Adaptive Variance Inflation in Thompson Sampling: Efficiency, Safety, Robustness, and Beyond

Feng Zhu
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139
fengzhu@mit.edu

David Simchi-Levi
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139
dslevi@mit.edu

Abstract

Thompson Sampling (TS) has emerged as a powerful algorithm for sequential decision-making, with strong empirical success and theoretical guarantees. However, it has been shown that its behavior under stringent safety and robustness criteria — such as safety of cumulative regret distribution and robustness to model mis-specification — can sometimes perform poorly. In this work, we try to address these aspects through the lens of adaptive variance inflation for Gaussian Thompson Sampling. Our one-line change introduces a time- and arm-dependent inflation factor into the sampling variance, and yields several compelling benefits. The resulting policy achieves provably worst-case optimal expected regret and worst-case optimal fast-decaying regret tail bounds, even in the presence of heavy-tailed (sub-exponential) noise or mis-specified environments. The policy is also robust to mis-specified noise variances. Beyond cumulative regret, we further demonstrate that our method ensures strong post-experiment guarantees: simple regret and estimation error per arm exhibit fast-decaying tail probabilities, contributing to more reliable and robust downstream decisions. Finally, we extend our policy to incorporate settings with unknown arm-specific variances and empirically validate the consistent performance of our approach across a range of environments.

1 Introduction

The stochastic multi-armed bandit (MAB) problem is a foundational framework for sequential decision-making under uncertainty, with broad applications ranging from recommendation systems [27] to clinical trials [33] and financial portfolio optimization [16]. A central challenge in MAB is balancing exploration and exploitation. Among the many algorithms proposed to address this trade-off, Thompson Sampling (TS), originally introduced in [32], has emerged as both a conceptually elegant and practically effective approach. As a Bayesian method, TS selects actions based on samples drawn from the posterior distributions over arm rewards. By naturally balancing exploration and exploitation through probabilistic sampling, TS enjoys near-optimal theoretical guarantees [19, 1, 2, 24, 3, 25, 20] and demonstrates excellent empirical performance [10, 25, 11]. Its simplicity of implementation and strong empirical results have contributed to its widespread adoption in real-world systems.

Despite its success, recent work has revealed notable limitations of Thompson Sampling, particularly regarding regret tail behavior [14, 30, 31] and statistical inference power [23, 29, 22]. Intuitively, this is because TS — like many other bandit algorithms — is primarily designed to maximize the *expected* cumulative reward *within* an experiment from an *instance-dependent* perspective. As a result, it tends to adapt quickly to a perceived optimal arm and reduce exploration prematurely. This limited exploration can lead to sub-optimal performance on other critical dimensions, such as safety under heavy-tailed risks and robustness to model mis-specification.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

While efficiency (i.e., good expected outcome) remains a central goal, safety (i.e., concentration of outcomes around the mean) and robustness (i.e., stability of both the mean and distribution with respect to hyperparameter tuning and environmental mis-specification) are equally important for practical deployment. In this work, we revisit Thompson Sampling from the perspective of efficiency, safety, and robustness, and ask:

How can we modify Thompson Sampling to achieve efficiency, safety, and robustness, ensuring strong performance both during and after an experiment?

1.1 Our contributions

Our work makes both methodological and practical contributions. We propose **Thompson Sampling with Variance Inflation (TS-VI)**, a simple yet effective extension of Gaussian Thompson Sampling that inflates the posterior sampling variance using a carefully designed time- and arm-dependent factor, with the goal of achieving strong theoretical guarantees and robust empirical performance — particularly in settings that involve multiple tasks and require safe decision-making.

Within-experiment regret control. We show that TS-VI achieves a worst-case optimal cumulative regret tail decay rate of $\exp(-\tilde{\Omega}(x/\sqrt{KT}))$, where x denotes the regret threshold. As a corollary, its worst-case expected regret grows at the optimal rate of $\tilde{O}(\sqrt{KT})$. Moreover, we demonstrate that TS-VI is robust: it maintains these performance guarantees even under mild mis-specification of policy hyper-parameters and environmental conditions.

Post-experiment decision quality. We establish that TS-VI facilitates high-quality post-experiment decisions. Specifically, both the simple regret (in best-arm identification) and the per-arm estimation error (in mean reward estimation) exhibit fast tail decay and optimal expected error rates. Our analysis further implies that TS-VI promotes exploratory behavior in worst-case environments.

Simulation validation and practical refinement. We complement theoretical results with numerical simulations, demonstrating that TS-VI is efficient, safe, and robust across diverse environments. To handle unknown arm-specific variances in practice, we incorporate Gamma-Normal Bayesian updates into our design and show that this amendment preserves the stability and effectiveness of the policy.

1.2 Other Related Work

Safety and robustness in bandit algorithms. Recent work has increasingly focused on understanding the safety and robustness of bandit algorithms, particularly in the context of within-experiment regret. Early studies by [6, 26] showed that regret concentration typically occurs only at a polynomial rate. More recently, [4] demonstrated that bandit algorithms targeting logarithmic expected regret can be fragile: a mis-specified risk parameter (e.g., the sub-Gaussian noise level) can lead to instance-dependent regret growing polynomially in T. Building on this line of work, [14, 30] established that standard bandit algorithms achieving instance-dependent $\tilde{O}(1)$ regret — such as TS and its variants — can suffer from poor tail behavior: the probability of incurring large regret decays slowly with the time horizon. [30] also show that this tail performance can be significantly improved under worst-case design. Further, [31] offers a comprehensive characterization of the trade-off between expected regret and tail risk, shedding light on the intrinsic tension between efficiency and safety.

Bandit experimental design. There is also a growing literature on understanding the quality of post-experiment decisions in bandit settings, shifting focus from cumulative regret to final outcomes. In best-arm identification [13, 8, 17, 28, 36, 35], a commonly used performance measure is simple regret, introduced by [7], which quantifies the gap between the optimal arm's mean and the selected arm's mean. This contrasts with the probability of selecting a sub-optimal arm [5], a metric that is highly sensitive to the smallest sub-optimality gap [9] and often becomes meaningful only at large sample sizes. For mean estimation tasks [12, 29, 34, 21], it has been shown that standard bandit policies achieving instance-dependent $\tilde{O}(1)$ regret often perform poorly due to insufficient exploration of sub-optimal arms. Much of this literature has focused on minimizing the estimation error in expectation or constructing (anytime) valid confidence intervals. However, questions of decision safety and robustness in these post-experiment settings remain largely unexplored.

Before proceeding, we introduce some other notations. Throughout the paper, we use $O(\cdot)$ ($\tilde{O}(\cdot)$) and $\Omega(\cdot)$ ($\tilde{\Omega}(\cdot)$) to present upper and lower bounds on the growth rate up to constant (logarithmic) factors, and $\Theta(\cdot)$ ($\tilde{\Theta}(\cdot)$) to characterize the rate when the upper and lower bounds match up to constant (logarithmic) factors. We use $o(\cdot)$ to present strictly dominating upper bounds. In addition, for any $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For any $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$.

2 Model and Setup

Let the number of arms be K. In each time $t = 1, 2, \cdots$, the decision maker (DM) needs to decide which arm $a_t \in [K]$ should be pulled. To be more precise, let $H_t = \{a_1, r_1, \cdots, a_{t-1}, r_{t-1}\}$ be the history prior to time t. When $t = 1$, $H_1 = \emptyset$. In time t, the DM adopts a policy $\pi_t : H_t \longmapsto a_t$ that maps the history H_t to an action a_t, where a_t follows a discrete probability distribution $\pi_t(a_t|H_t)$ on $[K]$. The environment then reveals an independent reward $r_{t,a_t} = \mu_{a_t} + \epsilon_{t,a_t}$ to the DM. Here, μ_{a_t} is the mean reward of arm a_t, and ϵ_{t,a_t} is an independent zero-mean noise term. We assume that ϵ_{t,a_t} is (σ, ν)-sub-exponential. That is, for any time t and arm k,

$$\mathbb{E}\left[\exp(\lambda \epsilon_{t,k})\right] \leq \exp\left(\lambda^2 \sigma^2 / 2\right), \quad \forall \lambda : |\lambda| < 1/\nu.$$

Let $\mu = (\mu_1, \cdots, \mu_K)$ be the mean vector. Let $\mu_* = \max\{\mu_1, \cdots, \mu_K\}$ be the optimal mean reward among the K arms. Note that DM does not know μ at the beginning. Let $\Delta_k = \mu_* - \mu_k$ be the gap between the optimal arm and the kth arm. For theoretical analysis, We assume $|\Delta_k| \leq 1$ (which is not necessarily known by the DM). Let Γ be all $\mu \in \mathbb{R}^K$ such that $|\Delta_k| \leq 1$. Let $n_{t,k}$ be the number of times arm k has been pulled *prior to* time t. That is, $n_{t,k} = \sum_{s=1}^{t-1} \mathbb{1}\{a_s = k\}$. We additionally define $t_k(n)$ as the time period that arm k is pulled for the nth time. Let $\mu_{t,k}$ be the empirical mean of arm k *prior to* time t. That is, $\hat{\mu}_{t,k} = \sum_{s=1}^{t-1} r_s \mathbb{1}\{a_s = k\} / \sum_{s=1}^{t-1} \mathbb{1}\{a_s = k\} = \sum_{s=1}^{n_{t,k}} r_{t_k(s)} / n_{t,k}$.

2.1 Evaluation metric

Denote $\mathcal{E} = (\mu; \sigma, \nu)$ as the environment parameter. Fix a time horizon $T \geq K$ (which may not be known a priori by the DM). We are interested in two types of tasks — within-experiment regret control and post-experiment decision quality, illustrated as follows.

Within-experiment regret. Define the cumulative regret of a policy π under the environment $\mathcal{E}$ up to time T as

$$R^\pi_{\mathcal{E}}(T) = \sum_{t=1}^{T} (\mu_* - \mu_{a_t}) = \sum_{k=1}^{K} n_{T+1,k} \Delta_k$$

For simplicity, we write $R^\pi_\mu(T)$ instead of $R^\pi_{\mathcal{E}}(T)$, but we need to keep in mind that $R^\pi_\mu(T)$ is dependent on the environment profile. We are interested in studying the efficiency metric $\sup_{\mu \in \Gamma} \mathbb{E}[R^\pi_\mu(T)]$ and the safety metric $\sup_{\mu \in \Gamma} \mathbb{P}(R^\pi_\mu(T) > x)$ for large x, and the robustness of these two metrics with respect to mis-specified policy hyper-parameters and environment parameters (such as (σ, ν)).

Post-experiment decision. We are interested in two post-experiment decisions: best arm selection and mean estimation.

Best arm selection. After T steps, the task is to select an arm $\hat{a}^*_T$ such that $\Delta_{\hat{a}^*_T}$ is as small as possible. In particular, we are interested in studying the efficiency metric $\sup_{\mu \in \Gamma} \mathbb{E}[\Delta_{\hat{a}^*_T}]$ and the safety metric $\sup_{\mu \in \Gamma} \mathbb{P}(\Delta_{\hat{a}^*_T} > y)$, and their robustness to policy hyper-parameters and environment parameters.

Mean estimation. After T steps, the task is to estimate the true mean of each arm such that the error is as small as possible. In particular, we are interested in studying the efficiency metric $\sup_{\mu \in \Gamma} \mathbb{E}[\|\hat{\mu}_{T+1} - \mu\|^2_\infty]$ and the safety metric $\sup_{\mu \in \Gamma} \mathbb{P}(\|\hat{\mu}_{T+1} - \mu\|^2_\infty\| > y)$, and their robustness to policy hyper-parameters and environment parameters. Here $\hat{\mu}_{T+1}$ is the estimated arm mean vector after the experiment.

Remarks on the worst-case analysis. The rationale for focusing on the worst-case scenario is twofold. First, worst-case analysis provides strong, uniform guarantees that hold across all environments. In particular, it allows us to investigate policy robustness with respect to both the mean reward vector μ and the broader environment profile (σ, ν). In contrast, instance-dependent optimal policies are often fragile, exhibiting unsafe and highly sensitive behavior in both within- and post-experiment performance [14, 30, 29].

Second, from a practical standpoint, the DM can only observe empirical quantities — namely, the cumulative reward and the sample mean of each arm:

$$\sum_{t=1}^{T} r_t = \sum_{k=1}^{K} n_{T+1,k}\mu_k + \sum_{t=1}^{T} \epsilon_{t,a_t} \quad \text{and} \quad \frac{1}{n_{t,k}} \sum_{\ell=1}^{n_{t,k}} r_{t_k(\ell)} = \mu_k + \frac{1}{n_{t,k}} \sum_{\ell=1}^{n_{t,k}} \epsilon_{t_k(s),k}.$$

The noise terms in both expressions are generally not controllable by the decision-maker. While these terms vanish in expectation, they remain significant when considering tail probabilities of high cumulative regret or large estimation error. In such cases, it is meaningful to focus on thresholds of the form $x = \Omega(\sqrt{T})$ or $y = \Omega(1/\sqrt{T})$, as any threshold below these (e.g., $x = o(\sqrt{T})$, $y = o(1/\sqrt{T})$) will typically be dominated by the noise, rendering tail bounds ineffective or uninformative.

2.2 Thompson Sampling with adaptive variance inflation

We present Gaussian Thompson Sampling (TS) in Algorithm 1 and introduce our one-line modification, TS with Variance Inflation (TS-VI), in Algorithm 2. Both algorithms follow a Bayesian updating framework that corresponds to placing an improper prior on each arm, $p(\mu_k) \propto 1$ (or, approximately, a weakly informative prior $\mu_k \sim \mathcal{N}(0, \sigma_*^2)$ with $\sigma_* \to \infty$), and assuming the reward distribution for arm k is Gaussian: $r_t \sim \mathcal{N}(\mu_k, \sigma_0^2)$ (see, e.g., [18]). The key distinction in TS-VI (Algorithm 2) is that each posterior sample is drawn from a Gaussian distribution whose variance is inflated relative to the standard posterior variance $\sigma_0^2/n_{t,k}$ by an adaptive factor $t/(Kn_{t,k})$. Importantly, this inflation affects only the sampling distribution; the Bayesian posterior update remains unchanged.

Algorithm 1 TS

Input: $\mathcal{A} = [K]$, σ_0^2.
Pull each arm once.
for $t = K+1, \cdots$ **do**
  For each arm k, draw a random sample

$$X_{t,k} \sim \mathcal{N}\left(\hat{\mu}_{t,k}, \frac{1}{n_{t,k}}\sigma_0^2\right). \tag{1}$$

  Take action $a_t = \arg\max_k \{X_{t,k}\}$.
  Collect reward $r_{t,a_t} = \mu_{a_t} + \epsilon_{t,a_t}$.
end for

Algorithm 2 TS-VI

Input: $\mathcal{A} = [K]$, σ_0^2.
Pull each arm once.
for $t = K+1, \cdots$ **do**
  For each arm k, draw a random sample

$$X_{t,k} \sim \mathcal{N}\left(\hat{\mu}_{t,k}, \frac{t/K}{n_{t,k}^2}\sigma_0^2\right). \tag{2}$$

  Take action $a_t = \arg\max_k \{X_{t,k}\}$.
  Collect reward $r_{t,a_t} = \mu_{a_t} + \epsilon_{t,a_t}$.
end for

3 Within-experiment regret

In this section, we study the safety and robustness behavior of TS-VI. Our goal is to build safety guarantees for the tail distribution of cumulative regret $R_\mu^\pi(T)$. Starting from the safety result, we also build guarantees for efficiency (low expected regret) and robustness (robust regret tail distribution).

3.1 Main results

Theorem 1 (Within-experiment regret) *Fix σ_0 and (σ, ν). Define $M(\sigma_0; \sigma, \nu) = 1 \vee \sigma_0 \vee \frac{(\sigma \vee \nu)^2}{\sigma_0}$. There exists absolute positive constants c and C such that for any $x \geq c \cdot M(\sigma_0; \sigma, \nu)\sqrt{KT \ln K \ln T}$, we have*

$$\sup_{\mu \in \Gamma} \mathbb{P}\left(R_\mu^\pi(T) > x\right) \leq \exp\left(-\frac{x}{C \cdot M(\sigma_0; \sigma, \nu)\sqrt{KT}}\right).$$

We provide a brief discussion on the proof for Theorem 1 and highlight how we address the proof challenges. Without loss of generality, we assume that arm 1 is the optimal arm. The proof is based on the fact that for *any* time t a sub-optimal arm k is pulled, we have

$$\hat{\mu}_{t,1} + \frac{\sqrt{t/K}}{n_{t,1}}\varepsilon_{t,1} \leq \hat{\mu}_{t,k} + \frac{\sqrt{t/K}}{n_{t,k}}\varepsilon_{t,k}$$

$$\Longleftarrow \left\{ \frac{\Delta_k}{2} \leq \frac{\sum_{\ell=1}^{n_{t,k}} \epsilon_{\ell,k}}{n_{t,k}} + \frac{\sqrt{t/K}}{n_{t,k}} \varepsilon_{t,k} \right\} \bigcup \left\{ \frac{\sum_{\ell=1}^{n_{t,1}} \epsilon_{t_1(\ell),1}}{n_{t,1}} + \frac{\sqrt{t/K}}{n_{t,1}} \varepsilon_{t,1} \leq -\frac{\Delta_k}{2} \right\} \tag{3}$$

To establish the desired bound, it suffices to control the probability of each of the two events in (3). However, there are two main challenges that render the techniques in [30, 31] insufficient for our setting.

The first challenge lies in relating Δ_k and $n_{t,k}$ to the time index t and the regret threshold x in a way that yields the desired tail decay dependence on x, K, and T simultaneously. This is nontrivial because $n_{t,k}$ and t are inherently intertwined. We address this by carefully designing a regret decomposition, showing that if the cumulative regret reaches a level x, then Δ_k must satisfy a precise lower bound that depends on K, T, and $t_k(n)$, and meanwhile, $n_{T+1,k}$ must be sufficiently large.

The second challenge arises from the noise term $\varepsilon_{t,k}$, which is a mean-zero random variable beyond the DM's control. This term could cause the second event in (3) to occur with non-negligible probability. To handle this, we analyze multiple values of t collectively: when $n_{T+1,k}$ is large enough, there exists — with high probability — at least one time t at which arm k is pulled and $\varepsilon_{t,1}$ exceeds a fixed constant $\eta > 0$. However, care must be taken in selecting the range of t considered — smaller t leads to smaller $\sqrt{t/K}$, which weakens the tail bound. Thus, bounding the overall tail probabilities is delicate, and we defer the full technical details to the supplementary material.

3.2 Implications

Efficiency. Theorem 1 implies that the expected regret of TS-VI is $O(\sqrt{KT \ln K \ln T})$, where in $O(\cdot)$ we are hiding a constant factor. In fact, we have that

$$\sup_{\mu \in \Gamma} \mathbb{E}[R_\mu^\pi(T)] \leq 2\bar{x} + \sup_{\mu \in \Gamma} \int_{x=\bar{x}}^{+\infty} \mathbb{P}\left(R_\mu^\pi(T) > x\right) dx \leq (2c + C) M(\sigma_0; \sigma, \nu) \sqrt{KT \ln K \ln T}.$$

The expected regret bound is worst-case optimal on both K and T up to a logarithmic factor. We would like to point out that in contrast to most approaches taken in the literature that obtain expected regret bound, the approach we take here is to first derive regret tail bound and then yield a guarantee in expectation. While the current work is not trying to obtain the best dependence on logarithmic factors, it would be interesting to see whether and how these logarithmic factors can be removed.

Safety. Theorem 1 gives the optimal regret tail decaying rate. As is shown by [30] through a two-armed bandit case, for the family of policies that obtains the worst-case expected regret performance guarantee $\tilde{O}(\sqrt{T})$, the tail probability $\mathbb{P}(R_\mu^\pi(T) > x)$ cannot be decaying faster than $\exp(-x/\sqrt{T})$ for large x. While [30, 31] only consider UCB-like deterministic policies, our result show that the standard TS policy, as a randomized policy, can also be amended to achieve the desired optimal safety guarantee.

Robustness. Lastly, we would like to emphasize the robustness performance of TS-VI, which can be of great importance in practice. Note that our policy is almost parameter-free (the only input parameter is K and σ_0) without assuming any knowledge to the distribution of the environment.

Knowledge of K. Typically K is known to the decision-maker. If K is not known a priori or not utilized in variance inflation, we can derive a worst-case guarantee similar to that in Theorem 1 with a sub-optimal dependence on K (but not on T) — by setting $\sigma_0' = \sqrt{K}\sigma_0$, the expected regret becomes $\tilde{O}(K\sqrt{T})$. In other words, the $1/K$ factor in the variance inflation term is a "recommended" scaling parameter that makes sure the worst-case rate has an optimal dependence on K.

Robustness to hyperparameter σ_0. TS-VI is robust to mis-specified σ_0, in the sense that for any σ_0 (as long it is positive), we can always achieve the optimal regret tail decaying rate as well as the optimal regret expectation growing rate, with only constant factors affected. Apparently, if we have prior knowledge on σ and ν (say in the sub-gaussian case we know σ and we know that $\nu = 0$), we can set $\sigma_0 = \sigma \vee \nu$ to obtain better constant factors. In Section 5, we will examine selection of σ_0 under various situations.

Robustness to environment mis-specification. A feature of our result is that we consider a sub-exponential environment, where the reward tail can be heavier than that from a sub-gaussian distribution. Regardless of whether the environment is sub-gaussian or sub-exponential, TS-VI does not

require any knowledge on environment profiles, and each random sample is drawn from a Gaussian distribution (instead of a heavier-tailed distribution that caters to the environment). As we will empirically show in Section 5, standard TS and some of it variants can lead to high hidden risk in the tail region under an exponential environment, while our policy can significantly alleviate the issue.

4 Post-experiment decision

In this section, we study the safety and robustness behavior of TS-VI for two types of post-experiment decisions. Our goal is to build theoretical safety guarantees for the tail distribution of simple regret $\Delta_{\hat{a}_T^*}$ and estimation error $\|\hat{\mu}_{T+1} - \mu\|_\infty$. Starting from the safety result, we build guarantees for efficiency (low expected simple regret and low expected estimation error) and robustness (robust simple regret tail distribution and robust estimation tail distribution). Our analysis also provide deeper insights on the exploration behavior of TS-VI.

For any $T \geq K$, we adopt the following decisions: (a) *Best arm selection.* We select the arm that is pulled the most often in the second half of the time horizon: $\hat{a}_T^* = \arg\max_k \{n_{T+1,k} - n_{\lceil T/2 \rceil + 1,k}\}$. (b) *Mean estimation.* For each arm k, we simply take the empirical mean $\hat{\mu}_{T+1,k}$.

4.1 Main results

Theorem 2 (Post-experiment best arm selection) *Fix σ_0 and (σ, ν). There exists absolute constants $c_1, C_1 > 0$ such that for any $y \geq c_1 \cdot M(\sigma_0; \sigma, \nu)\sqrt{\frac{K \ln K \ln T}{T}}$, we have*

$$\sup_{\mu \in \Gamma} \mathbb{P}\left(\Delta_{\hat{a}_T^*} > y\right) \leq \exp\left(-\frac{y}{C_1 \cdot M(\sigma_0; \sigma, \nu)}\sqrt{\frac{T}{K}}\right).$$

Theorem 3 (Post-experiment mean estimation) *Fix σ_0 and (σ, ν). There exists absolute constants $c_2, C_2 > 0$ dependent only on σ_0 and (σ, ν) such that for any $y \geq c_2\sqrt{\frac{K}{T}}\ln^4 T$, we have*

$$\sup_{\mu \in \Gamma} \mathbb{P}\left(\|\hat{\mu}_{T+1} - \mu\|_\infty^2 > y\right) \leq \exp\left(-\sqrt{\frac{y \wedge \sqrt{y}}{C_2}}\sqrt{\frac{T}{K}}\right).$$

The proof of Theorem 2 is based on Lemma 1, which indicates that the number of times that each sub-optimal arm is pulled for $\Omega(T/K)$ times is exponentially decaying with respect to the sub-optimality gap Δ_k. Intuitively, for the second half of the whole time horizon, with very low probability that TS-VI is continuously exploring or even sticking to any sub-optimal arm. Since we are selecting the arm that is pulled the most often, Theorem 2 then follows.

Lemma 1 *There exists absolute positive constants c_1', C_1' such that for any k and $\Delta_k \geq c_1' \cdot M(\sigma_0; \sigma, \nu)\sqrt{\frac{K \ln K \ln T}{T}}$,*

$$\mathbb{P}\left(n_{T+1,k} - n_{\lceil \gamma T \rceil + 1} > \frac{T}{2K}\right) \leq \exp\left(-\frac{\Delta_k}{C_1' \cdot M(\sigma_0; \sigma, \nu)}\sqrt{\frac{T}{K}}\right).$$

The proof of Theorem 3 is based on Lemma 2, which shows an interesting fact that TS-VI with very high probability explores each sub-optimal arm with at least $\Omega(\sqrt{T/K})$ times. Intuitively, TS-VI circumvents tail risk by inflating the variance and doing more exploration. Since the worst-case expected regret is $\tilde{O}(\sqrt{KT})$ (followed from Theorem 1), this leaves enough space for TS-VI to explore each arm on average for $\Omega(\sqrt{T/K})$ times, which leads to more estimation accuracy shown in Theorem 3.

Lemma 2 *There exists positive constants c_2', C_2' dependent only on σ_0, (σ, ν) such that for any k, we have*

$$\mathbb{P}\left(n_{T+1,k} < c_2'\sqrt{\frac{T}{K}}\right) \leq \exp\left(-\frac{1}{C_2'}\sqrt{\frac{T}{K}}\right).$$

4.2 Implications

Efficiency. Theorem 2 implies that the expected simple regret of TS-VI is $\tilde{O}(\sqrt{K/T})$, and Theorem 3 shows that the expected deviation between the empirical and true means is also $\tilde{O}(\sqrt{K/T})$. Both results follow from similar arguments as the expected regret bound in Theorem 1. These bounds are worst-case optimal in their dependence on K and T, up to logarithmic factors (see, e.g., [5, 29]). In particular, for Theorem 3, the $\tilde{O}(1/\sqrt{T})$ rate is known to be optimal for any policy achieving expected regret $\tilde{O}(\sqrt{T})$ [29].

Safety. Unlike prior works (e.g., [29]) that focus solely on expected performance, Theorems 2 and 3 also provide exponential tail bounds for simple regret and estimation error, respectively, offering reliability guarantees for decision quality. These are obtained by first deriving tail bounds — similar in spirit to Theorem 1 — and then translating them into expectation bounds. It remains an open question whether the logarithmic factors in these results can be improved or removed.

Robustness. TS-VI also exhibits robustness to both the prior variance σ_0 and environmental mis-specification. While additional environmental knowledge may improve empirical performance, incorrect assumptions only affect constant factors, without changing the asymptotic efficiency loss or tail decay rates.

5 Numerical Experiments

We conduct numerical experiments on the 2-armed bandit case to illustrate the benefits brought by our policy. The mean vector is fixed as $\mu = (-\delta, \delta)$ with $\delta = 0.3$. We focus on 4 empirical metrics:

(a) expected regret vs. t;

(b) log tail probability of cumulative reward < 0 vs. t;

(c) mean absolute estimation error per arm vs. t;

(d) log tail probability of absolute estimation error $> \delta$ vs. t.

For each policy considered below, we collect 10^4 trajectories. In (a) and (c), we also show the error bars (95% confidence interval) for the expected regret and the mean absolute estimation error. In the supplementary material, we also provide experiments for $\delta = 0.5$, and a 6-armed bandit case study with $\delta = 0.3, 0.5$.

5.1 Environments and Results

Well-specified environment with known variances. We first consider Gaussian environments where the noise variances are correctly specified. We consider $\sigma^2 = 2$ for both arms. Results are provided in Figure 1.

- For both TS and TS-VI, we assume the prior is $\mathcal{N}(0, 10^3)$.
- We consider the standard TS ($\sigma_0 = \sigma$), a slightly under-specified TS ($\sigma_0 = 0.9\sigma$), and a slightly over-specified TS ($\sigma_0 = 1.1\sigma$).
- For TS-VI, we consider $\sigma_0 = 0.3\sigma, 0.4\sigma, 0.5\sigma$ — we empirically find that a σ_0 being slightly less than the true σ yields empirically stronger performance.
- We also consider the UCB policy with the bonus term $\sigma_0\sqrt{2\ln t/n}$ [15], where $\sigma_0 = 0.9\sigma$ (slightly under-specified), $\sigma_0 = 1.0\sigma$ (standard), 1.1σ (slightly over-specified).

Mis-specified environment with known variances. We then consider Exponential environments with Laplacian noises — that is, the probability density function is $(2b)^{-1}\exp(-|x|/b)$. We consider $b = 1$ for both arms. Note that the variance of a Laplace distribution is $2b^2$. Results are provided in Figure 2.

- For both TS and TS-VI, we assume the prior is $\mathcal{N}(0, 10^3)$.
- For TS and UCB, we treat each sample as if it is drawn from a Gaussian distribution $\mathcal{N}(0, 2)$. We consider $\sigma_0 = 0.9\sigma, 1.0\sigma, 1.1\sigma$.

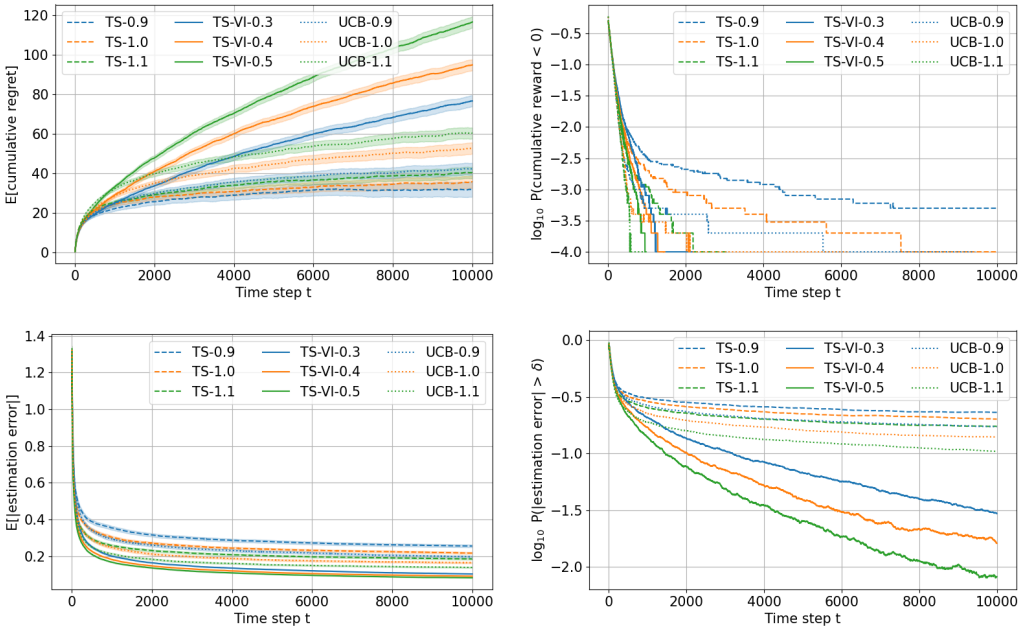

Figure 1: Results for the well-specified environment with known variances

- For TS-VI, we consider $\sigma_0 = 0.3\sigma, 0.4\sigma, 0.5\sigma$.

Mis-specified environment with unknown heterogeneous variances. Finally, we consider Exponential environments with noises following *unknown* Laplace distributions. We consider $b_1 = 1, b_2 = 2$. We focus exclusively on TS and TS-VI in this environment. Results are provided in Figure 3.

- Since variances are unknown, we adopt the Gamma-Normal TS where the variance of each arm σ_k^2 is modeled following an inverse Gamma distribution. The detailed paradigms of Gamma-Normal TS and its inflated version are provided in the supplementary material.
- For TS, we treat each sample as if it is drawn from a Gaussian distribution $\mathcal{N}(0, \sigma_k^2)$. The sampling variance is taken as $\sigma_{0,k} = 0.9\sigma_k, 1.0\sigma_k, 1.1\sigma_k$.
- For TS-VI, we consider $\sigma_{0,k} = 0.3\sigma_k, 0.4\sigma_k, 0.5\sigma_k$.

5.2 Observations and Implications

Efficiency. Across all environments, TS and UCB consistently exhibit lower expected within-experiment regret compared to TS-VI under various hyperparameter settings. This is expected, as TS-VI encourages greater exploration than TS, leading to more frequent pulls of sub-optimal arms — consistent with Lemma 2. However, we also observe that for smaller time horizons (e.g., $T \leq 1000$), the performance gap narrows and in some cases, TS-VI even achieves lower expected regret than TS and UCB. For larger horizons such as $T = 10000$, the increase in expected regret under TS-VI remains moderate, generally ranging from 20 to 100 (or 20 to 40 for TS-VI-0.3). This controlled sacrifice in regret is compensated by markedly improved estimation accuracy. In particular, across all settings, TS-VI consistently yields lower mean absolute estimation error compared to TS and UCB.

Safety. In terms of tail risk — both for within-experiment regret and post-experiment decision — TS-VI significantly outperforms TS and UCB. The empirical probability of incurring a negative cumulative reward or large estimation error for the best arm under TS-VI decays rapidly, typically approaching zero by $T \approx 5000$ (and in many cases as early as $T \approx 2000$). TS-VI also substantially mitigates the risk of large estimation errors for sub-optimal arms, whereas TS and UCB exhibits much slower tail decay.

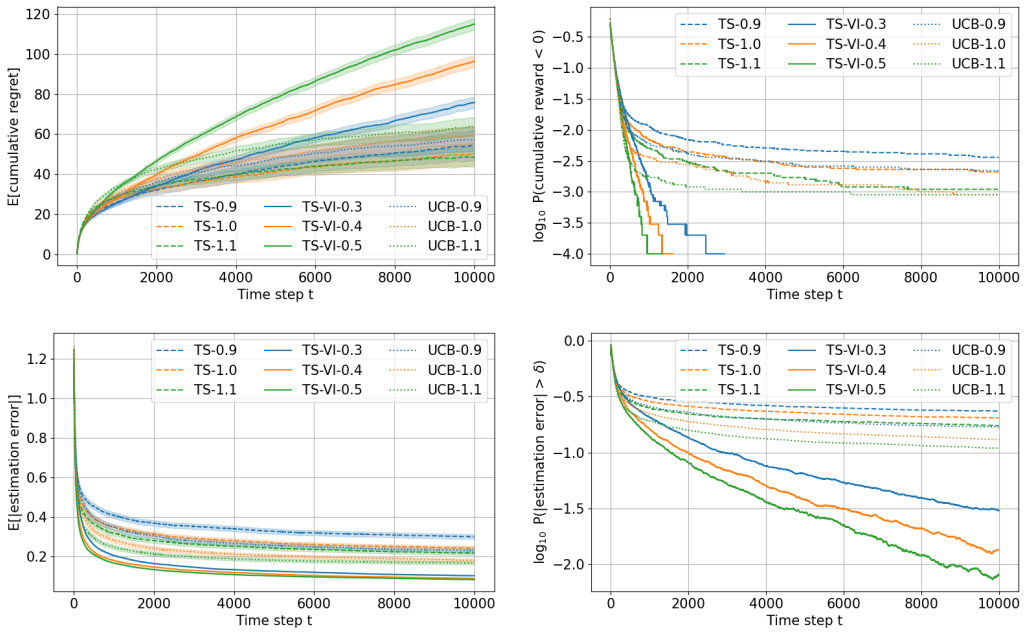

Figure 2: Results for the mis-specified environment with known variances

Robustness. TS-VI demonstrates stable performance across different hyperparameter settings and environmental conditions, both in terms of efficiency and safety, validating the theoretical results presented in Sections 3 and 4. In contrast, while TS reliably achieves low expected regret within an experiment, it consistently underperforms on the other three metrics: estimation accuracy, tail safety, and robustness. Notably, the safety performance of either TS or UCB deteriorates significantly in the presence of environment mis-specifications.

Finally, we would like to provide an additional remark on the performance of TS and UCB under under-specified ($\sigma_0 = 0.9\sigma$) and over-specified ($\sigma_0 = 1.1\sigma$) variances.

- For under-specified policies, while they perform well in terms of expected regret, they suffer significantly in terms of tail safety and mean estimation accuracy.
- For over-specified policies, they show slightly worse average regret but improved tail behavior and mean estimation. However, the tail decay remains slow as T grows, suggesting that while empirical gains are possible via over-specification, achieving intrinsic improvement in safety necessitates explicit variance inflation. We also discuss this point via a distribution visualization in the supplementary material.

6 Conclusion

In this work, we propose TS-VI, a modified version of Thompson Sampling in which the sampling posterior variance is inflated by an adaptive factor. We show that TS-VI achieves efficiency, safety, and robustness in the worst-case setting, both for within-experiment regret control and post-experiment decision quality. These theoretical findings are validated through simulations, and we further extend the policy design to handle heterogeneous, arm-specific variances via Gamma-Normal Bayesian updates. Several promising directions remain for future investigation, and we highlight three below:

Instance-dependent analysis. While this work focuses primarily on worst-case performance, it would be insightful to conduct an instance-dependent analysis. It remains unclear whether the safety and robustness guarantees established here continue to hold under such a perspective. An instance-dependent view may offer a more nuanced understanding of how efficiency, safety, and robustness interact and trade-off with each other in different environments.

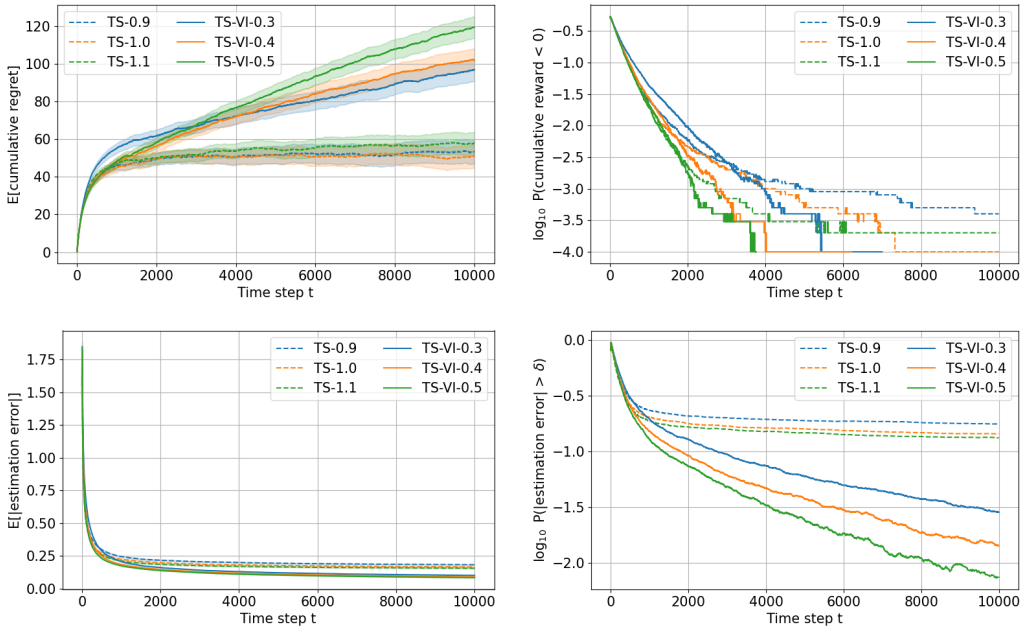

Figure 3: Results for the mis-specified environment with unknown variances

Asymptotic behavior. Our theoretical results provide finite-time error bounds for various within- and post-experiment objectives. However, questions remain about the asymptotic behavior of TS-VI — for example, what is the limiting proportion of times each arm is pulled? A deeper analysis of the long-run behavior could yield sharper characterizations of the policy and offer principled guidance for hyperparameter tuning.

More complex models. This work focuses on the standard stochastic multi-armed bandit setting. Extending our approach to more complex frameworks — such as linear bandits, nonparametric models, and Markov decision processes — is both challenging and worthwhile. It is also important to investigate how the presence of heavy-tailed reward distributions (beyond sub-exponential) affects the performance guarantees and policy behavior.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[2] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[3] S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

[4] K. Ashutosh, J. Nair, A. Kagrecha, and K. Jagannathan. Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2021.

[5] J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

[6] J.-Y. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[7] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International Conference on Algorithmic Learning Theory*, 2009.

[8] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

[9] A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[10] O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[11] D. Eckles and M. Kaptein. Bootstrap thompson sampling and sequential decision problems in the behavioral sciences. *Sage Open*, 9(2):2158244019851675, 2019.

[12] A. Erraqabi, A. Lazaric, M. Valko, E. Brunskill, and Y.-E. Liu. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR, 2017.

[13] E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pages 255–270. Springer, 2002.

[14] L. Fan and P. W. Glynn. The fragility of optimized bandit algorithms. *Operations Research*, 2024.

[15] A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[16] X. Huo and F. Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

[17] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[18] T. Jin, P. Xu, J. Shi, X. Xiao, and Q. Gu. Mots: Minimax optimal thompson sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

[19] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[20] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[21] B. Liang and I. Bojinov. An experimental design for anytime-valid causal inference on multi-armed bandits. *arXiv preprint arXiv:2311.05794*, 2023.

[22] C. Qin and D. Russo. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024.

[23] D. Russo. Simple bayesian algorithms for best arm identification. In *Conference on learning theory*, pages 1417–1418. PMLR, 2016.

[24] D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[25] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[26] A. Salomon and J.-Y. Audibert. Deviations of stochastic bandit regret. In *International Conference on Algorithmic Learning Theory*, pages 159–173. Springer, 2011.

[27] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*, 197:116669, 2022.

[28] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[29] D. Simchi-Levi and C. Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Available at SSRN 4224969*, 2022.

[30] D. Simchi-Levi, Z. Zheng, and F. Zhu. A simple and optimal policy design for online learning with safety against heavy-tailed risk. *Advances in Neural Information Processing Systems*, 35:33795–33805, 2022.

[31] D. Simchi-Levi, Z. Zheng, and F. Zhu. Stochastic multi-armed bandits: optimal trade-off among optimality, consistency, and tail risk. *Advances in Neural Information Processing Systems*, 36:35619–35630, 2023.

[32] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[33] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[34] I. Waudby-Smith, D. T. Arbour, R. Sinha, E. H. Kennedy, and A. Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics*, 2021.

[35] J. Yang, V. Y. Tan, and T. Jin. Best arm identification with minimal regret. *arXiv preprint arXiv:2409.18909*, 2024.

[36] Y. Zhao, C. J. Stephens, C. Szepesvari, and K.-S. Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims are correctly reflected via theory and experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Provided in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Provided in model, proof idea and supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Provided in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code provided in supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Provided in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper investigates tail distributions and runs 10000 samples for each environment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: N/A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: N/A

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for polishing paragraphs and speeding up experiments.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.