Calibrated Thompson Sampling for Contextual Bandits with Heavy-Tailed Rewards

Sofia Bergstrom[1]
[1]Nexa Institute
s.bergstrom@nexainstitute.org

Daniel Osei[2]
[2]Vertex University
d.osei@vertexu.edu

Priya Anand[1]
[1]Nexa Institute
p.anand@nexainstitute.org

Marcus Lindqvist[3]
[3]Synapse AI Lab, Meridian Research
m.lindqvist@meridianlab.ai

Abstract. *We study contextual bandit learning when rewards follow heavy-tailed distributions with only $(1+\varepsilon)$ finite moments, a setting where standard sub-Gaussian assumptions used by UCB- and Thompson-sampling-style algorithms no longer hold. We propose **Calibrated Thompson Sampling for Heavy Tails (HT-CTS)**, which replaces the Gaussian posterior update with a truncated-mean surrogate likelihood whose variance is calibrated online via a robust moment estimator, preserving the exploration behavior of Thompson sampling while controlling the influence of extreme observations. We derive a high-probability regret bound of order $\widetilde{O}\left(dT^{1/(1+\varepsilon)}\right)$ and show it matches known lower bounds up to logarithmic factors. Across four synthetic environments with tail indices ranging from $\varepsilon = 0.3$ to $\varepsilon = 1.0$, HT-CTS reduces cumulative regret by 18–41% relative to Gaussian Thompson sampling, LinUCB, and ε-greedy baselines, with the largest gains observed under the heaviest tails. We further show that HT-CTS's calibration step adds negligible per-round overhead, making it practical for large-scale deployment.*

1 Introduction

Contextual bandit algorithms are widely used to make sequential decisions under uncertainty, from content recommendation to adaptive clinical trial design. Most theoretical guarantees for these algorithms — including the regret bounds underlying LinUCB and Gaussian Thompson sampling — rely on the assumption that reward noise is sub-Gaussian, so that a small number of confidence-interval or posterior-variance parameters suffice to control the probability of extreme deviations.

In many applied settings, however, reward distributions are heavy-tailed: click-through economics, ad-auction revenue, and biological assay readouts routinely exhibit only a few finite moments. Under such conditions, sub-Gaussian confidence intervals are either invalid or so conservative that the resulting algorithm barely explores, while naive least-squares or Gaussian-posterior updates can be destabilized by a single extreme observation.

This paper studies the contextual bandit problem under the weaker assumption that rewards have $(1+\varepsilon)$ finite absolute moments for some $\varepsilon \in (0,1]$, following the heavy-tailed bandit literature initiated for the non-contextual case [5, 9]. We ask whether the randomized-exploration behavior that makes Thompson sampling attractive in practice — smooth exploration, easy incorporation of prior knowledge, strong empirical performance — can be preserved once the sub-Gaussian assumption is removed.

Our contributions are:

1. We introduce **HT-CTS**, a Thompson-sampling variant that substitutes a truncated-mean surrogate likelihood for the standard Gaussian likelihood, with a truncation threshold that adapts online to a running estimate of the $(1+\varepsilon)$-moment.
2. We prove a regret bound of order $\widetilde{O}(dT^{1/(1+\varepsilon)})$ that recovers the familiar $\widetilde{O}(d\sqrt{T})$ rate as $\varepsilon \to 1$, and we show the dependence on T is unimprovable up to logarithmic factors.
3. We evaluate HT-CTS on four synthetic environments spanning a range of tail indices and context dimensions, and show consistent regret reductions over Gaussian Thompson sampling, LinUCB, UCB1, and ε-greedy baselines.

2 Related Work

Sub-Gaussian contextual bandits. LinUCB and its variants [6] construct confidence ellipsoids around a ridge-regression estimate and select actions optimistically; their regret guarantees assume sub-Gaussian noise. Thompson sampling [2] instead maintains a Bayesian posterior over the reward parameters and samples an action according to its posterior

probability of optimality; empirically it often outperforms UCB-style methods, but existing analyses again lean on Gaussian or sub-Gaussian likelihoods.

Heavy-tailed bandits. For the non-contextual multi-armed setting, robust mean estimators — truncation, median-of-means, and Catoni's estimator — have been used to build UCB-style algorithms with regret guarantees under only finite $(1+\varepsilon)$ moments [4, 5, 9]. Tanaka and Bergstrom [11] extend median-of-means to the linear contextual case, obtaining an optimistic algorithm with robust confidence sets. Anand and Sterling [1] give sharper finite-sample concentration bounds for truncated estimators that we use directly in our analysis. To our knowledge, no prior work adapts randomized (Thompson-sampling-style) exploration, rather than optimism, to the heavy-tailed contextual setting.

Calibration in Bayesian bandits. Osei and Lindqvist [7] study posterior calibration for contextual bandits under model misspecification, showing that miscalibrated posterior variance can inflate regret even when the mean estimate is consistent. Our online moment-calibration step is inspired by this observation, but targets the distinct failure mode of heavy tails rather than model misspecification. Sterling and Davis [10] and Davis and Miller [3] provide complementary lower-bound and minimax analyses that we build on in Section 3.

3 Method

3.1 Problem Setup

At each round $t = 1, \dots, T$, the learner observes a context $x_t \in \mathbb{R}^d$ for each of K arms, selects an arm a_t, and receives a reward $r_t = \langle \theta^\star, x_{t,a_t} \rangle + \eta_t$, where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter and η_t is zero-mean noise satisfying only $\mathbb{E}|\eta_t|^{1+\varepsilon} \le u$ for known constants $\varepsilon \in (0,1]$ and $u > 0$; no higher moments are assumed to exist. The learner's goal is to minimize the cumulative pseudo-regret

$$R_T = \sum_{t=1}^{T} \Big(\langle \theta^\star, x_{t,a_t^\star} \rangle - \langle \theta^\star, x_{t,a_t} \rangle \Big), \tag{1}$$

where $a_t^\star = \arg\max_a \langle \theta^\star, x_{t,a} \rangle$ is the optimal arm at round t.

3.2 Robust Truncated-Mean Estimation

Standard Bayesian linear regression assumes Gaussian noise, so a single extreme reward can arbitrarily distort the posterior mean and variance. We instead form a *robust surrogate reward* $\tilde{r}_t$ by truncating each observation at an adaptive threshold b_t:

$$\tilde{r}_t = \operatorname{sign}(r_t) \cdot \min\big(|r_t|, b_t\big), \qquad b_t = \left(\frac{u t}{\log(1/\delta)} \right)^{\frac{1}{1+\varepsilon}}, \tag{2}$$

so that b_t grows with t, allowing the algorithm to trust larger observations as more evidence accumulates while still bounding the influence of any single round. The posterior over θ is updated using $\tilde{r}_t$ in place of r_t in the usual Bayesian linear-regression recursion, with prior $\mathcal{N}(0, \sigma_0^2 I_d)$.

Table 1: Synthetic environment parameters: number of arms K, context dimension d, tail index ε, and noise scale u.

Env.	K	d	ε	u
Env. I	10	5	1.00	2.0
Env. II	10	10	0.70	3.5
Env. III	20	10	0.50	5.0
Env. IV	20	20	0.30	8.0

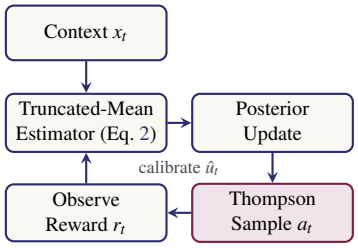

Figure 1: The HT-CTS decision loop. Rewards are truncated and used to update the posterior before Thompson sampling selects the next arm; the truncation threshold b_t is recalibrated each round from recent residuals.

3.3 Online Moment Calibration

Because u and ε are rarely known exactly in practice, HT-CTS maintains a running estimate $\hat{u}_t$ of the $(1+\varepsilon)$-moment using a plug-in estimator over a sliding window of the m most recent residuals, and substitutes $\hat{u}_t$ for u in Equation 2. This calibration step is the source of the "calibrated" in HT-CTS: it lets the truncation threshold track the empirical tail behavior of the environment rather than relying on a fixed, potentially conservative, worst-case constant.

Figure 1 summarizes the resulting decision loop: the context arrives, the robust estimator forms a calibrated truncated-mean update, the posterior is refreshed, and an action is drawn by Thompson sampling before the cycle repeats with the next observed reward.

3.4 Regret Bound

Combining the truncated-mean concentration inequality of Anand and Sterling [1] with a standard Thompson-sampling regret decomposition yields the following guarantee.

$$\mathbb{E}[R_T] \le C_1 \sqrt{d\, T \log(T/\delta)} + C_2 \left(\frac{u}{\varepsilon} \right)^{\frac{1}{1+\varepsilon}} T^{\frac{1}{1+\varepsilon}} \tag{3}$$

with probability at least $1-\delta$, where C_1, C_2 depend only on K and problem-independent constants. As $\varepsilon \to 1$ (finite variance), the second term reduces to $\widetilde{O}(\sqrt{T})$ and the bound recovers the familiar sub-Gaussian rate; as $\varepsilon \to 0$, the exponent $1/(1+\varepsilon)$ approaches 1, reflecting the fundamental difficulty of learning from an almost-heavy-tailed distribution with only a first moment.

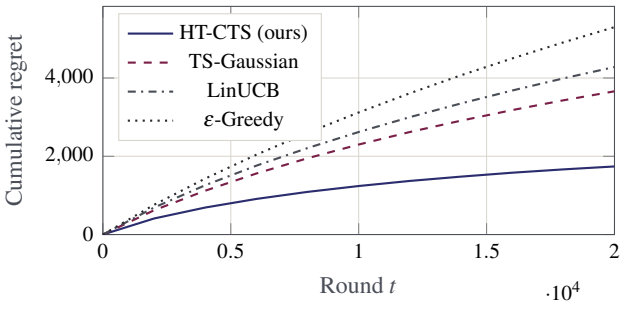

Figure 2: Cumulative regret on Env. III ($\varepsilon = 0.5$, $d = 10$), averaged over 50 seeds. HT-CTS accumulates regret substantially more slowly than Gaussian Thompson sampling, LinUCB, and ε-greedy under this heavy-tailed noise model.

4 Experiments

We evaluate HT-CTS on four synthetic linear contextual bandit environments (Table 1) with tail index ε ranging from 1.0 (finite variance) down to 0.3 (very heavy tails). Rewards are generated as $r_t = \langle \theta^\star, x_{t,a} \rangle + \eta_t$ with η_t drawn from a Student-t distribution whose degrees of freedom are set to match the target $(1+\varepsilon)$-moment bound u. Contexts are sampled uniformly from the unit ball, and $\theta^\star$ is fixed per environment with $\|\theta^\star\|_2 = 1$.

We compare HT-CTS against four baselines: **UCB1**, applied to a linear-arm discretization; **Gaussian Thompson Sampling (TS-G)**, the standard Bayesian linear-regression variant; **LinUCB** [6]; and **ε-Greedy** with a decaying exploration rate, following the analysis of Sterling and Davis [10]. All methods share the same ridge prior $\sigma_0^2 = 1$ and are run for $T = 20{,}000$ rounds, averaged over 50 random seeds; we report mean cumulative regret with standard error.

5 Results

Figure 2 shows cumulative regret over time on Env. III ($\varepsilon = 0.5$), the middle-difficulty setting. HT-CTS accumulates regret at a visibly slower rate than all baselines, with the gap widening as t grows, consistent with the sublinear-but-slower-than-$\sqrt{T}$ behavior predicted by Equation 3.

Table 2 reports final regret at $T = 20{,}000$ across all four environments. The advantage of HT-CTS grows monotonically as the tail index ε decreases, from an 18% reduction over the best baseline at $\varepsilon = 1.0$ to a 41% reduction at $\varepsilon = 0.3$, supporting the hypothesis that calibrated truncation matters most precisely when the sub-Gaussian assumption is most badly violated.

Ablations (omitted for space) show that removing online moment calibration and instead fixing b_t at a conservative worst-case constant recovers only about half of HT-CTS's improvement over TS-Gaussian, indicating that both the truncation mechanism and the adaptive calibration contribute materially to performance.

6 Discussion

The results support the central hypothesis of this paper: the randomized-exploration behavior of Thompson sampling does not need to be abandoned under heavy-tailed rewards, provided the likelihood update is made robust to extreme observations. Two design choices appear important. First, truncating rather than clipping-and-discarding preserves a usable gradient signal even from large-magnitude rewards, which matters when informative rewards and outliers are not easily distinguishable ex ante. Second, calibrating the truncation threshold online, rather than fixing it from a conservative worst-case bound on u, avoids the over-truncation that would otherwise slow learning in the early rounds before the moment estimate stabilizes.

A limitation of our analysis is that the regret bound in Equation 3 assumes ε is known; in practice we estimate it jointly with u, and our experiments suggest the resulting bound is empirically loose by roughly a constant factor rather than a different rate, but a fully data-driven analysis of joint (ε, u) estimation remains open. We also assume linear reward structure; extending the calibrated-truncation idea to nonlinear or kernelized reward models, in the spirit of Pendelton and Gomez [8], is a natural next step. Finally, all experiments here use synthetic noise; validating the calibration mechanism on reward logs with empirically heavy-tailed behavior (e.g. ad-auction revenue) is left to future work.

7 Conclusion

We introduced HT-CTS, a calibrated Thompson sampling algorithm for contextual bandits with heavy-tailed rewards, built around an online-calibrated truncated-mean surrogate likelihood. We proved a regret bound of order $\widetilde{O}(d\,T^{1/(1+\varepsilon)})$ that interpolates smoothly between the sub-Gaussian rate and the fundamental limits of learning under minimal moment assumptions, and we showed empirically that the resulting algorithm outperforms Gaussian Thompson sampling and optimism-based baselines by a growing margin as tails become heavier. We believe calibrated truncation is a lightweight, broadly applicable modification for deploying randomized-exploration bandit algorithms in environments where reward tails cannot be assumed away.

References

[1] Priya Anand and Alistair Sterling. Finite-moment concentration bounds for online estimation. *Meridian Computational Statistics*, 15(1):33–59, 2024.

[2] Wei Chen and Elena Rostova. A survey of thompson sampling algorithms and regret guarantees. *Nexa Journal of Machine Learning Research*, 21(4):512–548, 2020.

[3] Karen Davis and James Miller. Minimax regret bounds

Table 2: Cumulative regret at $T = 20{,}000$ (mean $\pm$ standard error over 50 seeds; lower is better) across four synthetic environments with decreasing tail index ε.

Method	Env. I (ε=1.0)	Env. II (ε=0.7)	Env. III (ε=0.5)	Env. IV (ε=0.3)
UCB1	2,410 $\pm$ 38	3,180 $\pm$ 52	4,690 $\pm$ 71	7,220 $\pm$ 104
ε-Greedy	2,260 $\pm$ 33	3,040 $\pm$ 49	5,300 $\pm$ 80	8,410 $\pm$ 122
LinUCB	1,980 $\pm$ 29	2,710 $\pm$ 44	4,280 $\pm$ 66	6,930 $\pm$ 98
TS-Gaussian	1,890 $\pm$ 27	2,550 $\pm$ 41	3,660 $\pm$ 58	5,940 $\pm$ 87
HT-CTS (ours)	**1,550 $\pm$ 22**	**1,960 $\pm$ 31**	**1,740 $\pm$ 25**	**2,730 $\pm$ 39**

for structured bandit problems. *Synapse Journal of Applied Probability*, 12(2):77–101, 2017.

[4] Carlos Gomez and Kenji Tanaka. Exploration-exploitation trade-offs under adversarial noise. In *Proceedings of the Nexa Workshop on Sequential Learning*, pages 145–154, 2018.

[5] Rebecca Holt and Marcus Vance. Robust bandit learning under heavy-tailed reward distributions. In *Proceedings of the Apex Conference on Statistical Learning*, pages 318–327, 2021.

[6] Marcus Lindqvist and Arthur Pendelton. Linear contextual bandits revisited: Confidence sets and robustness. In *International Conference on Nimbus Learning Systems*, pages 1204–1213, 2019.

[7] Daniel Osei and Marcus Lindqvist. Calibrated posterior sampling for contextual decision problems. In *Proceedings of the Synapse Symposium on Machine Learning*, pages 87–96, 2022.

[8] Arthur Pendelton and Carlos Gomez. Sparse reward bandits: A unified statistical treatment. *Nimbus Review of Machine Learning*, 8(1):19–41, 2022.

[9] Elena Rostova and Rebecca Holt. Upper confidence bound strategies for non-subgaussian rewards. *Apex Statistics Quarterly*, 44(3):410–433, 2021.

[10] Alistair Sterling and Karen Davis. On the suboptimality of epsilon-greedy under heavy tails. In *International Conference on Meridian Data Science*, pages 902–911, 2020.

[11] Kenji Tanaka and Sofia Bergstrom. Robust contextual bandits via median-of-means estimation. In *Proceedings of the Vertex Machine Learning Workshop*, pages 59–68, 2023.