-optimal (up to log terms). Please note that the optimal bound $\Theta(\sqrt{L|S||A|T})$ is only meaningful when $L|S||A| \leq T$ (since otherwise the regret is linear), and in this regime, the second term $L|S||A|$ in our upper bound is dominated by the first term. So indeed, our upper bound is simply $\widetilde{\mathcal{O}}(\sqrt{L|S||A|T})$, which is optimal.

"$0.04\sqrt{L|S||A|T}$ is the minimax lower bound when P is unknown." We assume that this specific bound is taken from Zimin and Neu, 2013 (Sec 5), which is a lower bound *even when P is known* (the entire paper of Zimin and Neu, 2013 is about known transition).

"In [23], the regret of their algorithm can be bounded by $2\sqrt{L|S||A|T}$ without the prior knowledge of the transition matrix." We are not sure we understand the comment — the work [23] (Simchowitz and Jamieson) is only for stochastic losses. For the adversarial case with unknown transition and bandit feedback, the best lower and upper bounds are $\Omega(L\sqrt{|S||A|T})$ and $\widetilde{\mathcal{O}}(L|S|\sqrt{|A|T})$ respectively as shown in Jin et al., 2020.

R6: Please see the last response to R2 on the implementation issue.

R7: Please see the last response to R2 on the implementation issue. Due to the complicated nature of the regularizer, our algorithm does not admit a "semi-closed form expression" unfortunately, but note that even though UC-O-REPS or UOB-REPS admits such a semi-closed form, it still requires solving a convex problem with as many positivity constraints (O-REPS, on the other hand, only requires solving an unconstrained convex problem).

"In Lemma 26, in line 709 ... should the expectation be removed?" No, the expectation should stay. This is because the expectation here is with respect to everything up to episode t (including episode t). What we do in line 709 is merely to first take the conditional expectation with respect to the randomness in episode t alone (so that $\mathbb{E}_t[\mathbb{I}\{s,a\}] = q_t(s,a)$), and after that, the expectation with respect to the past remains.

"do we need to consider the case $\|q' - q_t\|_H > 1$ as well?" No. If we can show that for all $\|q' - q_t\|_H = 1$, $G_t(q') \geq G_t(q_t)$ holds, then by convexity and the optimality of $\tilde{q}_t$, this implies $\|\tilde{q}_t - q_t\|_H \leq 1$, which is all we need. The same argument can be found in several earlier works, such as Lemma 9 of Lee et al. 2020.

"How is the first inequality in line 629 derived?" The inequality is equivalent to $-q_2(s')P(s|s',\pi_2(s')) \leq -q_2(s',\pi_1(s'))P(s|s',\pi_1(s'))$, which holds because the right-hand side is simply zero (to see this, note that $q_2(s',\pi_1(s')) = 0$ because π_2 is a deterministic policy that picks action $\pi_2(s')$ not equal to $\pi_1(s')$ at state s'). We will clarify this in the next version. Thanks for the question.