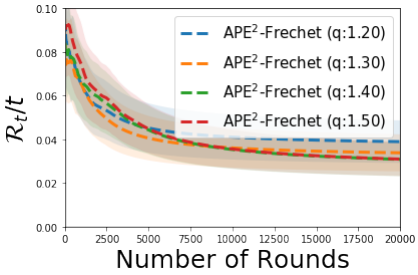
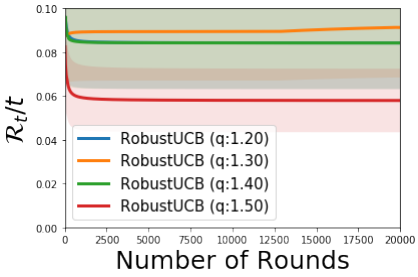
We are grateful for the valuable comments from reviewers, and provide the answer for each question as follows:

R1 (Q1) Role of perturbation in iid and adversarial setting: Perturbations determine the exploration tendency for both settings, and there exist suitable perturbations depending on the convergence speed of a reward estimator. Under the iid setting with the sub-Gaussian reward, the error of the empirical mean decreases squared exponentially fast, so it allows the sub-Weibull perturbation with $k \leq 2$ can achieve the near-optimal regret [8]. However, with the heavy-tail rewards, many estimators converge much slower, and it restricts the range of perturbations as sub-Weibull with $k \leq 1$ or heavy-tailed perturbations, as we demonstrated. For adversarial setting, there is no assumption on the distribution of rewards; hence, perturbations are required to cover the worst case. Thus, only heavy-tailed perturbations are used in the adversarial settings. Abernety et al. thoroughly analyzed the minimax regret bound of such perturbations.



(a) APE^2 with Frechet



(b) Robust UCB

R2 (Q1) Technical challenge and novelty: We extend the range of perturbation from the sub-Weibull to a broader class of distributions. In [8], the anti-concentration condition is a central assumption for the analysis of the regret bound under the sub-Weibull perturbation. However, the heavy-tailed perturbation, including GEV and Gamma, does not satisfy the anti-concentration condition. Hence, we propose a new framework (Assumption 2 in the main paper) which is a sufficient condition to ensure the bounded regret and generalizes the anti-concentration condition. To the best of our knowledge, this is the first result of heavy-tailed perturbations in the stochastic MAB. This discussion will be added to a revised version. **(Q2) Citation of (Medina & Yang 2016):** We appreciate for suggesting an important reference and will add it in the revised manuscript.

R3 (Q1) The idea of removing the need on prior information of the bound ν_p on the p-th moment: This idea has been first investigated in [6] (other related works did not address this problem mainly). However, there are significant differences from ours. We remark that [6] analyzes the upper bound of the *simple regret*, which focuses on finding the optimal action after T rounds, so it does not tell how much rewards will be lost during the exploration. We empirically observe that the algorithm in [6] shows the worst cumulative regret among all algorithms since minimizing the simple regret does not guarantee efficient exploration (See Figure E.1. in Appendix). On the contrary, we analyze the *cumulative regret*, which is an important metric to measure the efficient exploration. Hence, the proposed approach and analysis are independent of [6] while both works start with the same motivation. **(Q2) The theoretical contribution:** The proposed analysis is not incremental while it is related to [6, 8]. As we mentioned in Q1, our results are independent on [6]; and this is the first approach analyzing *heavy-tailed perturbations* in the stochastic MAB (See R2-Q1). The lower bound of the robust UCB is also an original contribution. **(Q3) A direct comparison between the *gap-dependent* bound of robust UCB and the proposed algorithm:** We mainly analyze the condition that the gap-dependent regret bound of the perturbation-based method is better than that of the robust-UCB, as explained in line 256-265. The main difference between perturbation-based methods and robust UCB is the dependency on Δ_a. It is the main reason why the proposed methods have better gap-independent bounds on T. **(Q4) Typo on line 259 & the figures size:** We deeply thank the reviewer for the comment; we will revise the typo and the figure size in the final paper.

R4 (Q1) The main driver for this research: From a practical perspective, reducing the number of tuning parameters makes the algorithm more robust. In particular, the perturbations do not depend on both bound and moment. So, the exploration tendency is not much sensitive to the mismatch of the moment parameter. To verify this, we add simple simulations by mismatching the moment parameter where all other settings are the same as the experiments in the manuscript. As shown in the above $\mathcal{R}_t/t$ plot, (a) APE^2 with Frechet perturbation shows a robust performance while (b) the robust UCB is sensitive depending on the choice of q, the moment parameter for the algorithm (here $p = 1.5$ is the true moment). Other perturbations show similar tendency. More extensive results will be included in the supplementary material. **(Q2) Originality of (C.6)-(C.8):** This trick itself appears in [1, 8]. Our contribution is utilizing the hazard function and proposing a new framework (Assumption 2) for a general class of reward distributions and perturbation strategies. **(Q3) Interpretation of (7):** (7) is not an equivalent assumption; it provides an interpretation under the assumption of Theorem 2 in the view of the ratio between the error and tail probability of the perturbation. We will revise the statement in the final paper.

References

[Abernety et al. 2015.] Abernethy, Jacob D., Chansoo Lee, and Ambuj Tewari. "Fighting bandits with a new kind of smoothness." Advances in Neural Information Processing Systems. 2015.