Optimism Stabilizes Thompson Sampling for Adaptive Inference

Shunxing Yan* SHUNXINGYAN@OUTLOOK.COM
Department of Statistics and Data Science, Tsinghua University

Han Zhong* HANZHONG.WORK@GMAIL.COM
Antai College of Economics and Management, Shanghai Jiao Tong University

Editors: Steve Hanneke and Tor Lattimore

Abstract

Thompson sampling (TS) is widely used for stochastic multi-armed bandits, yet its inferential behavior under adaptive data collection is subtle. Classical asymptotic theory for sample means can fail because arm-specific sample sizes are random, history-dependent, and coupled with the observed rewards through the action-selection rule. A useful sufficient condition for valid asymptotic inference is *stability*, which requires each arm's pull count to concentrate around a deterministic scale. While stability is now understood for several UCB-type algorithms, vanilla TS can be unstable, leading to nonstandard asymptotics and potentially invalid Wald-type confidence intervals.

We identify *optimism* as a general mechanism for stabilizing Thompson sampling. In the K-armed Gaussian bandit with any fixed $K \geq 2$, we study two optimistic TS variants. The first is TS with posterior variance inflation; the second keeps the posterior variance unchanged but adds an explicit optimism bonus to the posterior mean. For both variants, we prove stability: optimal arms asymptotically share the horizon uniformly, while each suboptimal arm is sampled on a sharp gap-dependent logarithmic scale. For variance-inflated TS, this resolves the open problem posed by Halder et al. (2025) by extending their two-armed stability theory to general K-armed bandits, including instances with multiple optimal arms. For the mean-bonus variant, our result shows that stability can also be achieved through a direct optimistic shift of the posterior center, without inflating the posterior variance.

The main technical novelty lies in the treatment of variance-inflated TS with multiple optimal arms. In this regime, stability requires proving that the randomized competition among statistically indistinguishable optimal arms converges to a deterministic allocation. We isolate a limiting pure-noise competition and prove a negative-feedback property: over-sampled optimal arms become less likely to win future posterior draws, while under-sampled ones become more likely to be selected. This yields a contraction toward the uniform allocation over the optimal set. Concentration and rare-event estimates then control the perturbations caused by empirical-mean errors and occasional suboptimal selections. For the mean-bonus variant, we use a separate argument based on posterior-sampling concentration and UCB-type comparisons, since optimism enters through a deterministic shift of the posterior mean rather than through variance inflation.

These stability results imply asymptotically valid adaptive inference. In particular, for either optimistic TS variant, the usual studentized sample mean is asymptotically standard normal, and standard Wald confidence intervals achieve the nominal coverage probability despite adaptive sampling. Thus, suitably implemented optimism stabilizes Thompson sampling and enables classical inference from adaptively collected bandit data, while incurring only a mild additional regret cost.[1]

Keywords: Thompson sampling, multi-armed bandits, adaptive inference, optimism

* Author names are listed in alphabetical order.

1. Extended abstract. Full version appears as https://arxiv.org/abs/2602.06014, v2; see also Yan and Zhong (2026).

References

Budhaditya Halder, Shubhayan Pan, and Koulik Khamaru. Stable thompson sampling: Valid inference via variance inflation. *arXiv preprint arXiv:2505.23260*, 2025.

Shunxing Yan and Han Zhong. Optimism stabilizes thompson sampling for adaptive inference. *arXiv preprint arXiv:2602.06014*, 2026.