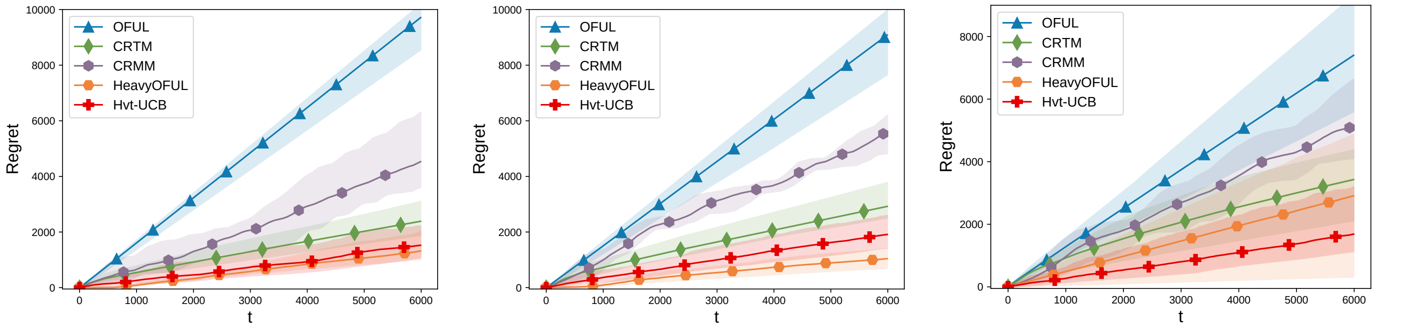
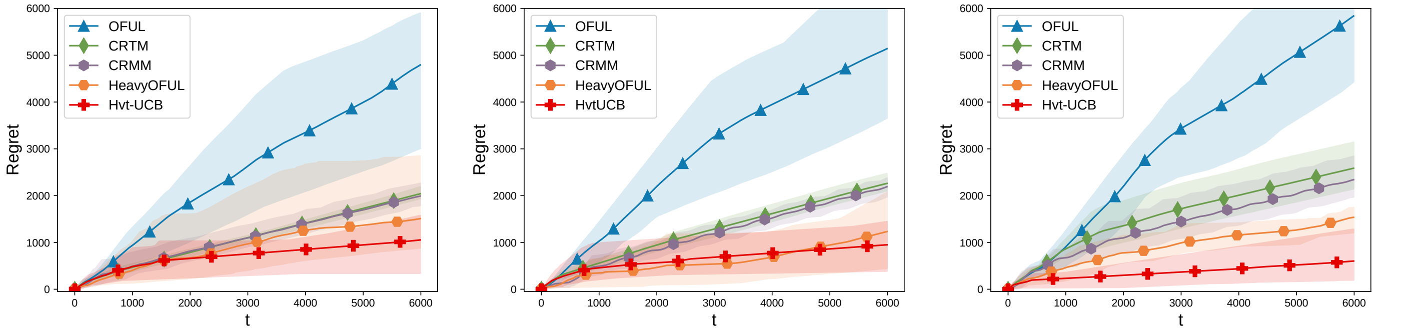


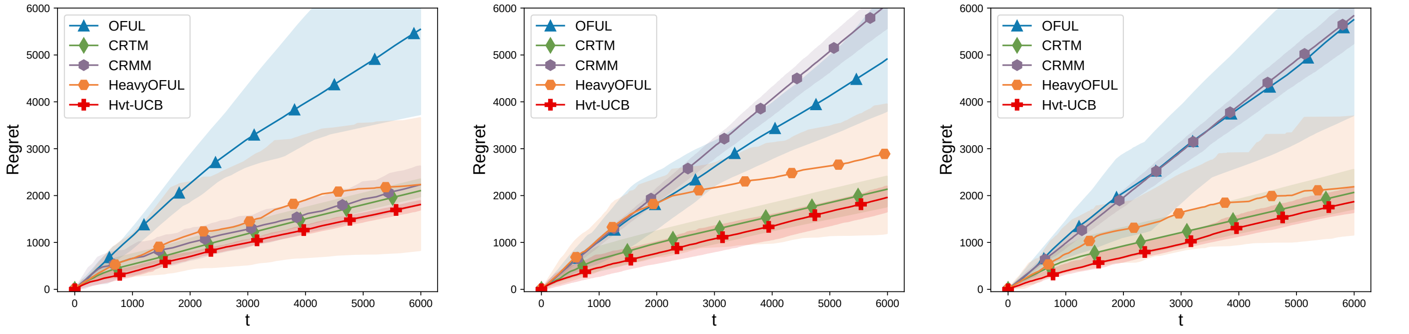
(a) Pareto distribution (b) Lomax distribution (c) Fisk distribution

Figure 1: [Reviewer C6wo, ZqBj] Experiment of Different Noise Distribution. *Configuration:* Besides the Student's t distribution, we consider three classic heavy-tailed distributions: (a) Pareto distribution, (b) Lomax distribution, and (c) Fisk distribution. We implement all three distributions using Python's scipy.stats. For the Pareto distribution, we set $b = 1.5$; for the Lomax and Fisk distributions, we set $c = 1.5$. We set $\epsilon = 1.49$, ensuring that the $(1 + \epsilon)$-moment exists. In the experiments, we use a time-varying noise scale, consistent with the setup in Figure 2. *Result:* We conducted 5 independent trials and averaged the results. The three figures show cumulative regret under different noise distributions. In both settings, our algorithm performs similarly to Heavy-OFUL, demonstrating competitive performance. Note that the CRMM algorithm performs poorly across all three distributions, as it can only handle symmetric noise, while these distributions are non-symmetric. In summary, this experiment validates the effectiveness of our algorithm in handling various noise distributions.



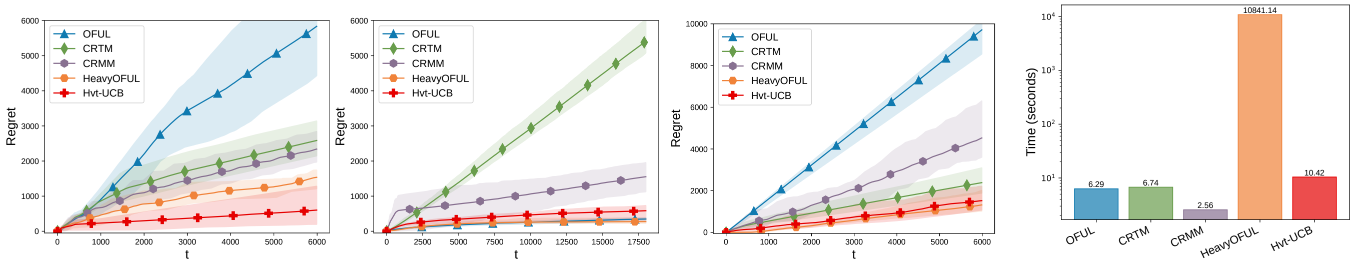
(a) scale of ν_t is 1.5 (b) scale of ν_t is 2 (c) scale of ν_t is 2.5

Figure 2: [Reviewer McYB, Reviewer ZqBj, Reviewer P7Rp] Experiment of Varying ν_t. *Configuration:* In this experiment, we consider a Student's t-distribution with a time-varying noise scale. Specifically, we fix the degrees of freedom as $df = 1.7$ and set $\epsilon = df - 0.01$. At each round, noise is first sampled from Student t(df) and then scaled by a factor α, where $\log_{10}(\alpha) \sim \text{Unif}(0, scale)$, so that the central moments of ε_t vary across rounds. In this setting, existing algorithms can only rely on an upper bound of the noise variance, whereas our proposed HvtLB algorithm leverages the actual per-round variance for more accurate and adaptive scheduling. This setting follows the experimental setup proposed by Huang et al. (2024). *Results:* We tested different scales of ν_t, with scales of 1.5, 2 and 2.5 in Figure 2(a), 2(b) and 2(c). In all environments, HvtLB and HeavyOFUL, which have variance-aware capabilities, outperform other algorithms significantly. In contrast, OFUL performs the worst because it lacks both variance-awareness and the ability to handle heavy-tailed noise.



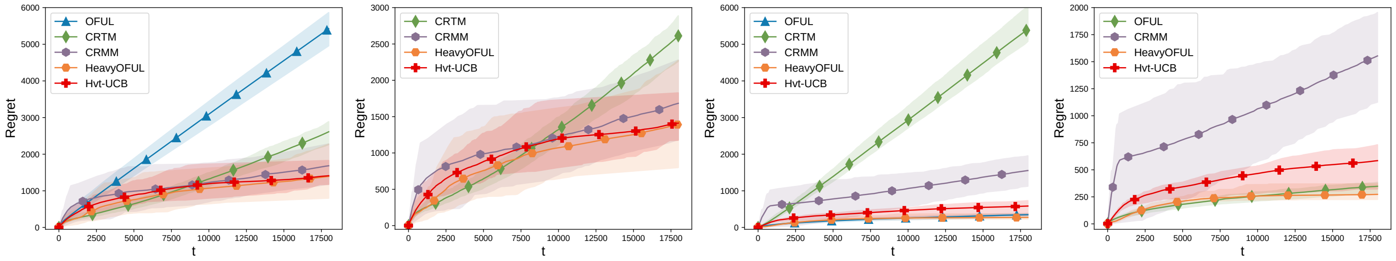
(a) fixed arm set (b) varying arm set 40/50 (c) varying arm set 30/50

Figure 3: [Reviewer C6wo] Experiment of Varying Arm Set Case. *Configuration:* In this experiment, we create a total arm set $\mathcal{X}$ with 50 arms, which are pre-sampled before the experiment begins. To make the arm set varying, in each round, we randomly select a subset from $\mathcal{X}$ as the current arm set $\mathcal{X}_t$, from which the algorithm can choose arms. In this case, the regret is based on the best arm in each round, rather than the global optimal arm. We consider three scenarios: (i) the arm set $\mathcal{X}_t = \mathcal{X}$ with 50 arms, (ii) randomly selecting 40 arms in each round, and (iii) randomly selecting 30 arms in each round. *Result:* We conducted 5 independent trials and averaged the results. The three figures show cumulative regret under different levels of arm set changes. In all settings, our algorithm performs the best, while HeavyOFUL and CRTM also achieve competitive results. However, when the arm set varies, the MOM-based algorithm fails. This is because MOM-based algorithms assume a fixed arm set and rely on resampling. When the arm set changes, resampling the same arms is not possible, so the algorithm selects the closest available arms instead. In summary, this experiment validates the effectiveness of our algorithm in scenarios where the arm set is time-varying.



(a) Time-varying ν case (b) Light-tailed case (c) Non-symmetric noise case (d) Average time cost

Figure 4: [Reviewer P7Rp] Questions of empirical performance and computational cost. In our paper, the performance of Hvt-UCB did not appear particularly advantages compared to previous methods. This is because the experiments were conducted under a fixed ν_t setting, where, theoretically, Hvt-UCB and previous methods share the same optimal regret guarantee. In our additional experiments, we introduce a variety of new settings to demonstrate the advantages of Hvt-UCB over existing methods. Figure 4(a) presents results under a time-varying ν_t setting with scale 2.5 (same as Figure 2(c)). We observe that both Hvt-UCB and HeavyOFUL outperform CRTM and CRMM. This is because Hvt-UCB and HeavyOFUL are variance-aware and can adapt to changes in ν_t, while CRTM and CRMM cannot. Figure 4(b) shows results in a sub-Gaussian scenario (origin Figure 2 in our paper). Here, CRTM performs poorly because its truncation mechanism is not well-suited for normally distributed data, rendering it ineffective. Figure 4(c) reports results under a Pareto distribution setting (same as Figure 1(a)). In this case, CRMM performs poorly due to its reliance on the median-of-means (MOM) estimator, which struggles with non-symmetric noise, leading to a drop in performance. Figure 4(d) presents the average time cost, with the vertical axis in logarithmic scale. It is evident that HeavyOFUL is significantly slower (about 800 times more computationally expensive) while our Hvt-UCB remains on the same order of magnitude in time complexity as existing one-pass algorithms (OFUL, CRTM, and CRMM). The efficiency of OFUL, CRTM, and CRMM comes from their use of closed-form least-squares solutions. Although Hvt-UCB requires a projection step in each round, its runtime remains competitive and is on the same order of magnitude as Least Squares-based algorithms. CRMM is faster than other one-pass methods, because it updates its estimates periodically rather than in every round, reducing the per-round overhead. In summary, our algorithm consistently demonstrates superior performance across diverse scenarios, while maintaining time complexity comparable to one-pass least-squares-based methods.



(a) Origin Figure 1 in our Paper (b) Scaled Figure 1 in our Paper (c) Origin Figure 2 in our Paper (d) Scaled Figure 2 in our Paper

Figure 5: [Reviewer P7Rp] Questions of Linear Appearance of Regret Curve. Some curves are linear due to algorithm failure, such as OFUL in Figure 5(a) and CRTM in Figure 5(c). Other curves appear flat because of the long time horizon of 18,000 rounds and compression of the y-axis. Similar linear appearance can be observed in the experiments of previous work (Xue et al., 2023; Huang et al., 2024). In Figure 5(b) and Figure 5(d), we removed the non-converging algorithms and scaled up the y-axis, clearly showing that the regret of algorithms are sub-linear.

References

Huang, J., Zhong, H., Wang, L., and Yang, L. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. to appear, 2024.

Xue, B., Wang, Y., Wan, Y., Yi, J., and Zhang, L. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. to appear, 2023.