Sample-Mean Anchored Thompson Sampling for Offline-to-Online Learning with Distribution Shift

Bochao Li [1] **Yao Fu** [1] **Wei Chen** [2] **Fang Kong** [1]

Abstract

Offline-to-online learning aims to improve online decision-making by leveraging offline logged data. A central challenge in this setting is the distribution shift between offline and online environments. While some existing works attempt to leverage shifted offline data, they largely rely on UCB-type algorithms. Thompson sampling (TS) represents another canonical class of bandit algorithms, well known for its strong empirical performance and naturally suited to offline-to-online learning through its Bayesian formulation. However, unlike UCB indices, posterior samples in TS are not guaranteed to be optimistic with respect to the true arm means. This makes indices constructed from purely online and hybrid data difficult to compare and complicates their use. To address this issue, we propose sample-mean anchored TS (Anchor-TS), which introduces a novel median-based anchoring rule that defines the arm index as the median of an online posterior sample, a hybrid posterior sample, and the online sample mean. The median anchoring systematically corrects bias induced by distribution shift by mitigating over-estimation for suboptimal arms and under-estimation for optimal arms, while exploiting offline information to obtain more accurate estimates when the shift is small. We establish theoretical guarantees showing that the proposed algorithm safely leverages offline data to accelerate online learning, and quantifying how the degree of distribution shift and the size of offline data affect the resulting regret reduction. Extensive experiments demonstrate consistent improvements of our algorithm over baselines.

[1]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China [2]Microsoft Research Asia, Beijing, China. Correspondence to: Fang Kong <kongf@sustech.edu.cn>.

ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Stochastic multi-armed bandits (MAB) provide a fundamental framework for sequential decision-making under uncertainty (Lai & Robbins, 1985; Lattimore & Szepesvári, 2020), where a learning agent repeatedly selects actions to learn the unknown environment and maximize cumulative reward. The upper confidence bound (UCB) (Auer et al., 2002) and Thompson sampling (TS) (Agrawal & Goyal, 2017) are two canonical algorithm types to solving the problem: the former relies on optimism-driven confidence intervals, and the latter selects actions by sampling from posterior distributions. Despite their importance, classical bandit algorithms typically start from scratch. However, learning purely from online interactions can be costly or risky. Many applications provide access to offline logged data collected by historical policies, which motivates the offline-to-online learning setting that leverages such data to accelerate subsequent online learning (Nair et al., 2020; Lee et al., 2022; Shivaswamy & Joachims, 2012; Wagenmaker & Pacchiano, 2023).

Recently, there has been increasing interest in offline-to-online algorithms under distribution shift (Cheung & Lyu, 2024; Yin & Fang, 2025; He et al., 2025; Qu et al., 2025), where the offline reward distribution differs from the online environment due to system updates, non-stationarity, or sim-to-real gaps. A recurring insight in this line of work is that achieving performance no worse than purely online learning requires some form of prior knowledge about the distribution shift. Existing approaches mainly adopt UCB-type algorithms. A representative strategy is to construct two UCBs: a hybrid UCB that incorporates offline data together with a bias correction term, and a purely online UCB based solely on online observations (Cheung & Lyu, 2024). Actions are selected according to the minimum of the two, which guarantees regret no worse than that of pure online UCB.

Thompson sampling (TS) represents another canonical class of bandit algorithms and has been extensively studied across a wide range of online learning settings (Agrawal & Goyal, 2012; Russo et al., 2018). It is well known for its strong empirical performance, often outperforming UCB-based methods in practice, and its Bayesian formulation makes it particularly appealing for offline-to-online learning as

unbiased offline data can be naturally incorporated into the prior distribution (Oetomo et al., 2023; Agnihotri et al., 2024).

Despite these advantages, the posterior-sampling nature of TS makes it fundamentally more challenging under distribution shift. In UCB-based methods, the index is optimistic, and regret can be attributed to inaccurate estimates of suboptimal arms. By contrast, TS relies on posterior samples that may fall on either side of the true arm means, and its regret depends jointly on the accuracy of both the optimal and suboptimal arms. This principal difference has important implications for incorporating offline data. For UCB-based methods, conservatively taking the minimum of a purely online index and a hybrid index provides a safe guard against distribution shift. For TS, however, no analogous comparison rule exists: taking the minimum risks underestimating the optimal arm, while taking the maximum may overestimate suboptimal arms. This lack of a principled mechanism for comparing posterior samples makes it nontrivial to exploit biased offline data for TS.

To address this challenge, we propose sample-mean anchored Thompson sampling (Anchor-TS), a modified TS algorithm based on a median-of-indices aggregation scheme. At each round, Anchor-TS assigns to each arm an index defined as the median of three statistics: (i) an online posterior sample, (ii) a hybrid posterior sample that integrates offline and online data, and (iii) the online sample mean, which serves as an unbiased anchor. This median-based aggregation induces an inherent robustness–efficiency trade-off. When the offline data are reliable, the hybrid posterior sample typically exhibits reduced variance and remains close to the anchor, thereby accelerating learning. Conversely, under severe bias in the offline component, the hybrid sample tends to behave as an outlier and is filtered out by the median, so that the resulting index is primarily driven by online observations. Furthermore, to prevent the underestimation of the optimal arms' online sample mean from being amplified under median aggregation when its offline reward is smaller, we additionally introduce a right-hand shift to the hybrid posterior distribution.

Our analysis develops a new regret decomposition tailored to median-based TS. In particular, we analyze the behavior of posterior samples for both suboptimal and optimal arms relative to the online sample mean, which serves as an anchor, and characterize their joint contributions to regret accordingly. This enables us to bound the leading regret terms in a way that captures the more favorable behavior between online TS and hybrid TS. Let $\alpha_i := (1 - 3\omega_i/\Delta_i)+$. Specifically, we obtain a regret upper bound

$$O\left(\sum_{i\neq 1}\Delta_i\left[(\log T/\Delta_i^2 - N_i\alpha_i)_+ + (\log T/\Delta_i^2 - N_1)_+\right]\right).$$

where T is online horizon, K is the arm set size, arm 1 is the optimal arm, Δ_i is arm i's sub-optimality gap compared with 1, ω_i is an effective discrepancy related to i's offline and online expected reward, N_i is the offline sample size of i, $(\cdot)_+$ represents $\max\{0, \cdot\}$.

The above regret guarantee formalizes the robustness–efficiency trade-off under distribution shift: it is no worse than the pure online TS (Agrawal & Goyal, 2017), and strictly improves when the distribution shift $\{\omega_i\}_{i\in[K]}$ is mild. Compared to existing UCB-based algorithms (Cheung & Lyu, 2024), the regret reduction attributable to suboptimal arms estimations is of similar order. Crucially, our analysis further reveals an additional source of improvement that is unique to TS: offline data also reduces regret by accelerating concentration on the optimal arm, as reflected in the dependence on N_1. Such an advantage is particularly relevant in practice, where offline data are often collected by prior learning policies or expert behavior and therefore tend to contain a large number of samples from the optimal arm. In such cases, Anchor-TS can effectively exploit this abundance of optimal-arm data, whereas UCB-based methods are unable to benefit in the same way. Extensive empirical results further validate our theory. Across all considered settings, Anchor-TS consistently and substantially outperforms UCB-based algorithms. The performance gap becomes even more pronounced when more offline data are on the optimal arm. Importantly, even in the unbiased and pure online settings, the sample-mean–based mechanism helps reduce the excessive exploration caused by the high variance of a single posterior sample in vanilla TS. As a result, it achieves better empirical performance than the corresponding TS baselines in these settings.

2. Related Work

Thompson sampling (TS). TS is a canonical Bayesian framework for stochastic MAB. It has been widely adopted since the seminal work of (Thompson, 1933), but the establishment of convergent regrets lagged for decades (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Agrawal & Goyal, 2017; Jin et al., 2021; 2023). In many problem settings, although UCB-based algorithms have been applied, researchers continue to devote significant effort to TS-type algorithms (Agrawal & Goyal, 2013; Komiyama et al., 2015; Wang & Chen, 2018; Verstraeten et al., 2020; Zhang et al., 2021), driven by their superior empirical performance and ease of implementation (Granmo, 2010; Scott, 2010; Chapelle & Li, 2011). More recently, TS-type algorithms are also proposed for offline-to-online learning problems with offline data naturally encoded in the prior (Oetomo et al., 2023; Agnihotri et al., 2024). These approaches typically do not account for distribution shift between offline and online environments.

Offline-to-online Learning. Offline-to-online learning aims to accelerate online decision-making by leveraging pre-collected logged data. This paradigm has been extensively explored empirically in both bandit and reinforcement learning (RL) settings (Lee et al., 2022; Nair et al., 2020; Ball et al., 2023; Yu & Zhang, 2023; Nakamoto et al., 2023; Xia et al., 2024), and has also motivated a growing body of theoretical work establishing performance guarantees.

Early theoretical studies primarily focused on the unbiased setting, where offline and online reward distributions coincide. In the bandit setting, Shivaswamy & Joachims (2012) show that historical data can yield constant regret, a result later extended to contextual bandits through informative priors (Oetomo et al., 2023) and meta-algorithmic approaches (Banerjee et al., 2022). More recently, Sentenac et al. (2025) analyze the trade-off between optimism in online learning and pessimism in offline learning in the offline-to-online context. For RL, theoretical frameworks have expanded from tabular Markov decision processes (Xie et al., 2021) to linear models (Wagenmaker & Pacchiano, 2023; Tan & Xu, 2024; Huang et al., 2025) and further to general function approximation (Song et al., 2023), with a common focus on relaxing offline data coverage assumptions.

In practice, offline data often originate from heterogeneous or outdated sources, making it essential to achieve both efficiency gains from informative offline data and robustness to distribution shift. Fundamental lower bounds establish that without a priori knowledge of the distribution shift, no policy can uniformly outperform purely online algorithms (Zhang et al., 2019; Cheung & Lyu, 2024; Zhang et al., 2025; Qu et al., 2025). Consequently, existing works typically incorporate some shift information into the decision process. Such information is encoded through lower bounds (Qu et al., 2025) or upper bounds on the shift (Cheung & Lyu, 2024; Ahn et al., 2025). Our work follows the last line of works. A representative strategy in this line is to construct a hybrid UCB index by combining offline data with a known upper bound on the bias, and to select actions according to the minimum of this hybrid index and a purely online UCB (Cheung & Lyu, 2024). This strategy is applied to a variety of settings, including best arm identification (BAI) (Yang et al., 2025), heterogeneous feedback structures (He et al., 2025), auxiliary rewards (Yin & Fang, 2025), combinatorial MAB (Zhou et al., 2025), and linear bandits (Zhang et al., 2025). To the best of our knowledge, TS-type algorithms for offline-to-online learning under distribution shift is still open.

TS with misspecified priors. Our work is also related to studies on misspecified priors in TS. Liu & Li (2016); Simchowitz et al. (2021) show that when TS is initialized with strongly biased priors, where the prior assigns very little probability to the true model, it can suffer linear regret and perform much worse than methods with uninformative priors. These results demonstrate the negative impact of using an incorrect prior. Compared with this line of work, our setting explicitly models the prior as being induced by offline data. Our setting requires not only robustness to biased priors, but also the ability to achieve regret improvement when offline data are informative.

3. Preliminaries

We begin with the classical stochastic multi-armed bandit problem. An agent interacts with an environment over a time horizon T by repeatedly selecting arms from a finite set $[K] = \{1, \dots, K\}$. Each arm $i \in [K]$ is associated with an unknown reward distribution $P_i^{(\mathrm{on})}$ supported on $[0, 1]$[1] and unknown mean $\mu_i^{(\mathrm{on})}$. At each round $t = 1, 2, \dots, T$, the agent selects an arm $A(t) \in [K]$ and observes a reward $R_{A(t)}(t)$ drawn independently from the corresponding online distribution $P_{A(t)}^{(\mathrm{on})}$. Without loss of generality, let arm $1 \in \arg\max_{j \in [K]} \mu_j^{(\mathrm{on})}$ denote the unique[2] optimal arm. For each suboptimal arm $i \neq 1$, define the sub-optimality gap $\Delta_i = \mu_1^{(\mathrm{on})} - \mu_i^{(\mathrm{on})}$.

In this work, we study a stochastic bandit setting augmented with *offline data*. In addition to the online interaction described above, each arm $i \in [K]$ is associated with an offline dataset $S_i = \{X_{i,1}, \dots, X_{i,N_i}\}$ of size N_i. The samples in S_i are assumed to be i.i.d. draws from an offline distribution $P_i^{(\mathrm{off})}$ supported on $[0, 1]$, with unknown mean $\mu_i^{(\mathrm{off})} = \mathbb{E}_{X_i \sim P_i^{(\mathrm{off})}}[X_i]$.

Importantly, for any arm i, the offline and online reward distributions are not assumed to coincide. To explicitly model the potential distribution shift between them, we assume that for each arm i, there exists a known upper bound $V_i \geq 0$ such that $|\mu_i^{(\mathrm{off})} - \mu_i^{(\mathrm{on})}| \leq V_i$. This assumption provides a priori control over the magnitude of the distribution shift and is information-theoretically necessary. Existing works establish fundamental lower bounds that no policy can uniformly outperform purely online algorithms without any prior knowledge (Zhang et al., 2019; Cheung & Lyu, 2024; Zhang et al., 2025; Qu et al., 2025) and such upper bound is widely adopted in previous works (Cheung & Lyu, 2024; Zhang et al., 2025; He et al., 2025; Ahn et al., 2025; Yin & Fang, 2025).

The performance of a policy π is measured by its cumulative regret, which quantifies the expected loss relative to always

[1] Our analysis extends to sub-Gaussian reward distributions by replacing Hoeffding-type concentration bounds with their sub-Gaussian counterparts. The regret order remains unchanged.

[2] The uniqueness assumption is only for the convenience of the analysis. The setting with multiple optimal arms can only decrease the regret as shown in (Agrawal & Goyal, 2012).

pulling the optimal arm:

$$\mathrm{Reg}_{\pi}(T) = T\mu_1^{(\mathrm{on})} - \mathbb{E}\left[\sum_{t=1}^{T} R_{A(t)}(t)\right] = \sum_{i\neq 1} \Delta_i \, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A(t)=i\}\right]. \quad (1)$$

where the expectation is taken over all sources of randomness, including the internal randomness of π, the offline data draw from $P^{(\mathrm{off})}$, and the online rewards generated under $P^{(\mathrm{on})}$.

Useful Notations. For each arm $i \in [K]$, let $T_i(t)$ denote the number of times arm i is pulled during rounds $1, \ldots, t-1$, and define $n_i(t) = T_i(t) + N_i$ as the total number of samples available for arm i, combining online and offline data. Let $\hat{\mu}_i^{(\mathrm{off})}$ denote the sample mean of the offline dataset for arm i, and let $\hat{\mu}_i^{(\mathrm{on})}(t)$ be the sample mean computed from the online rewards observed up to round $t-1$. We further define $\hat{\mu}_i^{(\mathrm{hyb})}(t)$ as the hybrid sample mean that aggregates both offline samples and online observations for arm i.

4. Algorithm

In the offline-to-online setting, it is natural to maintain two types of estimators: a purely online estimator constructed solely from online observations, and a hybrid estimator that aggregates offline data with online observations. Due to the distribution shift, the hybrid estimators may be biased and can mislead the learning process if used indiscriminately. The key algorithmic challenge is therefore how to balance the use of the hybrid estimator against the purely online estimator. In this section, we present our sample-mean anchored TS (Anchor-TS, Algorithm 1), an efficient and robust mechanism for combining online and hybrid estimators under the TS framework.

For each arm $i \in [K]$, the algorithm maintains an online posterior that is updated exclusively based on rewards observed during online interaction, which is the same as the traditional TS algorithm for the purely online setting (Agrawal & Goyal, 2017). At time t, this posterior is characterized by the online sample mean $\hat{\mu}_i^{(\mathrm{on})}(t)$ and an associated variance $\sigma_{i,\mathrm{on}}^2(t)$ (Line 12). An online posterior sample $\theta_i^{(\mathrm{on})}(t)$ is drawn from the corresponding Gaussian distribution (Line 5). Because it relies solely on online data, this posterior remains unbiased with respect to the true arm mean, although it may exhibit high variance in the early stages of learning.

In parallel, Anchor-TS constructs a hybrid posterior that integrates offline empirical information with online observations. Specifically, the hybrid posterior mean $\hat{\mu}_i^{(\mathrm{hyb})}(t)$ is obtained by combining the offline sample mean $\hat{\mu}_i^{(\mathrm{off})}$

Algorithm 1 Sample-Mean Anchored Thompson sampling (Anchor-TS)

1: **Input:** Arm set $[K]$; offline sample mean $\hat{\mu}_i^{(\mathrm{off})}$ and sample size N_i, bias bound V_i, $\forall i \in [K]$
2: Initialization: $T_i(1) \leftarrow 0$, $\hat{\mu}_i^{(\mathrm{on})}(1) \leftarrow 0$, $\hat{\sigma}_{i,\mathrm{on}}^2(t) \leftarrow 1$, $\hat{\mu}_i^{(\mathrm{hyb})}(1) \leftarrow \hat{\mu}_i^{(\mathrm{off})}$, $\hat{\sigma}_{i,\mathrm{hyb}}^2(1) \leftarrow 1/(N_i+1)$, $Z_{i,1} \leftarrow V_i$, $\forall i \in [K]$
3: **for** $t = 1, \ldots$ **do**
4: **for** each arm $i \in [K]$ **do**
5: Sample $\theta_i^{(\mathrm{on})}(t)$ from $\mathcal{N}\left(\hat{\mu}_i^{(\mathrm{on})}(t),\ \hat{\sigma}_{i,\mathrm{on}}^2(t)\right)$
6: Sample $\theta_i^{(\mathrm{hyb})}(t)$ from $\mathcal{N}\left(\hat{\mu}_i^{(\mathrm{hyb})}(t) + Z_{i,t},\ \hat{\sigma}_{i,\mathrm{hyb}}^2(t)\right)$
7: $\hat{\theta}_i(t) \leftarrow \mathrm{median}\{\hat{\mu}_i^{(\mathrm{on})}(t),\ \theta_i^{(\mathrm{on})}(t),\ \theta_i^{(\mathrm{hyb})}(t)\}$
8: **end for**
9: Select arm $A(t) \leftarrow \arg\max_{i\in[K]} \hat{\theta}_i(t)$ and observe reward $R_{A(t)}(t)$
10: **// Update online posterior of** $A(t)$
11: $T_{A(t)}(t+1) \leftarrow T_{A(t)}(t) + 1$
12: $\hat{\mu}_{A(t)}^{(\mathrm{on})}(t+1) \leftarrow \frac{T_{A(t)}(t)\cdot\hat{\mu}_{A(t)}^{(\mathrm{on})}(t) + R_{A(t)}(t)}{T_{A(t)}(t+1)+1}$, $\hat{\sigma}_{A(t),\mathrm{on}}^2(t+1) \leftarrow \frac{1}{T_{A(t)}(t+1)+1}$
13: **// Update hybrid posterior of** $A(t)$
14: $\hat{\mu}_{A(t)}^{(\mathrm{hyb})}(t+1) \leftarrow \frac{T_{A(t)}(t+1)\cdot\hat{\mu}_{A(t)}^{(\mathrm{on})}(t+1) + N_{A(t)}\cdot\hat{\mu}_{A(t)}^{(\mathrm{off})}}{T_{A(t)}(t+1)+N_{A(t)}+1}$
15: $\hat{\sigma}_{A(t),\mathrm{hyb}}^2(t+1) \leftarrow \frac{1}{N_{A(t)}(t+1)+T_{A(t)}+1}$, $Z_{A(t),t+1} \leftarrow \frac{N_{A(t)}V_{A(t)}}{T_{A(t)}(t+1)+N_{A(t)}}$
16: **for** $i \neq A(t)$ **do**
17: Update $\hat{\mu}_i^{(\mathrm{on})}(t+1) \leftarrow \hat{\mu}_i^{(\mathrm{on})}(t)$, $\hat{\mu}_i^{(\mathrm{hyb})}(t+1) \leftarrow \hat{\mu}_i^{(\mathrm{hyb})}(t)$, $\hat{\sigma}_{i,\mathrm{on}}^2(t+1) \leftarrow \hat{\sigma}_{i,\mathrm{on}}^2(t)$, $\hat{\sigma}_{i,\mathrm{hyb}}^2(t+1) \leftarrow \hat{\sigma}_{i,\mathrm{hyb}}^2(t)$, $Z_{i,t+1} \leftarrow Z_{i,t}$
18: **end for**
19: **end for**

and the current online sample mean $\hat{\mu}_i^{(\mathrm{on})}(t)$, weighted according to their respective sample sizes (Line 14). The corresponding variance $\hat{\sigma}_{i,\mathrm{hyb}}^2(t)$ reflects the increased effective sample size enabled by incorporating offline data (Line 15). Notably, to encourage sufficient exploration of the optimal arm 1 and reduce regret, we apply a rightward shift of magnitude $Z_{i,t}$ to the hybrid posterior distribution of each arm. A detailed discussion of the intuition behind this design choice will be provided in the end of this section. A hybrid posterior sample $\theta_i^{(\mathrm{hyb})}(t)$ is then drawn from the resulting Gaussian distribution (Line 6).

At each time step, Anchor-TS computes three indices for every arm: the online sample mean $\hat{\mu}_i^{(\mathrm{on})}(t)$, an online posterior sample $\theta_i^{(\mathrm{on})}(t)$, and a hybrid posterior sample $\theta_i^{(\mathrm{hyb})}(t)$. The arm score is obtained by taking the *median* of these three quantities (Line 7). And Anchor-TS selects the arm

with the highest score in each round (Line 9).

Intuition behind the median aggregation of three indices. Intuitively, the role of median aggregation is to adaptively select the index that is closest to the true mean rather than to favor large or small posterior realizations like UCB-type algorithms (Cheung & Lyu, 2024). This distinction is intrinsic to TS: posterior samples in TS are neither guaranteed to be optimistic nor pessimistic. Naively selecting the minimum or maximum of multiple indices would either underestimating the optimal arm or overestimating suboptimal arms, leading to uncontrolled regret. The sample mean $\hat{\mu}^{(\mathrm{on})}$, estimated from online observations, therefore serves as a natural stabilizing anchor as it concentrates rapidly around the true arm mean. By taking the median of $\hat{\mu}^{(\mathrm{on})}$, an online-only and a hybrid posterior sample, Anchor-TS selects the index that is most consistent with this anchor.

When the bias is small, the hybrid posterior, benefiting from a larger effective sample size, tends to concentrate more tightly around the true mean than the online-only posterior. In this regime, the median naturally favors the hybrid posterior sample, allowing Anchor-TS to exploit offline data for faster learning. Conversely, when the offline bias is large, the hybrid posterior may deviate from the true mean, while the online-only posterior remains centered around the unbiased online signal. In this case, the median suppresses the biased hybrid sample and tends to select the online sample instead, yielding robustness against offline-induced distortion. Through this adaptive selection mechanism, Anchor-TS automatically interpolates between efficiency and robustness, leveraging offline data when it is reliable and reverting to online evidence when offline bias is substantial.

Intuition behind the right-hand shift on the hybrid posterior distribution. Recall that for TS-type algorithms, the regret can still accumulate when the estimates of suboptimal arms are accurate, as long as the optimal arm is poorly estimated. Specific to the offline-to-online setting, if the offline data underestimate the optimal arm, i.e., $\mu_1^{(\mathrm{off})} < \mu_1^{(\mathrm{on})}$, and the online sample mean of arm 1 happens to be pessimistic, arm 1 may receive too little posterior probability due to the three-index voting mechanism. In this case, it is difficult to correct the estimation error of arm 1 through additional online observations. As a result, the persistent underestimation of arm 1 amplifies the number of selections of suboptimal arms, leading to increased regret despite their accurate estimation.

Our right-shift operation counteracts this effect by preventing the hybrid posterior distribution from underestimating the online mean of arm 1. This guarantees adequate exploration of the optimal arm and prevents regret from being dominated by prolonged under-exploration of the optimal arm.

Since the identity of the optimal arm is unknown, the right-hand shift is applied uniformly across arms. Importantly, this shift does not compromise the benefit of offline data for suboptimal arms: although overestimation by the hybrid posterior may initially increase their selection frequency, the influence of offline information quickly diminishes as online observations accumulate, and the estimate becomes dominated by online samples through median aggregation.

5. Theoretical Results and Discussions

The following theorem summarizes the regret bound of our algorithm:

Theorem 5.1. *The cumulative regret of Algorithm 1 can be bounded by*

$$\mathrm{Reg}(T) \le O\Bigg(\sum_{i\neq 1} \Delta_i \Bigg(\Bigg(\frac{C_1 \log T}{\Delta_i^2} - N_i \Big(1 - \tfrac{3\omega_i}{\Delta_i}\Big)_+ \Bigg)_+ + \Bigg(\frac{C_2 \log T}{\Delta_i^2} - N_1 \Bigg)_+ + \frac{C_3}{\Delta_i^2} \Bigg) \Bigg).$$

where C_1, C_2, C_3 are constants, $(\cdot)_+$ represents $\max\{\cdot, 0\}$, $\omega_i := V_i + \mu_i^{(\mathrm{off})} - \mu_i^{(\mathrm{on})}$ represents the effective discrepancy by adding V_i in the hybrid posterior distribution.

Due to the space limit, the complete proof of Theorem 5.1 is deferred to Appendix A. In the following, we first provide some discussions and then show the proof sketch.

Intuition of the regret upper bound. This upper bound consists of three terms. The first term arises from inaccurate estimation of suboptimal arms, the second from inaccurate estimation of the optimal arm, and the third is a constant term required for convergence of the online sample mean.

This upper bound can be interpreted as the regret in a purely online setting minus the benefit provided by the offline data. When no offline data are available, the regret reduces to the standard purely online regret (Agrawal & Goyal, 2017). When the distribution shift $\mu_i^{(\mathrm{on})} - \mu_i^{(\mathrm{off})}$ is zero, the offline samples for both the sub-optimal arms and the optimal arm can be viewed as offsetting a portion of the regret. In particular, if the sub-optimal arms have sufficiently many offline samples, their means can be accurately estimated from the outset, rendering the corresponding regret term constant. Similarly, if the optimal arm has sufficiently many offline samples, the regret arising from inaccurate estimation of the optimal arm is likewise reduced to a constant.

Regarding the bias, for a sub-optimal arm i, if the offline mean is smaller than the online mean and V is a tight bound, i.e., $V = \mu_i^{(\mathrm{on})} - \mu_i^{(\mathrm{off})} > 0$, then $\omega_i = 0$ and the offline improvement scales linearly with N_i. However, if the offline mean $\mu_i^{(\mathrm{off})}$ is larger, the hybrid samples may

overestimate the true online mean $\mu_i^{(\text{on})}$. In this case, the contribution of the offline data is attenuated by a discount factor $(1 - 3\omega_i/\Delta_i)$, which decreases as ω_i increases, resulting in diminishing effective information from the offline samples. When $3\omega_i > \Delta_i$, the hybrid samples can mislead the identification of the optimal arm, and as the result the offline benefit vanishes and the regret reduces to that of the purely online setting. In contrast, for the optimal arm (arm 1), adding the correction term V to the hybrid samples always leads to an overestimation of the true mean $\mu_1^{(\text{on})}$. Such overestimation consistently favors the identification of the optimal arm. Consequently, the contribution of the offline data for the optimal arm is not subject to any discount factor.

Comparison with Cheung & Lyu (2024). Cheung & Lyu (2024) study the same offline-to-online setting as ours, but focus on a UCB-based approach. They derive an upper bound on the sub-optimal i's selection time $O(C \log T/\Delta_i^2 - N_i \cdot \max\{1 - \omega_i/\Delta_i, 0\}^2)_+$. The first term in our Theorem 5.1, which captures the regret due to inaccurate estimation of suboptimal arms, is of the same order as this bound. The main difference lies in how the size of the offline dataset N_i is discounted. Specifically, their bound uses the discount factor $(1 - \omega_i/\Delta_i)^2$ whereas ours involves the factor $(1 - 3\omega_i/\Delta_i)$. This constant coefficient discrepancy primarily stems from differences in the analysis techniques. In TS, posterior samples may be either optimistic or pessimistic, which requires partitioning the gap between μ_1 and μ_i into three regions, leading to the $\Delta_i/3$ factor in our formula. Importantly, our analysis avoids an additional squaring of the discount factor.

Compared with this result, our Theorem 5.1 further reveals an additional improvement arising from the offline data N_1 on the optimal arm, as reflected in the second term of our bound. This improvement stems from the property that TS incurs regret due to inaccurate estimation of both suboptimal and optimal arms. When offline data are available for arm 1, the estimation of the optimal arm becomes more accurate, thereby reducing this source of regret. Such a scenario is common in practical applications, since offline data are typically collected using expert or near-optimal policies, and therefore observations of the optimal arm are often abundant. Our Anchor-TS algorithm is able to exploit this advantage, whereas UCB-based methods do not benefit from offline data on the optimal arm in the same way.

Another aspect worth discussion is the right-shift operation applied to the hybrid samples. Although both the MINUCB algorithm and our approach incorporate a bias-correction term into the hybrid index, the underlying motivations are fundamentally different. In MINUCB, the correction term is introduced to preserve the optimism of the hybrid UCB index, ensuring that it upper-bounds the true reward. In contrast, TS does not intrinsically rely on optimism. The primary purpose of introducing the right-shift in our method is to encourage exploration of arm 1, thereby mitigating the regret caused by underestimation of this arm. While omitting this correction term would in fact improve the first regret term from over-estimation of sub-optimal arms, it would significantly complicate the analysis of $\hat{\mu}_1^{(\text{on})}$. In particular, inaccuracies in estimating $\hat{\mu}_1^{(\text{on})}$ would introduce regret terms that grow exponentially with N_1, since underestimation of the hybrid sample $\theta_1^{(\text{hyb})}$ would make arm 1 increasingly unlikely to be selected, leaving little opportunity for correction.

Reduction to the pure online setting and the role of median aggregation. In the pure online setting, where no offline data are available, the hybrid posterior distribution coincides with the online posterior distribution. In this case, Anchor-TS can be interpreted as sampling two online indices, $\theta_{i,1}^{(\text{on})}$, $\theta_{i,2}^{(\text{on})}$, and selecting the median among these two samples and the online sample mean $\hat{\mu}_i^{(\text{on})}$. We analyze this setting separately and show that this modification preserves the classical instance-dependent regret bound, while improving the leading regret order by a factor of $1/2$. We further provide empirical results that illustrate this advantage. The corresponding theoretical analysis and experimental results are presented in Appendix B. This observation may, to some extent, reflect a similar idea in Jin et al. (2023), who show that TS with less exploration can perform better. But we leverage different ideas to reduce exploration: they divert a fixed probability mass ϵ from vanilla TS to selecting the sample mean, whereas our method incorporates the sample mean in a more adaptive manner through median aggregation.

6. Proof Sketch

In this section, we present a proof sketch for Theorem 5.1 and highlight the key ideas of our analysis. Our technical contributions operate at two levels: at a high level, we tightly couple the median-based algorithmic design with the regret decomposition, enabling the regret bound to adapt to the more informative of the online and hybrid posterior distributions; at a more technical level, we develop a refined probabilistic analysis that decouples multiple random indices appearing within probability and expectation operators, allowing their individual regret contributions to be explicitly controlled.

Standard TS analyses consider regret contributions from a single posterior distribution. In contrast, our setting involves two posterior sources, where a naive extension yields additive regret bounds and fails to exploit the more informative one. To address this, we use the online sample

mean as an anchor: its estimation errors contribute only a constant-order regret of $O(1/\Delta_i)$, allowing us to condition on accurate estimation and isolate the effects of the two posterior distributions.

For convenience, define $x_i := \mu_i^{(\text{on})} + \Delta_i/3$ and $y_i := \mu_i^{(\text{on})} + 2\Delta_i/3$ as two thresholds, $E_i^{\mu(\text{on})}(t) := \{\hat{\mu}_i^{(\text{on})}(t) \leq x_i\}$ as the good event that arm i's online sample mean is accurate. Then,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A(t)=i\}\right] = \sum_{t=1}^{T} \Pr\big(A(t)=i,\ \neg E_i^{\mu(\text{on})}(t)\big) + \sum_{t=1}^{T} \Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t)\big).$$

The first term can be upper bounded by $O(1/\Delta_i^2)$ without relying on T. The second term representing regret from selecting i when its online sample mean is accurate can be decomposed as

$$\sum_{t=1}^{T} \Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t),\ \hat{\theta}_i(t) > y_i\big) + \sum_{t=1}^{T} \Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t),\ \hat{\theta}_i(t) \leq y_i\big). \quad (2)$$

The first term in (2) corresponds to the case where arm i's median index is over-estimated. Conditional on the good event $\{\hat{\mu}_i^{(\text{on})}(t) \leq x_i\}$, the event $\{\hat{\theta}_i(t) > y_i\}$ implies that $\{\theta_i^{(\text{on})}(t) > y_i\} \cap \{\theta_i^{(\text{hyb})}(t) > y_i\}$. The corresponding regret event is therefore the intersection of two unfavorable events, whose probability is strictly smaller than that induced by either posterior alone,

$$\sum_{t=1}^{T} \min\Big\{ \Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t),\ \theta_i^{(\text{on})}(t) > y_i\big), \Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t),\ \theta_i^{(\text{hyb})}(t) > y_i\big) \Big\}.$$

This yields the first term in Theorem 5.1 by taking the minimum of the corresponding regret contributions from the online and hybrid posterior distributions.

The second term in (2) can be transformed to the case where arm 1's median index is under-estimated. Dealing with this term is commonly recognized as a key challenge in TS analyses (Agrawal & Goyal, 2017; Jin et al., 2021; 2023), which is typically upper bounded by $\mathbb{E}[1/p - 1]$ where p is the conditional probability (given the history) that $\hat{\theta}_1 > y_i$. In vanilla TS, p is the tail probability of a single posterior sample $\hat{\theta}_1 := \theta_1^{(\text{on})}$, so one can directly convert $\mathbb{E}[1/p]$ into the expectation of a geometric hitting time related to the behavior of the index. In our setting, however, the arm index is the median of three random quantities with distinct concentration behaviors, which prevents a direct reduction to a one-dimensional geometric hitting-time argument as in vanilla TS.

To handle this term, we further condition on the behavior of $\hat{\mu}_1^{(\text{on})}$. Under the good event $\{\hat{\mu}_1^{(\text{on})}(t) > y_i\}$, it suffices that either the online or the hybrid posterior sample of arm 1 exceeds y_i for the optimal arm to be selected. This yields a union of favorable events, implying that

$$p \geq \max\{p^{(\text{on})},\, p^{(\text{hyb})}\},$$
$$\text{thus } \frac{1}{p} - 1 \leq \min\Big\{\frac{1}{p^{(\text{on})}} - 1,\ \frac{1}{p^{(\text{hyb})}} - 1\Big\}.$$

where $p^{(\text{on})}$ and $p^{(\text{hyb})}$ represent the conditional probability that $\{\theta_1^{(\text{on})}(t) > y_i\}$ and $\{\theta_1^{(\text{hyb})}(t) > y_i\}$, respectively. This leads to the second term in Theorem 5.1, where the right-hand shift of the hybrid posterior distribution ensures that $p^{(\text{hyb})} > p^{(\text{on})}$ when offline data arm 1 are available.

Another key difficulty arises when $\hat{\mu}_1^{(\text{on})}(t) < y_i$. In this regime, the median index can no longer exploit the more favorable of the two posterior samples, and the success probability p reduces to an intersection event involving both samples. This joint event can significantly shrink p and amplify regret. Moreover, since $\hat{\mu}_1^{(\text{on})}(t)$ and $\hat{\theta}_1(t)$ are statistically coupled, their contributions cannot be bounded separately and combined multiplicatively.

To address this challenge, we first apply Hölder's inequality to decompose the regret term into two components that can be handled separately. The first component involves the success probability $1/p$ corresponding to a joint hitting event of two posterior samples. Here, the right-hand shift of the hybrid posterior distribution plays a crucial role by preventing this term from becoming excessively large, ensuring that it remains bounded by a constant. The second component captures the probability of inaccurate estimation of the online sample mean, which decreases exponentially as the selection of arm 1. As a result, the combined contribution of these two components can be controlled by an $O(1/\Delta_i^2)$ upper bound.

7. Experiments

In this section, we compare our Anchor-TS with three classes of baselines: online-only methods including standard TS (Agrawal & Goyal, 2017) and standard UCB (Auer et al., 2002); naive offline-to-online methods including hybrid TS and hybrid UCB (Shivaswamy & Joachims, 2012) which use offline data for initialization while treating it as unbiased; and bias-aware offline-to-online method MIN-UCB (Cheung & Lyu, 2024). For each experiment, we report cumulative regret over $10k$ rounds, averaged over 50 runs with error bars showing the standard error. We consider a basic setting with $K = 10$ arms, where the optimal arm

has online mean reward 0.8 and all suboptimal arms have mean 0.5, yielding a gap of 0.3. Rewards are drawn from a Gaussian distribution with unit variance. The total number of offline samples is $2k$. To investigate the effect of offline data coverage, we consider three coverage patterns: uniform coverage across arms, coverage concentrated on the optimal arm 1 (with 80% of samples), and coverage concentrated on a suboptimal arm 2 (with 80% of samples). We further explore algorithm performance by varying parameters of this basic setting.

Unbiased offline data. We first consider a simple setting with zero bias under different sub-optimality gaps $\Delta \in \{0.3, 0.1\}$. The results are reported in Figure 1. There are two common observations: methods that leverage offline data consistently outperform purely online methods, and TS-type algorithms typically outperform their UCB-type counterparts.

In the easiest setting with uniformly distributed and sufficiently abundant offline data (Figure 1 left(a)), both hybrid UCB and hybrid TS achieve zero regret. In contrast, Anchor-TS incurs a small constant regret, as it deliberately requires a constant number of online samples to form a reliable anchor. In other settings, Anchor-TS consistently achieves the lowest regret and outperforms hybrid TS, despite both methods leveraging the same offline information. This gap comes from different exploration control: hybrid TS relies on a single high-variance posterior sample and may tend to over-explore, whereas Anchor-TS reduces exploration through sample-mean–based anchoring. A similar phenomenon was also observed in Jin et al. (2023), where incorporating the sample mean into the decision process leads to improved performance.

Compared with UCB-based algorithms, Anchor-TS consistently achieves substantial gains, particularly when offline data are concentrated on the optimal arm. This behavior aligns with our theoretical analysis, which predicts that the benefit of Anchor-TS scales with the amount of offline data available for the optimal arm, while UCB-based methods do not benefit in the same way. This advantage becomes even more pronounced in harder settings with a smaller gap where hybrid UCB and MINUCB even underperform pure online UCB (Figure 1 right (b)). This is because the UCB of the optimal arm is excessively reduced when incorporating offline data, forcing other arms to be pulled more often before their sub-optimality can be identified.

Biased offline data. We then evaluate the performance of algorithms under distribution shift between offline and online environments. We include both pure online methods and bias-aware approaches as baselines. To ensure that the offline data is sufficiently misleading, we consider a base setting in which $\mu_1^{(\text{off})} = 0.5$ and $\mu_i^{(\text{off})} = 0.6$ for $i \neq 1$, so

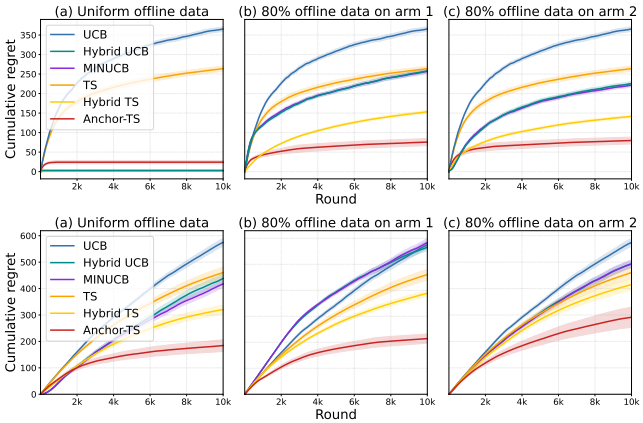

Figure 1. Cumulative regret in the unbiased setting under varying offline coverage regimes and suboptimality gaps (left: $\Delta = 0.3$; right: $\Delta = 0.1$).

that the optimal arm is underestimated in the offline data while the suboptimal arms are overestimated. Figure 2 summarizes the algorithms performances under biased offline data by varying key problem parameters, including the total offline sample size, the real bias level, the hyper-parameter V, and the arm set size. Unless otherwise specified, the hyperparameter V is set to the true bias magnitude for each arm. Across all settings, Anchor-TS consistently achieves the lowest regret, demonstrating robustness to offline bias and markedly outperforming UCB-based methods in its ability to effectively utilize offline information.

The top-left panel shows the algorithms performance under different offline sample sizes $\sum_i N_i$. Increasing offline data generally improves bias-aware methods, especially under uniform coverage.

The top-right panel studies the effect of the bias level δ. For ease of implementation, we fix the offline mean of sub-optimal arms and vary $\mu_1^{(\text{off})}$ by setting different choices of $\delta := \mu_1^{(\text{on})} - \mu_1^{(\text{off})}$. Larger δ corresponds to stronger underestimation of the optimal arm in the offline data. Nevertheless, Anchor-TS maintains a stable low-regret profile and consistently outperforms baselines.

The bottom-left panel examines the impact of hyper-parameter V. For MINUCB and Anchor-TS, we set V to the maximum of the true bias magnitude and the tested value. When V is small, tighter prior control allows Anchor-TS to exploit offline data more aggressively. As V increases, its performance degrades but never falls below pure online TS. Even when offline data are uninformative with large V, Anchor-TS retains a slight advantage due to sample-mean based exploration.

The bottom-right panel evaluates scalability with respect to the number of arms K. While regret increases with K

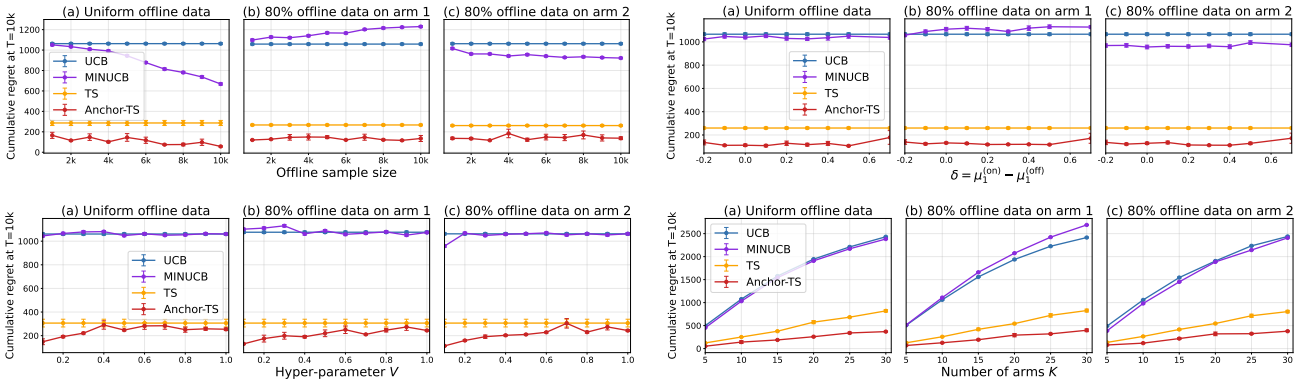

Figure 2. Cumulative regret under varying offline coverage regimes and problem parameters (top-left: total offline sample size; top-right: bias level δ; bottom-left: hyperparameter V; bottom-right: number of arms K).

for all methods, UCB-based algorithms grow much faster, whereas Anchor-TS scales more favorably by suppressing unnecessary exploration.

8. Conclusion

In this work, we study the TS-type algorithm for offline-to-online learning. Such algorithms face a fundamental challenge in assessing the reliability of the hybrid posterior distribution through comparison with the online posterior, stemming from the lack of inherent optimism in posterior samples. We address this challenge by introducing a novel median-based aggregation mechanism that combines the online sample mean, an online posterior sample, and a hybrid posterior sample, enabling adaptive exploitation of informative offline data. Our theoretical results demonstrate that our algorithm can effectively cope with bias in offline data, while achieving strictly better regret improvement when the offline data are sufficiently informative. Compared with UCB-type algorithms, our TS-based approach can further leverage abundant offline data on the optimal arm to obtain larger performance gains. Moreover, even in the pure online setting, the proposed median-index mechanism yields provable improvement in the leading constant of the regret bound compared with pure online TS. Extensive empirical evaluations further validate our theoretical findings, demonstrating that our algorithm consistently outperforms state-of-the-art baselines across various settings of data coverage and bias magnitudes.

Several promising directions merit further investigation. First, it would be valuable to establish gap-independent regret bounds that explicitly quantify the benefits of offline data. Second, extending the proposed anchor-based framework to contextual TS is an important direction for addressing high-dimensional, real-world decision-making problems.

Impact Statement

This paper develops theory and algorithms for offline-to-online bandit learning under distribution shift, with experiments only on synthetic instances. The work is methodological and we do not release datasets or models. We do not foresee direct dual-use risks from the theoretical results themselves.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. National Bureau of Standards, Washington, DC, 1964.

Agnihotri, A., Jain, R., Ramachandran, D., and Wen, Z. Online bandit learning with offline preference data for improved RLHF. *arXiv preprint arXiv:2406.09574*, 2024.

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

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

Ahn, H.-S., Zhang, M., and Zhang, Z. Online decisions with (biased) offline data. *Available at SSRN 5350921*, 2025.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Ball, P. J., Smith, L., Kostrikov, I., and Levine, S. Efficient online reinforcement learning with offline data. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Banerjee, S., Sinclair, S. R., Tambe, M., Xu, L., and Yu, C. L. Artificial replay: a meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 2249–2257, 2011.

Cheung, W. C. and Lyu, L. Leveraging (biased) information: multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 8286–8309. PMLR, 2024.

Granmo, O.-C. Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2): 207–234, 2010.

He, Q., Wang, M., Liu, X., Wang, Z., and Kong, F. Learning across the gap: Hybrid multi-armed bandits with heterogeneous offline and online data. In *The 39th Annual Conference on Neural Information Processing Systems*, 2025.

Huang, R., Li, D., Shi, C., Shen, C., and Yang, J. Augmenting online RL with offline data is all you need: A unified hybrid RL algorithm design and analysis. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Jin, T., Xu, P., Shi, J., Xiao, X., and Gu, Q. MOTS: Minimax optimal Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5074–5083. PMLR, 2021.

Jin, T., Yang, X., Xiao, X., and Xu, P. Thompson sampling with less exploration is fast and optimal. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15239–15261. PMLR, 2023.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2012.

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1152–1161. PMLR, 2015.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

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

Lee, S., Seo, Y., Lee, K., Abbeel, P., and Shin, J. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In *Conference on Robot Learning*, pp. 1702–1712. PMLR, 2022.

Liu, C.-Y. and Li, L. On the prior sensitivity of Thompson sampling. In *International Conference on Algorithmic Learning Theory*, pp. 321–336. Springer, 2016.

Nair, A., Gupta, A., Dalal, M., and Levine, S. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Nakamoto, M., Zhai, Y., Singh, A., Sobol Mark, M., Ma, Y., Finn, C., Kumar, A., and Levine, S. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 62244–62269, 2023.

Oetomo, B., Perera, R. M., Borovica-Gajić, R., and Rubinstein, B. I. Cutting to the chase with warm-start contextual bandits. *Knowledge and Information Systems*, 65(9): 3533–3565, 2023.

Qu, C., Shi, L., Panaganti, K., You, P., and Wierman, A. Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1054–1062. PMLR, 2025.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–99, 2018.

Scott, S. L. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Sentenac, F., Lee, I., and Szepesvári, C. Balancing optimism and pessimism in offline-to-online learning. *arXiv preprint arXiv:2502.08259*, 2025.

Shivaswamy, P. and Joachims, T. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pp. 1046–1054. PMLR, 2012.

Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D., Lykouris, T., Dudík, M., and Schapire, R. Bayesian decision-making under misspecified priors with applications to meta-learning. In *Proceedings of the 35th International*

Conference on Neural Information Processing Systems, pp. 26382–26394, 2021.

Song, Y., Zhou, Y., Sekhari, A., Bagnell, D., Krishnamurthy, A., and Sun, W. Hybrid RL: Using both offline and online data can make RL efficient. In *The 11th International Conference on Learning Representations*, 2023.

Tan, K. and Xu, Z. A natural extension to online algorithms for hybrid RL with limited coverage. *Reinforcement Learning Journal*, 1, 2024.

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

Verstraeten, T., Bargiacchi, E., Libin, P. J., Helsen, J., Roijers, D. M., and Nowé, A. Multi-agent Thompson sampling for bandit applications with sparse neighbourhood structures. *Scientific Reports*, 10(1):6728, 2020.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35300–35338. PMLR, 2023.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5114–5122. PMLR, 2018.

Xia, Y., Xie, Z., Yu, T., Zhao, C., and Li, S. Toward joint utilization of absolute and relative bandit feedback for conversational recommendation. *User Modeling and User-Adapted Interaction*, 34(5):1707–1744, 2024.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *35th Conference on Neural Information Processing Systems*, pp. 27395–27407, 2021.

Yang, L., Tan, V. Y., and Cheung, W. C. Best arm identification with possibly biased offline data. *arXiv preprint arXiv:2505.23165*, 2025.

Yin, Z. and Fang, Z. Multi-armed bandits with biased and heteroscedastic auxiliary rewards. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3899–3908, 2025.

Yu, Z. and Zhang, X. Actor-critic alignment for offline-to-online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40452–40474. PMLR, 2023.

Zhang, C., Agarwal, A., Daumé III, H., Langford, J., and Negahban, S. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7335–7344. PMLR, 2019.

Zhang, W., Zhou, D., Li, L., and Gu, Q. Neural Thompson sampling. In *The 9th International Conference on Learning Representations*, 2021.

Zhang, Y., Zhu, R., and Xie, Q. Contextual online pricing with (biased) offline data. In *The 39th Annual Conference on Neural Information Processing Systems*, 2025.

Zhou, K., Zhang, T., Chen, W., and Kong, F. Hybrid combinatorial multi-armed bandits with probabilistically triggered arms. *arXiv preprint arXiv:2512.21925*, 2025.

A. Proof of Theorem 5.1

We first introduce some useful definitions that will be used in the proof.

For each sub-optimal arm $i \neq 1$, we define two thresholds: $x_i := \mu_i^{(\text{on})} + \Delta_i/3$ and $y_i := \mu_i^{(\text{on})} + 2\Delta_i/3 = \mu_1^{(\text{on})} - \Delta_i/3$. It holds that $x_i < y_i$ and $y_i - x_i = \Delta_i/3$.

Define the confidence radius $L_t := 2\log(2t/\delta_t)$ with uncertainty level $\delta_t = 1/(t^2\Delta_i^2)$.

Let $\mathcal{F}_t := \cup_{i=1}^K S_i \cup \{(A(\tau), R_{A(\tau)}(\tau)) : 1 \leq \tau \leq t\}$ to represent the history comprising the offline dataset and the sequence of online observations collected up to round t.

For the optimal arm 1 and arm i, define $p_{i,t} := \Pr\big(\hat{\theta}_1(t) > y_i \mid \mathcal{F}_{t-1}\big)$ to represent the conditional probability that the median-index is larger than y_i. Similarly, define $p_{i,t}^{(\text{on})} := \Pr\big(\theta_1^{(\text{on})}(t) > y_i \mid \mathcal{F}_{t-1}\big)$ and $p_{i,t}^{(\text{hyb})} := \Pr\big(\theta_1^{(\text{hyb})}(t) > y_i \mid \mathcal{F}_{t-1}\big)$.

Further define the following good events representing the estimators are not far from their centers.

Definition A.1 (Good events $E_i^{\mu(\text{on})}(t), E_i^{\mu(\text{hyb})}(t), E_i^{\theta(\text{hyb})}(t), \mathcal{G}_i(t), E_1^{\mu(\text{on})}(t)$)**.** For $i \neq 1$, define the following good event

$$
\begin{aligned}
E_i^{\mu(\text{on})}(t) &:= \left\{\hat{\mu}_i^{(\text{on})}(t) \leq x_i\right\}, \\
E_i^{\mu(\text{hyb})}(t) &:= \left\{\left|\hat{\mu}_i^{(\text{hyb})}(t) - \mu_i^{(\text{on})}\right| \leq \sqrt{\frac{L_t}{N_i + T_i(t)}} + \frac{T_i(t)V_i}{N_i + T_i(t)}\right\}, \\
E_i^{\theta(\text{hyb})}(t) &:= \left\{\left|\theta_i^{(\text{hyb})}(t) - \hat{\mu}_i^{(\text{hyb})}(t)\right| \leq \sqrt{\frac{L_t}{N_i + T_i(t)}} + \frac{T_i(t)V_i}{N_i + T_i(t)}\right\}.
\end{aligned}
$$

and the global good event

$$
\mathcal{G}_i(t) := E_i^{\mu(\text{on})}(t) \cap E_i^{\mu(\text{hyb})}(t) \cap E_i^{\theta(\text{hyb})}(t)\,.
$$

Similarly, for optimal arm 1, define

$$
E_1^{\mu(\text{on})}(t) := \left\{\hat{\mu}_1^{(\text{on})}(t) > y_i\right\}.
$$

In the following, we provide the detailed proof of Theorem 5.1.

Proof of Theorem 5.1. We first analyze the regret by bounding the number of selections of each sub-optimal arm $i \neq 1$.

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{A(t) = i\}\right] &= \sum_{t=1}^T \Pr(A(t) = i,\ \neg E_i^{\mu(\text{on})}(t)) + \sum_{t=1}^T \Pr(A(t) = i,\ \neg E_i^{\mu(\text{hyb})}(t)) \\
&\quad + \sum_{t=1}^T \Pr(A(t) = i,\ \neg E_i^{\theta(\text{hyb})}(t)) + \sum_{t=1}^T \Pr(A(t) = i, \mathcal{G}_i(t)) \qquad (3) \\
&\leq O\left(\frac{1}{\Delta_i^2}\right) + \sum_{t=1}^T \Pr(A(t) = i, \mathcal{G}_i(t)) \qquad (4) \\
&\leq O\left(\frac{1}{\Delta_i^2}\right) + \sum_{t=1}^T \Pr\underbrace{(A(t) = i, \mathcal{G}_i(t), \hat{\theta}_i(t) > y_i)}_{\mathcal{E}_{1,t}} \\
&\qquad + \sum_{t=1}^T \Pr\underbrace{(A(t) = i, \mathcal{G}_i(t), \hat{\theta}_i(t) \leq y_i)}_{\mathcal{E}_{2,t}} \\
&\leq O\left(\frac{1}{\Delta_i^2}\right) + \sum_{t=1}^T \Pr(\mathcal{E}_{1,t}) + \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{E_1^{\mu(\text{on})}(\tau_k + 1)\right\}\right]
\end{aligned}
$$

$$+\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\mathbb{1}\left\{\neg E_1^{\mu(\text{on})}(\tau_k+1)\right\}\right] \tag{5}$$

$$\leq \frac{C_3}{\Delta_i^2}+\left(\frac{C_1\log T}{\Delta_i^2}-N_i\max\left\{\left(1-\frac{3\omega_i}{\Delta_i}\right),0\right\}\right)_+ + c\left(\frac{C_2\log T}{\Delta_i^2}-N_1\right)_+ . \tag{6}$$

Here (4) follows from Lemma D.4, Lemma A.2, and Lemma A.3 for bounding the bad events. (5) is due to Lemma A.5. (6) comes from the upper bound for $\mathcal{E}_1$ in Lemma A.4 and the upper bound for two terms of $\mathcal{E}_2$ in Lemma A.6 and Lemma A.9. C_1, C_2, C_3 are constant terms independent of the problem parameters, $c=\max\left\{e^{11},\exp\left\{\frac{28+16\sqrt{3}}{N_1+1}\right\}\right\}+5$ is a constant smaller than the coeffient in the pure online TS (Agrawal & Goyal, 2017) and decreases as N_1 increases. O hides constant terms.

The final regret can then be obtained by

$$\begin{aligned}\text{Reg}(T) &= \sum_{i\neq 1}\Delta_i\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{A(t)=i\}\right]\\ &\leq \sum_{i\neq 1}\Delta_i\left(\frac{C_3}{\Delta_i^2}+\left(\frac{C_1\log T}{\Delta_i^2}-N_i\max\left\{\left(1-\frac{3\omega_i}{\Delta_i}\right),0\right\}\right)_+ + c\left(\frac{C_2\log T}{\Delta_i^2}-N_1\right)_+\right).\end{aligned}$$

□

We next introduce the lemmas used in the above proof and provide the proof of lemmas in Appendix C. The first two lemmas establish constant upper bounds on the probability of bad events.

Lemma A.2. *For sub-optimal arm $i\neq 1$, the second term in (3) can be bounded by*

$$\sum_{t=1}^{T}\Pr(A(t)=i,\ \neg E_i^{\mu(\text{hyb})}(t))\leq\sum_{t=1}^{T}\delta_t=O\left(\frac{1}{\Delta_i^2}\right).$$

Lemma A.3. *For sub-optimal arm $i\neq 1$, the third term in (3) can be bounded by*

$$\sum_{t=1}^{T}\Pr(A(t)=i,\ \neg E_i^{\theta(\text{hyb})}(t))\leq\sum_{t=1}^{T}\frac{\delta_t}{t}=O\left(\frac{1}{\Delta_i^2}\right).$$

The following lemma establishes an upper bound on $\sum_{t=1}^{T}\Pr(\mathcal{E}_{1,t})$ in the regret, corresponding to the regret incurred when the global good event $\mathcal{G}_i(t)$ holds and the index $\hat{\theta}_i(t)$ of arm i is inaccurate.

Lemma A.4. *For sub-optimal arm $i\neq 1$,*

$$\sum_{t=1}^{T}\Pr(\mathcal{E}_{1,t})\leq\left(36\frac{\log(T\Delta_i^2+e^6)}{\Delta_i^2}-N_i\max\left\{\left(1-\frac{3\omega_i}{\Delta_i}\right),0\right\}\right)_+ +\frac{1}{\Delta_i^2}.$$

The following lemma establishes an upper bound on $\sum_{t=1}^{T}\Pr(\mathcal{E}_{2,t})$ in the regret. This term can be transformed into the case where the index $\hat{\theta}_1(t)$ of arm 1 is inaccurate, which may lead to the selection of arm i.

Lemma A.5. *For all t, $i\neq 1$ and all instantiations F_{t-1} of $\mathcal{F}_{t-1}$,*

$$\Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \hat{\theta}_i(t)\leq y_i\mid F_{t-1}\big)\ \leq\ \frac{1-p_{i,t}}{p_{i,t}}\ \Pr\big(A(t)=1,\ \mathcal{G}_i(t),\ \hat{\theta}_i(t)\leq y_i\mid F_{t-1}\big).$$

Further,

$$\sum_{t=1}^{T}\Pr\big(\mathcal{E}_{2,t}\big)\leq\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\mathbb{1}\left\{E_1^{\mu(\text{on})}(\tau_k+1)\right\}\right]+\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\mathbb{1}\left\{\neg E_1^{\mu(\text{on})}(\tau_k+1)\right\}\right].$$

Lemma A.6. *For any $t \geq 0$,*

$$\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{E_1^{\mu(\mathrm{on})}(\tau_k+1)\right\}\right] \;\leq\; \sum_{k=0}^{T-1} \min\left\{\mathbb{E}\big[\tfrac{1}{p^{(\mathrm{on})}_{i,\tau_k+1}}-1\big],\; \mathbb{E}\big[\tfrac{1}{p^{(\mathrm{hyb})}_{i,\tau_k+1}}-1\big]\right\}.$$

We then prove a similar lemma corresponding to Lemma D.6 for the hybrid estimator.

Lemma A.7. *Let τ_k be the time of the k-th play of arm 1. Then*

$$\mathbb{E}\Big[\tfrac{1}{p^{(\mathrm{hyb})}_{i,\tau_k+1}}-1\Big] \leq \begin{cases} \max\left\{e^{11}, \exp\left(\frac{28+16\sqrt{3}}{N_1+1}\right)\right\}+5, & \forall\, k, \\ \dfrac{5}{T\Delta_i^2}, & k > L_i^{(\mathrm{hyb})}(T), \end{cases}$$

where $L_i^{(\mathrm{hyb})}(T) = \left(\frac{288\ln(T\Delta_i^2+e^6)}{\Delta_i^2} - N_1\right)_+$.

Lemma A.8.

$$\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{E_1^{\mu(\mathrm{on})}(\tau_k+1)\right\}\right] \;\leq\; C_1\left(\frac{\log(T\Delta_i^2+e^{32})}{\Delta_i^2} - N_1\right)_+ + C_2\,\frac{1}{\Delta_i^2},$$

for suitable constants $C_1, C_2 > 0$ independent of T.

Proof of Lemma A.8. Combining Lemma A.6, Lemma D.6 and Lemma A.7 we get the bound. □

Lemma A.9.

$$\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\right\}\right] \;\leq\; \frac{C}{\Delta_i^2},$$

for some constant C.

B. Theoretical Analysis and Experiments in the Pure Online Setting

In this section, we provide a tighter analysis to the pure online setting, where no offline data are available. In this case, the hybrid posterior coincides with the online posterior, and Anchor-TS reduces to selecting arms according to the median index

$$\mathrm{median}\left\{\theta_{i,1}^{(\mathrm{on})}(t), \theta_{i,2}^{(\mathrm{on})}(t), \hat{\mu}_i^{(\mathrm{on})}\right\},$$

where $\theta_{i,1}^{(\mathrm{on})}$ and $\theta_{i,2}^{(\mathrm{on})}$ are i.i.d. samples from the standard online posterior of arm i at time t. We show that this median rule preserves the classical instance-dependent regret guarantee while improving the leading logarithmic term by a factor $1/2$ compared with vanilla TS.

B.1. $1/2$ improvement on the leading $\log T$-term

We reuse the notation in Appendix A. Recall that the $\log T$ contribution in Theorem 5.1 arises from two parts: (i) controlling over-optimistic indices of suboptimal arms (captured by the term $\sum_{t=1}^{T}\Pr(\mathcal{E}_{1,t})$ as in Lemma A.4), and (ii) controlling under-exploration of the optimal arm (via the ratio term in Lemma A.5). We bound these two parts separately.

Suboptimal-arm term. Recall that for a suboptimal arm $i \neq 1$,

$$\sum_{t=1}^{T}\Pr(\mathcal{E}_{1,t}) = \sum_{t=1}^{T}\Pr\left(A(t)=i, \mathcal{G}_i(t), \hat{\theta}_i(t) > y_i\right). \tag{7}$$

Conditioned on the global good event $\mathcal{G}_i(t)$, we have $\hat{\mu}_i^{(\mathrm{on})} < x_i < y_i$. Therefore, the event $\{\hat{\theta}_i(t) > y_i\}$ can only happen when both posterior samples exceed y_i, i.e.,

$$\left\{\mathrm{median}\{\theta_{i,1}^{(\mathrm{on})}(t), \theta_{i,2}^{(\mathrm{on})}(t), \hat{\mu}_i^{(\mathrm{on})}(t)\} > y_i\right\} = \left\{\theta_{i,1}^{(\mathrm{on})}(t) > y_i \;\cap\; \theta_{i,2}^{(\mathrm{on})}(t) > y_i\right\}.$$

Since $\theta_{i,1}^{(\mathrm{on})}(t)$, $\theta_{i,2}^{(\mathrm{on})}(t)$ are i.i.d. conditional on F_{t-1},

$$\Pr\Big(\hat{\theta}_i(t) > y_i \mid \mathcal{F}_{t-1}\Big) = \Pr\Big(\theta_i^{(\mathrm{on})}(t) > y_i \mid \mathcal{F}_{t-1}\Big)^2 . \tag{8}$$

Using (8) squares the Gaussian tail probability (Lemma D.2) in the vanilla TS analysis (Lemma D.4) and yields a halved exploration threshold as

$$\begin{aligned}
\sum_{t=1}^{T} \Pr\Big(T_i(t) > \tfrac{1}{2}L_i(T),\ G_i(t),\ \hat{\theta}_i(t) > y_i\Big) &\le \sum_{t=\frac{1}{2}L_i(t)}^{T} \left(\exp\left\{-\frac{T_i(t)(y_i - x_i)^2}{2}\right\}\right)^2 \\
&\le \sum_{t=1}^{T} \left(\exp\left\{-\frac{\frac{1}{2}L_i(t)(y_i - x_i)^2}{2}\right\}\right)^2 \\
&\le \sum_{t=1}^{T} \frac{1}{T\Delta_i^2} \\
&\le \frac{1}{\Delta_i^2} ,
\end{aligned} \tag{9}$$

where the threshold $L_i(T)$ is the same as in Lemma D.4.

Thus take (9), we get

$$\begin{aligned}
\sum_{t=1}^{T} \Pr(\mathcal{E}_{1,t}) &\le \sum_{t=1}^{T} \Pr\Big(A(t) = i,\ T_i(t) \le \tfrac{1}{2}L_i(T),\ G_i(t),\ \hat{\theta}_i(t) > y_i\Big) \\
&\qquad + \sum_{t=1}^{T} \Pr\Big(A(t) = i,\ T_i(t) > \tfrac{1}{2}L_i(T),\ G_i(t),\ \hat{\theta}_i(t) > y_i\Big) \\
&\le \tfrac{1}{2}L_i(T) \ + \ \sum_{t=1}^{T} \Pr\Big(T_i(t) > \tfrac{1}{2}L_i(T),\ G_i(t),\ \hat{\theta}_i(t) > y_i\Big) \\
&\le \tfrac{1}{2}L_i(T) + \frac{1}{\Delta_i^2} ,
\end{aligned}$$

which matches the classical bound but with a factor $1/2$ on the leading logarithmic term.

Optimal-arm term. Next we bound the term arising from Lemma A.5, namely

$$\sum_{t=1}^{T} \Pr\big(\mathcal{E}_{2,t}\big) \le \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\Big\{E_1^{\mu(\mathrm{on})}(\tau_k + 1)\Big\}\right] + \sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\Big\{\neg E_1^{\mu(\mathrm{on})}(\tau_k + 1)\Big\}\right] . \tag{10}$$

We mainly focus on the first term since the second term upper bound does not depend on T. Under $E_1^{\mu(\mathrm{on})}(\tau_k + 1)$, we have $\hat{\mu}_1^{(\mathrm{on})}(\tau_k + 1) > y_i$, hence

$$p_{i,\tau_k+1} = \Pr\Big(\hat{\theta}_1(\tau_k + 1) > y_i \mid \mathcal{F}_{\tau_k}\Big) = \Pr\Big(\theta_{1,1}^{(\mathrm{on})}(\tau_k + 1) > y_i \ \cup\ \theta_{1,2}^{(\mathrm{on})}(\tau_k + 1) > y_i \mid \mathcal{F}_{\tau_k}\Big) .$$

Let $p_{i,\tau_k+1}^{(\mathrm{on})} := \Pr(\theta_1^{(\mathrm{on})}(\tau_k + 1) > y_i \mid \mathcal{F}_{\tau_k})$. Then

$$\begin{aligned}
p_{i,\tau_k+1} &= 1 - \Big(1 - p_{i,\tau_k+1}^{(\mathrm{on})}\Big)^2 \\
&= 2\, p_{i,\tau_k+1}^{(\mathrm{on})} - \Big(p_{i,\tau_k+1}^{(\mathrm{on})}\Big)^2 .
\end{aligned}$$

Therefore,

$$\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}} = \frac{(1-p^{(\mathrm{on})}_{i,\tau_k+1})^2}{p^{(\mathrm{on})}_{i,\tau_k+1}(2-p^{(\mathrm{on})}_{i,\tau_k+1})} = \frac{1-p^{(\mathrm{on})}_{i,\tau_k+1}}{p^{(\mathrm{on})}_{i,\tau_k+1}} \cdot \frac{1-p^{(\mathrm{on})}_{i,\tau_k+1}}{2-p^{(\mathrm{on})}_{i,\tau_k+1}} \le \frac{1}{2} \cdot \frac{1-p^{(\mathrm{on})}_{i,\tau_k+1}}{p^{(\mathrm{on})}_{i,\tau_k+1}}, \tag{11}$$

where we used $(1-p)/(2-p) \le 1/2$ for $p \in [0,1]$. Plugging (11) into the first term of (10) directly yields a factor-$1/2$ reduction on the leading logarithmic contribution in the optimal-arm analysis. The second term in (10) is handled identically to Appendix A and remains a lower-order constant term.

B.2. Experiments in the pure online setting

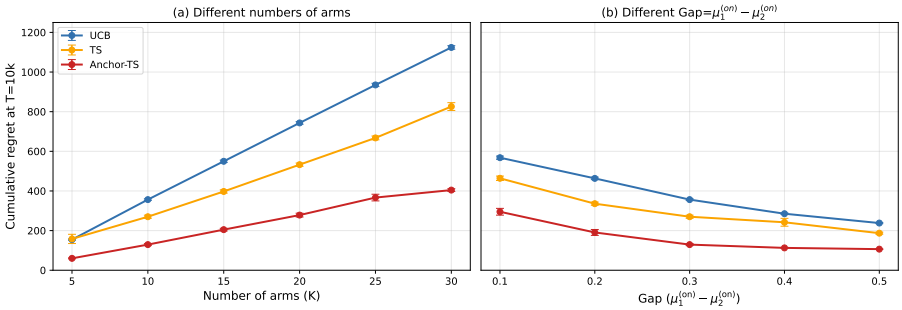

Figure 3. Cumulative regret of our Anchor-TS and baselines in the pure online setting with different arm set size K and sub-optimality gap Δ.

We examine the pure online setting in Figure 3, where no offline data are available. In this case, MIN-UCB, hybrid UCB, and hybrid TS reduce to their respective pure online variants. The experimental setting follows Section 7. We test different choices of arm set size K and sub-optimality gap $\Delta = \mu_1^{(\mathrm{on})} - \mu_2^{(\mathrm{on})}$. Anchor-TS makes decisions based on the median of two posterior samples and the empirical mean from online data. As shown in the figure, Anchor-TS consistently achieves the lowest cumulative regret across all tested settings. This demonstrates that even without offline data, the median-based decision rule employed by Anchor-TS effectively suppresses over-exploration caused by the randomness of posterior sampling, thereby exhibiting greater learning efficiency compared with vanilla TS and UCB-based methods.

C. Proof of Lemmas

Proof of Lemma A.2. We expand the difference between the empirical hybrid mean and the true mean of the online distribution in $E_i^{\mu(\mathrm{hyb})}(t)$ as follows:

$$\begin{aligned}\hat{\mu}_i^{(\mathrm{hyb})}(t) - \mu_i^{(\mathrm{on})} &= \left(\hat{\mu}_i^{(\mathrm{hyb})}(t) - \frac{T_i(t)\mu_i^{(\mathrm{on})}(t) + N_i\mu_i^{(\mathrm{off})}}{T_i(t)+N_i}\right) + \left(\frac{T_i(t)\mu_i^{(\mathrm{on})}(t) + N_i\mu_i^{(\mathrm{off})}}{T_i(t)+N_i} - \mu_i^{(\mathrm{on})}\right) \\ &= \left(\hat{\mu}_i^{(\mathrm{hyb})}(t) - \frac{T_i(t)\mu_i^{(\mathrm{on})}(t) + N_i\mu_i^{(\mathrm{off})}}{T_i(t)+N_i}\right) + \frac{N_i}{T_i(t)+N_i}\big(\mu_i^{(\mathrm{off})} - \mu_i^{(\mathrm{on})}\big).\end{aligned}$$

Using $|\mu_i^{(\mathrm{off})} - \mu_i^{(\mathrm{on})}| \le V_i$, we obtain

$$\big|\hat{\mu}_i^{(\mathrm{hyb})}(t) - \mu_i^{(\mathrm{on})}\big| \le \left|\frac{T_i(t)\hat{\mu}_i^{(\mathrm{on})}(t) + N_i\hat{\mu}_i^{(\mathrm{off})}}{T_i(t)+N_i} - \frac{T_i(t)\mu_i^{(\mathrm{on})}(t) + N_i\mu_i^{(\mathrm{off})}}{T_i(t)+N_i}\right| + \frac{N_iV_i}{T_i(t)+N_i}. \tag{12}$$

Combining the definition of $\neg E_i^{\mu(\mathrm{hyb})}(t)$

$$\big|\hat{\mu}_i^{(\mathrm{hyb})}(t) - \mu_i^{(\mathrm{on})}\big| > \sqrt{\frac{L_t}{T_i(t)+N_i}} + \frac{N_iV_i}{T_i(t)+N_i},$$

with (12), we can derive the following relation:

$$\neg E_i^{\mu(\text{hyb})}(t) \subseteq \left\{ \left| \frac{T_i(t)\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{T_i(t)+N_i} - \frac{T_i(t)\mu_i^{(\text{on})}(t) + N_i\mu_i^{(\text{off})}}{T_i(t)+N_i} \right| > \sqrt{\frac{L_t}{T_i(t)+N_i}} \right\}. \tag{13}$$

Noting that in (13),

$$\mathbb{E}\left[\frac{T_i(t)\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{T_i(t)+N_i}\right] = \frac{T_i(t)\mu_i^{(\text{on})}(t) + N_i\mu_i^{(\text{off})}}{T_i(t)+N_i}.$$

Given $T_i(t) = s$, by Chernoff-Hoeffding bounds (Lemma D.1), we have

$$\begin{aligned}\Pr\left(\left|\frac{s\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{s+N_i} - \mathbb{E}\left[\frac{s\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{s+N_i}\right]\right| > \sqrt{\frac{L_t}{s+N_i}}\right) &\le 2e^{-(s+N_i)\frac{L_t}{s+N_i}} \\ &\le \frac{\delta_t}{t}.\end{aligned}$$

Thus

$$\begin{aligned}\Pr\left(\neg E_i^{\mu(\text{hyb})}(t)\right) &= \Pr\left(\exists s \in \{0,\dots,t-1\} : T_i(t) = s \text{ and } \neg E_i^{\mu(\text{hyb})}(t)\right) \\ &\le \sum_{s=0}^{t-1} \Pr\left(\left|\frac{s\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{s+N_i} - \mathbb{E}\left[\frac{s\hat{\mu}_i^{(\text{on})}(t) + N_i\hat{\mu}_i^{(\text{off})}}{s+N_i}\right]\right| > \sqrt{\frac{L_t}{s+N_i}}\right) \\ &\le \sum_{s=0}^{t-1} \frac{\delta_t}{t} \le \delta_t.\end{aligned}$$

Then we have

$$\sum_{t=1}^{T} \Pr(A(t) = i,\ \neg E_i^{\mu(\text{hyb})}(t)) \le \sum_{t=1}^{T} \Pr\left(\neg E_i^{\mu(\text{hyb})}(t)\right) \le \sum_{t=1}^{T} \delta_t = O\left(\frac{1}{\Delta_i^2}\right).$$

□

Proof of Lemma A.3. Recall that

$$\theta_i^{(\text{hyb})}(t) \sim \mathcal{N}\left(\hat{\mu}_i^{(\text{hyb})}(t) + \frac{N_iV_i}{T_i(t)+N_i}, \frac{1}{T_i(t)+N_i+1}\right),$$

thus $\mathbb{E}[\theta_i^{(\text{hyb})}(t)] = \hat{\mu}_i^{(\text{hyb})}(t) + \frac{N_iV_i}{T_i(t)+N_i}$. By the triangle inequality,

$$\left|\theta_i^{(\text{hyb})}(t) - \hat{\mu}_i^{(\text{hyb})}(t)\right| \le \left|\theta_i^{(\text{hyb})}(t) - \mathbb{E}[\theta_i^{(\text{hyb})}(t)]\right| + \left|\mathbb{E}[\theta_i^{(\text{hyb})}(t)] - \hat{\mu}_i^{(\text{hyb})}(t)\right|$$

Thus

$$\begin{aligned}\neg E_i^{\theta(\text{hyb})}(t) &= \left\{\left|\theta_i^{(\text{hyb})}(t) - \hat{\mu}_i^{(\text{hyb})}(t)\right| > \sqrt{\frac{L_t}{T_i(t)+N_i}} + \frac{N_iV_i}{T_i(t)+N_i}\right\} \\ &\subseteq \left\{\left|\theta_i^{(\text{hyb})}(t) - \mathbb{E}[\theta_i^{(\text{hyb})}(t)]\right| \ge \sqrt{\frac{L_t}{T_i(t)+N_i}}\right\}.\end{aligned} \tag{14}$$

Conditional on the filtration $\mathcal{F}_t$, $\hat{\mu}_i^{(\text{hyb})}(t)$ and $T_i(t)$ are deterministic. Thus the right-hand of (14) can be bounded by Gaussian tail inequality (Lemma D.2) as

$$\Pr\left(\neg E_i^{\theta(\text{hyb})}(t) | \mathcal{F}_t\right) \le 2\exp\left(-\frac{1}{2}\frac{L_t}{T_i(t)+N_i} \cdot (T_i(t)+N_i+1)\right) \le \frac{\delta_t}{t}.$$

Then we have

$$\begin{aligned}\sum_{t=1}^{T}\Pr(A(t)=i,\ \neg E_i^{\theta(\text{hyb})}(t)) &= \sum_{t=1}^{T}\mathbb{E}\left[\Pr(A(t)=i,\ \neg E_i^{\theta(\text{hyb})}(t)|\mathcal{F}_t)\right]\\ &\le \sum_{t=1}^{T}\mathbb{E}\left[\Pr(\ \neg E_i^{\theta(\text{hyb})}(t)|\mathcal{F}_t)\right]\\ &\le \sum_{t=1}^{T}\frac{\delta_t}{t}\\ &= O\left(\frac{1}{\Delta_i^2}\right).\end{aligned}$$

□

Proof of Lemma A.4. $\hat{\theta}_i(t) > y_i$ indicates that the median among three indices is above y_i, which implies that both the online and hybrid indexes are above y_i based on the event $E_i^{\mu(\text{on})}(t)$, i.e.,

$$\{\hat{\theta}_i(t) > y_i\} \subseteq \{\theta_i^{(\text{on})}(t) > y_i\} \cap \{\theta_i^{(\text{hyb})}(t) > y_i\} \text{ when } \mathcal{G}_i(t) \text{ holds.}$$

Therefore

$$\begin{aligned}\Pr(\mathcal{E}_{1,t}) &= \Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \hat{\theta}_i(t) > y_i\big)\\ &\le \min\left\{\Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \theta_i^{(\text{on})}(t) > y_i\big), \Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \theta_i^{(\text{hyb})}(t) > y_i\big)\right\}.\end{aligned}$$

Then

$$\begin{aligned}\sum_{t=1}^{T}\Pr(\mathcal{E}_{1,t}) &\le \sum_{t=1}^{T}\min\left\{\Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \theta_i^{(\text{on})}(t) > y_i\big), \Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \theta_i^{(\text{hyb})}(t) > y_i\big)\right\}\\ &\le \min\left\{\underbrace{\sum_{t=1}^{T}\Pr\big(A(t)=i,\ E_i^{\mu(\text{on})}(t),\ \theta_i^{(\text{on})}(t) > y_i\big)}_{S_1}, \underbrace{\sum_{t=1}^{T}\Pr\big(A(t)=i,\ \mathcal{G}_i(t),\ \theta_i^{(\text{hyb})}(t) > y_i\big)}_{S_2}\right\}. \quad (15)\end{aligned}$$

Note that S_1 is exactly the same as the corresponding term in vanilla TS (Agrawal & Goyal, 2017). By Lemma D.5,

$$S_1 \le \frac{2\log(T\Delta_i^2)}{(y_i - x_i)^2} + \frac{1}{\Delta_i^2} = \frac{18\log(T\Delta_i^2)}{\Delta_i^2} + \frac{1}{\Delta_i^2}. \quad (16)$$

Next we bound term S_2. Based on the good event $\mathcal{G}_i(t)$ and triangle inequality,

$$\begin{aligned}\theta_i^{(\text{hyb})}(t) - \mu_i^{(\text{on})} &= \theta_i^{(\text{hyb})}(t) - \mu_i^{(\text{hyb})} + \mu_i^{(\text{hyb})} - \mu_i^{(\text{on})}\\ &\le 2\sqrt{\frac{L_t}{T_i(t)+N_i}} + \frac{N_i V_i}{T_i(t)+N_i} + \frac{N_i}{T_i(t)+N_i}(\mu_i^{(\text{off})} - \mu_i^{(\text{on})})\\ &= 2\sqrt{\frac{L_t}{T_i(t)+N_i}} + \frac{N_i\omega_i}{T_i(t)+N_i}. \quad (17)\end{aligned}$$

Conditional on the event $\{\theta_i^{(\text{hyb})}(t) > y_i\}$,

$$\theta_i^{(\text{hyb})}(t) - \mu_i^{(\text{on})} > y_i - \mu_i^{(\text{on})} = 2\Delta_i/3. \quad (18)$$

Combining (17) and (18),

$$
\begin{aligned}
S_2 &\leq \sum_{t=1}^{T} \Pr\left(A(t)=i,\ 2\sqrt{\frac{L_t}{T_i(t)+N_i}} + \frac{N_i\omega_i}{T_i(t)+N_i} \geq \frac{2\Delta_i}{3}\right) \\
&\leq \sum_{t=1}^{T} \Pr\left(A(t)=i,\ 2\sqrt{\frac{L_t}{T_i(t)+N_i}} \geq \frac{\Delta_i}{3} \ \text{ or } \ \frac{N_i\omega_i}{T_i(t)+N_i} \geq \frac{\Delta_i}{3}\right) \\
&= \sum_{t=1}^{T} \Pr\left(A(t)=i,\ T_i(t)+N_i \leq \frac{36L_t}{\Delta_i^2} \ \text{ or } \ T_i(t)+N_i \leq \frac{3N_i\omega_i}{\Delta_i}\right) \\
&\leq \sum_{t=1}^{T} \Pr\left(A(t)=i,\ T_i(t)+N_i \leq \max\left\{\frac{36L_t}{\Delta_i^2}, \frac{3N_i\omega_i}{\Delta_i}\right\}\right) \\
&\leq \left(\max\left\{36\frac{L_T}{\Delta_i^2}, 3\frac{N_i\omega_i}{\Delta_i}\right\} - N_i\right)_+ \leq \left(36\frac{\log(T\Delta_i^2)}{\Delta_i^2} - N_i\Big(1-3\frac{\omega_i}{\Delta_i}\Big)\right)_+
\end{aligned} \tag{19}
$$

Above all, based on (15), (16) and (19),

$$
\begin{aligned}
\sum_{t=1}^{T} \Pr(\mathcal{E}_{1,t}) &\leq \min\{S_1, S_2\} \\
&\leq \left(36\frac{\log(T\Delta_i^2)}{\Delta_i^2} - N_i \max\left\{\Big(1-3\frac{\omega_i}{\Delta_i}\Big), 0\right\}\right)_+ + \frac{1}{\Delta_i^2}.
\end{aligned} \tag{20}
$$

□

Proof of Lemma A.5. Our goal is to demonstrate

$$
\Pr\big(A(t)=i,\ \hat{\theta}_i(t) \leq y_i \mid \mathcal{G}_i(t),\ F_{t-1}\big) \ \leq \ \frac{1-p_{i,t}}{p_{i,t}} \Pr\big(A(t)=1,\ \hat{\theta}_i(t) \leq y_i \mid \mathcal{G}_i(t),\ F_{t-1}\big). \tag{21}
$$

For sub-optimal arm i to be chosen given this constraint, every other arm j must satisfy $\hat{\theta}_j(t) \leq \hat{\theta}_i(t) \leq y_i$. Besides, noting that given instantiation F_{t-1}, the posterior distribution between optimal arm 1 and other sub-optimal arms are independent, and $\Pr\left(\hat{\theta}_1(t) \leq y_i \mid \mathcal{G}_i(t),\ F_{t-1}\right) = \Pr\left(\hat{\theta}_1(t) \leq y_i \mid F_{t-1}\right)$. Therefore,

$$
\begin{aligned}
\text{LHS of (21)} &\leq \Pr\left(\hat{\theta}_j(t) \leq y_i, \forall\, j \mid \mathcal{G}_i(t),\ F_{t-1}\right) \\
&= \Pr\left(\hat{\theta}_1(t) \leq y_i \mid F_{t-1}\right) \cdot \Pr\left(\hat{\theta}_j(t) \leq y_i, \forall\, j \neq 1 \mid \mathcal{G}_i(t),\ F_{t-1}\right) \\
&= (1-p_{i,t}) \cdot \Pr\left(\hat{\theta}_j(t) \leq y_i, \forall\, j \neq 1 \mid \mathcal{G}_i(t),\ F_{t-1}\right).
\end{aligned}
$$

Next, consider the probability of selecting the optimal arm 1 under the same conditions. Arm 1 is chosen if its median index $\hat{\theta}_1(t)$ exceeds all others. Thus we have

$$
\begin{aligned}
\Pr\big(A(t)=1 \mid \mathcal{G}_i(t),\ F_{t-1}\big) &\geq \Pr\left(\hat{\theta}_1(t) > y_i \geq \hat{\theta}_j(t), \forall j \mid \mathcal{G}_i(t),\ F_{t-1}\right) \\
&= \Pr\left(\hat{\theta}_1(t) > y_i \mid F_{t-1}\right) \\
&\qquad \cdot \Pr\left(\hat{\theta}_j(t) \leq y_i, \forall\, j \neq 1 \mid \mathcal{G}_i(t),\ F_{t-1}\right) \\
&= p_{i,t} \cdot \Pr\left(\hat{\theta}_j(t) \leq y_i, \forall\, j \neq 1 \mid \mathcal{G}_i(t),\ F_{t-1}\right).
\end{aligned}
$$

Combining the above two inequalities, we get the first result in the lemma.

For the term $\mathcal{E}_{2,t}$, taking expectations and summing over t yields

$$\begin{aligned}\sum_{t=1}^{T}\Pr\big(\mathcal{E}_{2,t}\big) &= \sum_{t=1}^{T}\mathbb{E}\Big[\Pr\big(\mathcal{E}_{2,t}\mid\mathcal{F}_{t-1}\big)\Big]\\ &\le \sum_{t=1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\,\Pr\big(A(t)=1,\mathcal{G}_i(t)\mid\mathcal{F}_{t-1}\big)\right] \qquad (22)\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\,\mathbb{1}\{A(t)=1,\mathcal{G}_i(t)\}\mid\mathcal{F}_{t-1}\right]\right] \qquad (23)\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\frac{1-p_{i,t}}{p_{i,t}}\,\mathbb{1}\{A(t)=1,\mathcal{G}_i(t)\}\right]. \qquad (24)\end{aligned}$$

(22) is derived from the first result in the lemma, (23) above uses that $p_{i,t}$ is fixed given $\mathcal{F}_{t-1}$.

Let τ_k be the time of the k-th pull of arm 1 ($k \ge 1$) and set $\tau_0 := 0$. Between two consecutive pulls of arm 1, neither its posteriors nor their empirical mean estimations change, so both $p_{i,t}$ and $E_1^{\mu(\mathrm{on})}(t)$ is invariant within each block $\{\tau_k+1,\dots,\tau_{k+1}\}$. Summing up the sum in (24) block wise we obtain

$$\begin{aligned}\text{RHS of (24)} &\le \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\sum_{t=\tau_k+1}^{\tau_{k+1}}\mathbb{1}\{A(t)=1,\mathcal{G}_i(t)\}\right]\\ &\le \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\right] \qquad (25)\\ &\le \sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\,\mathbb{1}\Big\{E_1^{\mu(\mathrm{on})}(\tau_k+1)\Big\}\right]+\sum_{k=0}^{T-1}\mathbb{E}\left[\frac{1-p_{i,\tau_k+1}}{p_{i,\tau_k+1}}\,\mathbb{1}\Big\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\Big\}\right],\end{aligned}$$

(25) is derived because each block contains at most one round with $A(t)=1$ and $E_1^{\mu(\mathrm{on})}(t)$ coincides with $E_1^{\mu(\mathrm{on})}(\tau_k+1)$ throughout the block. □

Proof of Lemma A.6. For any time t, define $\mathrm{MIX}_t := \max\{p_{i,t}^{(\mathrm{on})},\, p_{i,t}^{(\mathrm{hyb})}\}$. We first prove that when $E_1^{\mu(\mathrm{on})}(t)$ holds,

$$p_{i,t} = \Pr\big(\hat\theta_1(t) > y_i \mid \mathcal{F}_{t-1}\big) \;\ge\; \mathrm{MIX}_t\,. \qquad (26)$$

Recall that event $E_1^{\mu(\mathrm{on})}(t) = \{\hat\mu_1^{(\mathrm{on})}(t) > y_i\}$. This implies with $E_1^{\mu(\mathrm{on})}(t)$ being true,

$$\Big(\{\theta_1^{(\mathrm{hyb})}(t) > y_i\}\cup\{\theta_1^{(\mathrm{on})}(t) > y_i\}\Big) \subseteq \{\hat\theta_1(t) > y_i\}.$$

Taking conditional probabilities given $\mathcal{F}_{t-1}$ and using that $\theta_1^{(\mathrm{on})}(t)$ and $\theta_1^{(\mathrm{hyb})}(t)$ are independent of $\hat\mu_1^{(\mathrm{on})}(t)$ conditional on $\mathcal{F}_{t-1}$, we obtain

$$p_{i,t} \;\ge \Pr\Big(\{\theta_1^{(\mathrm{hyb})}(t) > y_i\}\cup\{\theta_1^{(\mathrm{on})}(t) > y_i\}\mid\mathcal{F}_{t-1}\Big) \ge\; \max\{p_{i,t}^{(\mathrm{on})}, p_{i,t}^{(\mathrm{hyb})}\} = \mathrm{MIX}_t \text{ on } E_1^{\mu(\mathrm{on})}(t).$$

Further, using $\frac{1-p}{p} = \frac{1}{p} - 1$ and the monotonicity of $x \mapsto 1/x$ on $(0,1]$, (26) implies that on $E_1^{\mu(\mathrm{on})}(t)$

$$\frac{1-p_{i,t}}{p_{i,t}} \le \frac{1-\mathrm{MIX}_t}{\mathrm{MIX}_t} = \frac{1}{\mathrm{MIX}_t} - 1,$$

and multiplying by $\mathbb{1}\Big\{E_1^{\mu(\mathrm{on})}(t)\Big\}$ gets the point-wise bound

$$\frac{1-p_{i,t}}{p_{i,t}}\mathbb{1}\Big\{E_1^{\mu(\mathrm{on})}(t)\Big\} \;\le\; \left(\frac{1}{\mathrm{MIX}_t}-1\right)\mathbb{1}\Big\{E_1^{\mu(\mathrm{on})}(t)\Big\} \le \frac{1}{\mathrm{MIX}_t} - 1.$$

The last inequality comes from $\mathbb{1}\left\{E_1^{\mu(\mathrm{on})}(t)\right\} \le 1$.

Since $\mathrm{MIX}_t = \max\{p_{i,t}^{(\mathrm{on})}, p_{i,t}^{(\mathrm{hyb})}\}$,

$$\frac{1}{\mathrm{MIX}_t} - 1 \;\le\; \min\left\{\frac{1}{p_{i,t}^{(\mathrm{on})}} - 1,\; \frac{1}{p_{i,t}^{(\mathrm{hyb})}} - 1\right\}. \tag{27}$$

Applying (26)–(27) at time $\tau_k + 1$, we obtain that for each k,

$$\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{E_1^{\mu(\mathrm{on})}(\tau_k + 1)\right\} \;\le\; \min\Big\{\frac{1}{p_{i,\tau_k+1}^{(\mathrm{on})}} - 1,\; \frac{1}{p_{i,\tau_k+1}^{(\mathrm{hyb})}} - 1\Big\}, \tag{28}$$

and hence

$$\sum_{k=0}^{T-1} \mathbb{E}\left[\frac{1 - p_{i,\tau_k+1}}{p_{i,\tau_k+1}} \mathbb{1}\left\{E_1^{\mu(\mathrm{on})}(\tau_k + 1)\right\}\right] \;\le\; \sum_{k=0}^{T-1} \min\Big\{\mathbb{E}\big[\tfrac{1}{p_{i,\tau_k+1}^{(\mathrm{on})}} - 1\big],\; \mathbb{E}\big[\tfrac{1}{p_{i,\tau_k+1}^{(\mathrm{hyb})}} - 1\big]\Big\}. \tag{29}$$

□

Proof of Lemma A.7. Recall that $p_{i,t}^{(\mathrm{hyb})}$ denotes the probability that $\theta_1^{(\mathrm{hyb})}(t)$ exceeds y_i given F_{t-1}, and for the algorithm with Gaussian priors we have

$$\theta_1^{(\mathrm{hyb})}(t) \sim \mathcal{N}\left(\hat{\mu}_1^{(\mathrm{hyb})}(t) + \frac{N_1 V_1}{T_1(t) + N_1},\; \frac{1}{T_1(t) + N_1 + 1}\right).$$

Given F_{τ_k}, let Θ_k denote a Gaussian random variable sampled from the above gaussian distribution with $T_1(t) = k$. Let G_k be the geometric random variable denoting the number of consecutive independent trials until a sample of Θ_k becomes greater than y_i. Then

$$p_{i,\tau_k+1} \;=\; \Pr(\Theta_k > y_i \mid \mathcal{F}_{\tau_k})$$

and hence

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}}\right] = \mathbb{E}\big[\mathbb{E}[G_k \mid \mathcal{F}_{\tau_k}]\big] = \mathbb{E}[G_k].$$

We first bound $\mathbb{E}[G_k]$ by a constant for all k.

Fix any integer $r \ge 1$. Let $z = \sqrt{\ln r}$ and let the random variable MAX_r denote the maximum of r independent samples of Θ_k. For brevity, write $\hat{\mu}_1^{(\mathrm{hyb})} = \hat{\mu}_1^{(\mathrm{hyb})}(\tau_k + 1)$. Then, for any integer $r \ge 1$,

$$\begin{aligned}
\Pr(G_k \le r) \ge& \Pr(\mathrm{MAX}_r > y_i) \\
\ge& \Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \ge y_i\right) \\
=& \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}\Big\{\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \ge y_i\Big\} \mid \mathcal{F}_{\tau_k}\big]\Big] \\
=& \mathbb{E}\Big[\mathbb{1}\Big\{\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \ge y_i\Big\} \cdot \\
& \Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \mid \mathcal{F}_{\tau_k}\right)\Big].
\end{aligned} \tag{30}$$

Given F_{τ_k}, Θ_k is Gaussian with distribution $\mathcal{N}\left(\hat{\mu}_1^{(\mathrm{hyb})}(t) + \frac{N_1 V_1}{k+N_1},\, 1/(k + N_1 + 1)\right)$, using a Gaussian tail (Lemma D.2) we have

$$\begin{aligned}
\Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \,\middle|\, \mathcal{F}_{\tau_k}\right) &\ge 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{z}{z^2 + 1} e^{-z^2/2}\right)^r \\
&= 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln r}}{\ln r + 1} \frac{1}{\sqrt{r}}\right)^r.
\end{aligned}$$

For sufficiently large r (in particular, for all $r \geq e^{11}$), we have

$$\left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\ln r}}{\ln r + 1} \frac{1}{\sqrt{r}}\right)^r \leq \exp\left(-\frac{\sqrt{r}}{4\pi r \ln r}\right) \leq \frac{1}{r^2},$$

and hence for all $r \geq e^{11}$,

$$\Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{j + N_1 + 1}} + \frac{N_1 V_1}{j + N_1} \,\middle|\, \mathcal{F}_{\tau_j}\right) \;\geq\; 1 - \frac{1}{r^2}. \tag{31}$$

Substituting (31) into (30) yields, for $r \geq e^{11}$,

$$\Pr(G_k \leq r) \;\geq\; \left(1 - \frac{1}{r^2}\right) \Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \geq y_i\right).$$

Next we obtain a lower bound on the second term using Chernoff–Hoeffding bound (Lemma D.1).

$$\begin{aligned}\Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \geq y_i\right) &\geq \Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \geq \mu_1^{(\mathrm{on})}\right) \\ &\geq \Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} \geq \frac{k \cdot \mu_1^{(\mathrm{on})} + N_1 \mu_1^{(\mathrm{off})}}{k + N_1}\right),\end{aligned}$$

where the last inequality is from a simple decomposition for the hybrid mean like (12).

Applying the Chernoff bound at time $t = \tau_k + 1$ (so that $k_1(t) = k$), we get, for any $x > 0$,

$$\Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{1}{k + N_1} + \frac{x}{\sqrt{k + N_1 + 1}} \geq \mu_1^{(\mathrm{hyb})}\right) \geq 1 - e^{-2x^2}.$$

Here, the term $\frac{1}{k+N_1+1}$ was added to $\hat{\mu}_1^{(\mathrm{hyb})}$ to adjust for the fact that $\hat{\mu}_1^{(\mathrm{hyb})}$ is not simply average of the past $k+N_1$ samples, instead, it is the sum of past $k + N_1$ samples divided by $k + N_1 + 1$. Now, we use $x := z - \frac{1}{\sqrt{k+N_1+1}} \geq z - \frac{1}{\sqrt{N_1+1}}$for all $k \geq 0$, we obtain

$$\begin{aligned}\Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \geq \mu_1\right) &\geq 1 - \exp\left(-2\left(z - \frac{1}{\sqrt{N_1 + 1}}\right)^2\right) \\ &= 1 - \frac{1}{r^2} \exp\left(\frac{4}{\sqrt{N_1 + 1}} \sqrt{\log r} - \frac{2}{N_1 + 1}\right).\end{aligned}$$

Since $y_i \leq \mu_1$, this further implies

$$\Pr\left(\hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} \geq y_i\right) \geq 1 - \frac{1}{r^2} \exp\left(\frac{4}{\sqrt{N_1 + 1}} \sqrt{\log r} - \frac{2}{N_1 + 1}\right). \tag{32}$$

Noting that for $r \geq \exp\left(\frac{28+16\sqrt{3}}{N_1+1}\right)$, we have

$$\frac{1}{r^2} \exp\left(\frac{4}{\sqrt{N_1 + 1}} \sqrt{\log r} - \frac{2}{N_1 + 1}\right) \leq \frac{1}{r^{1.5}} \tag{33}$$

and combining this with (32), (32) and (33) yields, for all $r \geq \max\left\{e^{11}, \exp\left(\frac{28+16\sqrt{3}}{N_1+1}\right)\right\}$,

$$\Pr(G_k \leq r) \geq 1 - \frac{1}{r^2} - \frac{1}{r^{1.5}}.$$

We can now bound $\mathbb{E}[G_k]$:

$$\begin{aligned}
\mathbb{E}[G_k] &= \sum_{r=0}^{\infty} \Pr(G_k \geq r) \\
&= 1 + \sum_{r=1}^{\infty} \Pr(G_k \geq r) \\
&\leq 1 + e^{11} + \sum_{r \geq 1} \left(\frac{1}{r^2} + \frac{1}{r^{1.5}} \right) \\
&\leq 1 + \max\left\{ e^{11}, \exp\left(\frac{28 + 16\sqrt{3}}{N_1 + 1} \right) \right\} + 2 + 2.7 \\
&\leq \max\left\{ e^{11}, \exp\left(\frac{28 + 16\sqrt{3}}{N_1 + 1} \right) \right\} + 6. \qquad (34)
\end{aligned}$$

Hence

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}} - 1 \right] = \mathbb{E}[G_k] - 1 \leq \max\left\{ e^{11}, \exp\left(\frac{28 + 16\sqrt{3}}{N_1 + 1} \right) \right\} + 5 \quad \text{for all } k.$$

Next we derive a tighter bound for large k.

$$k > L_i^{(\mathrm{hyb})}(T) = \left(\frac{288 \ln(T\Delta_i^2 + e^6)}{\Delta_i^2} - N_1 \right)_+ .$$

Again fix $r \geq 1$, let $z = \sqrt{\ln r}$, and define MAX_r as before. Then

$$\begin{aligned}
\Pr(G_k \leq r) &\geq \Pr(\mathrm{MAX}_r > y_i) \\
&\geq \Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} - \frac{\Delta_i}{6} \geq y_i \right) \\
&= \mathbb{E}\Big[\mathbb{1}\left\{ \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} + \frac{\Delta_i}{6} \geq \mu_1^{(\mathrm{on})} \right\} \qquad (35) \\
&\qquad \cdot \Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})} + \frac{z}{\sqrt{k + N_1 + 1}} + \frac{N_1 V_1}{k + N_1} - \frac{\Delta_i}{6} \;\middle|\; \mathcal{F}_{\tau_k} \right) \Big], \qquad (36)
\end{aligned}$$

where we used that $y_i = \mu_1 - \Delta_i/3$.

By the definition of $L_i^{(\mathrm{hyb})}(T)$ we have

$$k + N_1 \;\geq\; \frac{288 \ln(T\Delta_i^2 + e^6)}{\Delta_i^2},$$

and hence

$$\frac{\sqrt{2 \ln(T\Delta^2 + e^6)}}{\sqrt{k + N_1 + 1}} \leq \frac{\Delta_i}{12}.$$

Therefore, for all $r \leq (T\Delta_i^2 + e^6)^2$,

$$\frac{z}{\sqrt{k + N_1 + 1}} - \frac{\Delta_i}{6} = \frac{\sqrt{\log r}}{\sqrt{k + N_1 + 1}} - \frac{\Delta_i}{6} \leq -\frac{\Delta_i}{12}.$$

Using the upper tail bound for a Gaussian random variable (Lemma D.2) we obtain, for any realization F_{τ_k},

$$\Pr\left(\Theta_k > \hat{\mu}_1^{(\mathrm{hyb})}(\tau_k + 1) + \frac{N_1 V_1}{k + N_1} - \frac{\Delta_i}{12} \;\middle|\; F_{\tau_k} \right) \geq 1 - \frac{1}{2} \exp\left(-\frac{(k+1)\Delta_i^2}{288} \right) \geq 1 - \frac{1}{2(T\Delta_i^2 + e^{32})},$$

which implies

$$\Pr\left(\mathrm{MAX}_r > \hat{\mu}_1^{(\mathrm{hyb})}(\tau_k+1) + \frac{z}{\sqrt{k+N_1+1}} + \frac{N_1 V_1}{k+N_1} - \frac{\Delta_i}{6} \;\middle|\; F_{\tau_k}\right) \geq 1 - \frac{1}{2^r (T\Delta_i^2 + e^6)^r}.$$

Moreover, for any $t \geq \tau_k + 1$ we have $T_1(t) \geq k$, and by Chernoff–Hoeffding bound (Lemma D.1),

$$\begin{aligned}
\Pr\left(\hat{\mu}_1^{(\mathrm{hyb})}(t) + \frac{z}{\sqrt{k+N_1+1}} + \frac{N_1 V_1}{k+N_1} - \frac{\Delta_i}{6} \geq y_i\right) &\geq \Pr\left(\hat{\mu}_1^{(\mathrm{hyb})}(t) \geq \mu_1^{(\mathrm{hyb})} - \frac{\Delta_i}{6}\right) \\
&\geq 1 - \exp\left(-\frac{2(T_1(t)+N_1)\Delta_i^2}{36}\right) \\
&\geq 1 - \frac{1}{(T\Delta_i^2 + e^6)^{16}}.
\end{aligned}$$

Let $T' = (T\Delta_i^2 + e^6)^2$. Therefore,

$$\begin{aligned}
\mathbb{E}[G_k] &\leq \sum_{r=1}^{\infty} \Pr(G_k \geq r) \\
&\leq 1 + \sum_{r=1}^{T'} \Pr(G_k \geq r) + \sum_{r=T'+1}^{\infty} \Pr(G_k \geq r) \\
&\leq 1 + \sum_{r=1}^{T'} \left(\frac{1}{(2\sqrt{T'})^r} + \frac{1}{(T')^8}\right) + \sum_{r=T'+1}^{\infty} \left(\frac{1}{r^2} + \frac{1}{r^{1.5}}\right) \\
&\leq 1 + \frac{1}{\sqrt{T'}} + \frac{1}{(T')^7} + \frac{2}{T'} + \frac{3}{\sqrt{T'}} \\
&\leq 1 + \frac{5}{T\Delta_i^2 + e^6}.
\end{aligned}$$

The above upper bound shows that $\mathbb{E}\left[\frac{1}{p_{i,\tau_k+1}^{(\mathrm{hyb})}}\right] - 1 = \mathbb{E}\left[G_k\right] - 1 \leq \frac{5}{T\Delta_i^2}$ for $k > L_i^{(\mathrm{hyb})}(T)$. □

Proof of Lemma A.9. On $E_1^{\mu(\mathrm{on})}(\tau_k+1)$ we have $\hat{\mu}_1^{(\mathrm{on})}(\tau_k+1) \leq y_i$, hence

$$\left\{\mathrm{median}\{\theta_1^{(\mathrm{on})}(\tau_k{+}1),\, \hat{\mu}_1^{(\mathrm{on})}(\tau_k{+}1),\, \theta_1^{(\mathrm{hyb})}(\tau_k{+}1)\} > y_i\right\} = \left\{\theta_1^{(\mathrm{on})}(\tau_k{+}1) > y_i,\, \theta_1^{(\mathrm{hyb})}(\tau_k{+}1) > y_i\right\}. \tag{37}$$

We denote the right-hand of (37) as

$$p_k^{\wedge} := \Pr\left(\theta_1^{(\mathrm{on})}(\tau_k{+}1) > y_i,\, \theta_1^{(\mathrm{hyb})}(\tau_k{+}1) > y_i \mid \mathcal{F}_{\tau_k}\right).$$

Conditioned on $\mathcal{F}_{\tau_k}$, draw i.i.d. copies

$$\left(\Theta_{k,\ell}^{(\mathrm{on})}, \Theta_{k,\ell}^{(\mathrm{hyb})}\right)_{\ell \geq 1} \overset{\text{i.i.d.}}{\sim} \left(\theta_1^{(\mathrm{on})}(\tau_k{+}1),\, \theta_1^{(\mathrm{hyb})}(\tau_k{+}1)\right) \mid \mathcal{F}_{\tau_k},$$

and define the (conditional) geometric hitting time

$$G_k := \min\left\{\ell \geq 1 :\; \Theta_{k,\ell}^{(\mathrm{on})} > y_i,\, \Theta_{k,\ell}^{(\mathrm{hyb})} > y_i\right\}.$$

Then $G_k \mid \mathcal{F}_k \sim \mathrm{Geom}(p_k^{\wedge})$ (supported on $\{1, 2, \dots\}$), so

$$\frac{1}{p_k^{\wedge}} = \mathbb{E}[G_k \mid \mathcal{F}_k].$$

Fix any $\alpha \in (1, \frac{3}{2})$ and let $\beta := \frac{\alpha}{\alpha-1}$. By Jensen's inequality (since $x \mapsto x^\alpha$ is convex for $\alpha > 1$),

$$\left(\frac{1}{p_k^\wedge}\right)^\alpha = \left(\mathbb{E}[G_k \mid \mathcal{F}_k]\right)^\alpha \le \mathbb{E}[G_k^\alpha \mid \mathcal{F}_k], \qquad \Rightarrow \qquad \mathbb{E}\Big[\Big(\frac{1}{p_k^\wedge}\Big)^\alpha\Big] \le \mathbb{E}[G_k^\alpha].$$

Next we bound $\mathbb{E}[G_k^\alpha]$ uniformly in k. For any integer $r \ge 1$,

$$\{G_k > r\} \subseteq \Big\{\max_{1\le\ell\le r} \Theta_{k,\ell}^{(\mathrm{on})} \le y_i\Big\} \cup \Big\{\max_{1\le\ell\le r} \Theta_{k,\ell}^{(\mathrm{hyb})} \le y_i\Big\}.$$

Recall in Lemma D.6 and Lemma A.7 where we prove a constant upper bound for all $k \ge 0$, for $r \ge r_0 = e^{64}$, we have

$$\Pr\left(\max_{1\le\ell\le r} \Theta_{k,\ell}^{(\mathrm{on})} \le y_i\right) \le \frac{1}{r^2} + \frac{1}{r^{3/2}}\,,$$
$$\Pr\left(\max_{1\le\ell\le r} \Theta_{k,\ell}^{(\mathrm{hyb})} \le y_i\right) \le \frac{1}{r^2} + \frac{1}{r^{3/2}}\,.$$

Therefore,

$$\Pr(G_k > r) \le \frac{2}{r^2} + \frac{2}{r^{3/2}} \le \frac{4}{r^{3/2}}.$$

For integer-valued $X \ge 1$, we first note that

$$(r+1)^\alpha - r^\alpha \le \alpha(r+1)^{\alpha-1} \le \alpha 2^{\alpha-1} r^{\alpha-1},$$

Then we have

$$\begin{aligned}
X^\alpha &= \sum_{r=1}^{X-1} \left((r+1)^\alpha - r^\alpha\right) + 1 \\
&\le \sum_{r=1}^{X} \left((r+1)^\alpha - r^\alpha\right) \\
&\le \alpha 2^{\alpha-1} \sum_{r=1}^{X} r^{\alpha-1} \\
&\le \alpha 2^{\alpha-1} \sum_{r=1}^{\infty} r^{\alpha-1} \mathbb{1}\{r \le X\} \\
&= \alpha 2^{\alpha-1} \sum_{r=1}^{\infty} r^{\alpha-1} \mathbb{1}\{X \ge r\}
\end{aligned}$$

Taking exception on both sides, let $C_\alpha = \alpha 2^{\alpha-1}$,we have

$$\mathbb{E}[X^\alpha] \le C_\alpha \sum_{r=1}^{\infty} r^{\alpha-1} \mathbb{P}(X \ge r),$$

hence with $X = G_k$,

$$\mathbb{E}[G_k^\alpha] \le C_\alpha \left(\sum_{r=1}^{r_0-1} r^{\alpha-1} + \sum_{r=r_0}^{\infty} r^{\alpha-1} \cdot \frac{4}{r^{3/2}}\right) = C_\alpha \left(C_{r_0} + 4\sum_{r=r_0}^{\infty} r^{\alpha-\frac{5}{2}}\right) \le M_\alpha < \infty,$$

where the last inequality uses $\alpha < \frac{3}{2}$ so that $\alpha - \frac{5}{2} < -1$, and we use C_{r_0} to denote $\sum_{r=1}^{r_0-1} r^{\alpha-1}$ as a constant.

Consequently,

$$\sup_{k\geq 0} \mathbb{E}\Big[\Big(\frac{1}{p_k^{\wedge}}\Big)^{\alpha}\Big] \leq M_\alpha < \infty. \tag{38}$$

By Hölder's inequality (Lemma D.3) with exponents (α, β),

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{p_{i,\tau_k+1}}\mathbb{1}\Big\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\Big\}\Big] &= \mathbb{E}\Big[\frac{1}{p_k^{\wedge}}\mathbb{1}\Big\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\Big\}\Big] \\ &\leq \Big(\mathbb{E}\Big[\Big(\frac{1}{p_k^{\wedge}}\Big)^{\alpha}\Big]\Big)^{1/\alpha} \cdot \Pr(\neg E_1^{\mu(\mathrm{on})}(\tau_k+1))^{1/\beta} \\ &\leq M_\alpha^{1/\alpha}\ \Pr(\neg E_1^{\mu(\mathrm{on})}(\tau_k+1))^{1/\beta}.\end{aligned}$$

Noting that, by approximation, taking $r_0 = e^{64}$, $(C_{r_0})^{1/\alpha} \approx e^{64}$.

Finally, since at time $\tau_k + 1$ the online empirical mean of arm 1 is computed from exactly k i.i.d. Gaussian samples with mean μ_1 and variance 1, we have the standard tail bound

$$\Pr(\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)) = \mathbb{P}\big(\hat{\mu}_{1,k}^{(\mathrm{on})} \leq y_i\big) \leq \exp\Big(-\frac{k(\mu_1 - y_i)^2}{2}\Big) = \exp\Big(-\frac{k\Delta_i^2}{18}\Big),$$

where we used $y_i = \mu_1 - \Delta_i/3$. Therefore,

$$\sum_{k=1}^{T-1} \mathbb{E}\Big[\frac{1}{p_{i,\tau_k+1}}\mathbb{1}\Big\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\Big\}\Big] \leq M_\alpha^{1/\alpha} \sum_{k=1}^{\infty} \exp\Big(-\frac{k\Delta_i^2}{18\beta}\Big) \leq \frac{C}{\Delta_i^2}, \tag{39}$$

for some constant $C > 0$ depending only on α (hence on β), which completes the bound for the $\mathbb{I}\{\neg E_1^{\mu(\mathrm{on})}(\tau_k+1)\}$-part of Lemma A.5. □

D. Useful Lemmas

Lemma D.1 (Chernoff-Hoeffding Bound)**.** *Let $X_1, \ldots, X_n$ be independent random variables in $[0,1]$ with $\mathbb{E}[X_i] = \mu_i$ (not necessarily equal). Let $X = \frac{1}{n}\sum_{i=1}^n X_i$, $\mu = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^n \mu_i$. Then for any $0 < \epsilon < 1 - \mu$,*

$$\Pr(X \geq \mu + \epsilon) \leq e^{-2n\epsilon^2},$$

and, for any $0 < \epsilon < \mu$,

$$\Pr(X \leq \mu - \epsilon) \leq e^{-2n\epsilon^2}.$$

Lemma D.2 (Concentration inequality of Gaussian variables (Abramowitz & Stegun, 1964))**.** *For a Gaussian distributed random variable Z with mean μ and variance σ^2, for any $x > 0$,*

$$\Pr(Z > \mu + x\sigma) \geq \frac{1}{\sqrt{2\pi}}\frac{x}{x^2+1}e^{-x^2/2}.$$

Lemma D.3 (Hölder's inequality)**.** *Let $p, q \in (1, \infty)$ be conjugate exponents such that $\frac{1}{p} + \frac{1}{q} = 1$. Let X and Y be real-valued random variables defined on the same probability space. If $\mathbb{E}[|X|^p] < \infty$ and $\mathbb{E}[|Y|^q] < \infty$, then*

$$\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}}(\mathbb{E}[|Y|^q])^{\frac{1}{q}}.$$

Lemma D.4 (Lemma 2.15 in (Agrawal & Goyal, 2017))**.** *For any sub-optimal arm $i \neq 1$,*

$$\sum_{t=1}^{T} \Pr(A(t) = i, \hat{\mu}_i^{(\mathrm{on})}(t) \leq x_i) \leq \frac{1}{d(x_i, y_i)} + 1 \leq \frac{9}{2\Delta_i^2} + 1,$$

where $d(a,b) = a\ln\frac{a}{b} + (1-a)\ln\frac{(1-a)}{(1-b)}$.

Lemma D.5 (Lemma 2.16 in (Agrawal & Goyal, 2017))**.** *For any sub-optimal arm $i \neq 1$,*

$$\sum_{t=1}^{T} \Pr(A(t) = i, \theta_i^{(\mathrm{on})}(t) > y_i, \hat{\mu}_i^{(\mathrm{on})}(t) \leq x_i) \leq L_i(T) + \frac{1}{\Delta_i^2},$$

where $L_i(T) \geq \frac{2\log(T\Delta_i^2)}{(y_i - x_i)^2}$.

Lemma D.6 (Lemma 2.13 in (Agrawal & Goyal, 2017))**.** *Let τ_j denote the time of the jth pull of the optimal arm 1. Then*

$$\mathbb{E}\left[\frac{1}{p_{i,\tau_j+1}^{(\mathrm{on})}} - 1\right] \leq \begin{cases} e^{64} + 5 & \forall j, \\ \frac{5}{T\Delta_i^2}, & j > L_i(T), \end{cases}$$

where $L_i^{(\mathrm{on})}(T) = \frac{288\log(T\Delta_i^2 + e^{32})}{\Delta_i^2}$.