OPT2023: 15th Annual Workshop on Optimization for Machine Learning

From 623514908081161688290923870 8 to 29: Vanilla Thompson Sampling Revisited

Bingshan Hu BINGSHA1@CS.UBC.CA
University of British Columbia, Canada
Tianyue H. Zhang TIANYUE.ZHANG@MILA.QUEBEC
Mila-Quebec AI Institute, Université de Montréal, Canada

Abstract

In this work, we derive a new problem-dependent regret bound for Thompson Sampling with Gaussian priors (Algorithm 2 in [1]), one of the classical stochastic bandit algorithms that has demonstrated excellent empirical performance and been widely deployed in real-world applications. The existing regret bound is $\sum_{i\in[K]:\Delta_i>0} \frac{288\left(e^{64}+6\right)\ln\left(T\Delta_i^2+e^{32}\right)}{\Delta_i} + \frac{10.5}{\Delta_i} + \Delta_i$, where $[K]$ denotes the arm set, Δ_i denotes the single round performance loss when pulling a sub-optimal arm i instead of the optimal arm, and T is the time horizon. Since real-world learning tasks care about learning algorithms' performance when T is finite, the existing regret bound is only non-vacuous when $T > 288 \cdot e^{64}$, which may not be practical. Our new regret bound is $\sum_{i\in[K]:\Delta_i>0} \frac{1252\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i} + \frac{18\ln\left(T\Delta_i^2\right)}{\Delta_i} + \frac{182.5}{\Delta_i} + \Delta_i$, which tightens the leading term's coefficient significantly. Despite having made some improvements, we would like to emphasize that the goal of this work is to deepen the understanding of Thompson Sampling from a theoretical perspective to unlock the full potential of this classical learning algorithm in order to solve challenging real-world learning problems.

1. Introduction

We study the learning problem of stochastic multi-armed bandits (MAB) specified by $(K; p_1, \ldots, p_K)$, where K is the number of arms and p_i is the reward distribution associated with arm i with its mean denoted as μ_i. In this learning problem, a player chooses an arm to pull in each round $t = 1, 2, \ldots, T$ without knowledge of the reward distributions. At the end of the round, the player obtains and observes a reward that is drawn from the reward distribution associated with the pulled arm. The goal of the player is to choose arms sequentially to maximize the cumulative reward over T rounds. Since only the pulled arm has the chance to be observed at the end of each round, the main challenge for solving bandit problems is the balance between exploitation and exploration in each round. Exploitation involves pulling the arms that have the potential to produce high rewards based on past experience, whereas exploration involves pulling the arms that can help the player to better learn the reward distributions or the means of the reward distributions.

Many applications that require the balance between exploitation and exploration can be framed as bandit learning problems. In healthcare, bandit algorithms can be used for clinical trials where exploitation refers to prescribing known effective treatments and exploration refers to prescribing new medicine or procedures to discover potentially more effective or personalized treatments. In content recommendation or online advertising systems, bandit algorithms can help to decide whether to show users familiar content (exploitation) or to introduce users to new and potentially

more interesting content (exploration) to optimize overall satisfaction. In inventory management, bandit algorithms can be used to decide whether to reorder products that have sold well historically (exploitation) or to switch to new products or suppliers (exploration) to maximize overall profit.

Successful stochastic bandit algorithms such as the Upper Confidence Bound (UCB)-based [2, 3, 6, 11, 13] and the Thompson Sampling-based [1, 4, 7–10, 12] have been extensively studied in literature. The key difference between the UCB-based and the Thompson Sampling-based algorithms lies in the exploration mechanism. In UCB-based algorithms, the exploration is driven by adding a deterministic bonus to the empirical estimates, whereas in Thompson Sampling-based algorithms, the exploration is achieved by injecting random noise into the empirical estimates.

The problem-dependent regret bound with a finite time horizon T is of great interest as real-world applications cannot run learning algorithms infinitely. As implied by the name, a problem-dependent regret bound takes the problem instance $(K; p_1, p_2, \ldots, p_K)$ (or $(K; \mu_1, \mu_2, \ldots, \mu_K)$) into account and usually takes the $\sum_{i:\Delta_i>0} O\left(\frac{\ln(T)}{\Delta_i}\right)$ form [1–4, 7, 9, 13], where Δ_i denotes the single round performance loss when pulling a sub-optimal arm i instead of the optimal arm, and the big-Oh notation hides a universal constant. For the aforementioned UCB-based learning algorithms, the hidden constants are in a reasonable range depending on the added bonus, for example, the constant in the big-Oh notation for UCB1 [3] is 8.

As empirically studied in [5], Thompson Sampling performs excellently for practical learning tasks. We revisit one of the two original versions of Thompson Sampling for stochastic bandits with bounded rewards, Thompson Sampling with Gaussian Priors (Algorithm 2) in [1]. For ease of presentation, we rename it as *Vanilla Thompson Sampling*.[1] The key idea of Thompson Sampling is to maintain a posterior distribution (a Gaussian distribution) to model the mean μ_i of the reward distribution p_i. The exploitation component is from the mean of the posterior distribution while the spread of the posterior distribution contributes to the exploration. Although learning algorithms from [4, 7–10, 12] are all Thompson Sampling-based, they either use Beta priors (Beta distributions to model μ_i) [12] or design new distributions to model μ_i [4, 7–10]. None of them revisits Vanilla Thompson Sampling.

Now, we review the existing problem-dependent regret bound and discuss why it may not be enough for some real-world applications.

Existing regret bound derived in [1]. The regret bound of Vanilla Thompson Sampling is

$$\sum_{i:\Delta_i>0} \frac{288\left(e^{64}+6\right)\ln\left(T\Delta_i^2+e^{32}\right)}{\Delta_i} + \frac{10.5}{\Delta_i} + \Delta_i \quad . \tag{1}$$

Since the coefficient for the leading term is at least $288 \cdot e^{64} \approx 1.8 \times 10^{30}$, this regret bound is vacuous for learning problems when $T \leq 288 \cdot e^{64}$. Note that when $T \leq 288 \cdot e^{64}$, the regret is at most T, and thus, the existing regret bound does not take the bandit problem instance $(K; \mu_1, \mu_2, \ldots, \mu_K)$ into account due to the extremely large coefficient for the $\ln(T)$ term. Therefore, we are motivated to derive a new regret bound with a more acceptable coefficient for the $\ln(T)$ leading term.

1. The other original version is to use Beta distributions to model the mean rewards and the problem-dependent regret bounds are derived in [1, 12].

Preview of our new bound (Theorem 1). The regret bound of Vanilla Thompson Sampling is

$$\sum_{i:\Delta_i>0} \frac{1252\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i}+\frac{18\ln\left(T\Delta_i^2\right)}{\Delta_i}+\frac{182.5}{\Delta_i}+\Delta_i \quad . \tag{2}$$

Despite the fact the coefficient is significantly improved, we would like to emphasize that the purpose of this work is not to find the optimal coefficient. Instead, we intend to answer the following fundamental question in Vanilla Thompson Sampling.

When the posterior distribution of the optimal arm is not concentrated, that is, the optimal arm has not been sufficiently observed, what is the the expected number of rounds needed before the optimal arm has a good posterior sample?

Intuitively, this quantity indicates the concentration speed of optimal arm's posterior distribution. After the optimal arm's posterior distribution is concentrated, the player is unlikely to pull a sub-optimal arm. In Lemma 2.13 of [1], the derived upper bound on the expected number of rounds needed is e^{64}. As will be shown in Lemma 2, our new upper bound is 29.

2. Stochastic Bandit Problems

Consider a classical stochastic bandit problem where we have an arm set $[K]$ with size K and each arm $i \in [K]$ is associated with a fixed but unknown reward distribution p_i with $[0,1]$ support. Let μ_i denote the mean of distribution p_i. Without loss of generality, we assume that the first arm is the unique optimal arm. In other words, we assume $\mu_1 > \mu_i$ for all $i \neq 1$. In each round $t = 1, 2, \ldots, T$, the player pulls an arm $i_t \in [K]$ and receives a reward $X_{i_t}(t) \sim p_{i_t}$. The goal of the player is to pull arms sequentially to maximize the cumulative reward, or equivalently, minimize the *(cumulative)-regret*, defined as

$$\mathcal{R}(T) \quad = \quad T\cdot\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T}\mu_{i_t}\right] \quad , \tag{3}$$

where the expectation is taken over i_t. The regret measures the cumulative performance loss between always pulling the optimal arm and the player's choices of which arms to pull.

Notation. Let $n_i(t)$ denote the number of pulls of arm i by the end of round t and $\hat{\mu}_{i,n_i(t)}(t)$ denote the empirical mean of arm i by the end of round t. For ease of presentation, we write $\hat{\mu}_{i,n_i(t)}$ for short. Let $\mathcal{N}\left(\mu,\sigma^2\right)$ denote a Gaussian distribution with μ as the mean parameter and σ^2 as the variance parameter.

3. Vanilla Thompson Sampling

Vanilla Thompson Sampling is described in Algorithm 1 below.[2] The core idea of it is to maintain a posterior distribution $\mathcal{N}\left(\hat{\mu}_{i,n_i(t-1)},\frac{1}{n_i(t-1)}\right)$ to model the mean μ_i for each arm i and use a random

2. Only for the purpose of practical implementation only, we make minor modifications as compared to the original version presented in [1]. We add an initialization phase to initialize the empirical mean of each arm. However, it does not change the algorithm fundamentally. More details can be found in Appendix A.

sample $\theta_i(t) \sim \mathcal{N}\left(\hat{\mu}_{i,n_i(t-1)}, \frac{1}{n_i(t-1)}\right)$ in the learning. With all posterior samples $\theta_i(t)$ in hand, the player is safe to behave greedily and pull the arm with the highest posterior sample. It is important to note that the posterior sample $\theta_i(t)$ already takes into account the exploitation-exploration balance. As mentioned in Section 1, the mean of the posterior distribution, $\hat{\mu}_{i,n_i(t-1)}$, is for the exploitation purpose, and, the variance of the posterior distribution, $\frac{1}{n_i(t-1)}$, controls the level of exploration.

Algorithm 1 Thompson Sampling with Gaussian Priors [1]

1: **Initialization:** for each $i \in [K]$: pull it once to initialize n_i and the empirical mean $\hat{\mu}_{i,n_i}$
2: **for** $t = K+1, K+2, \cdots$ **do**
3: Draw $\theta_i(t) \sim \mathcal{N}\left(\hat{\mu}_{i,n_i}, \frac{1}{n_i}\right)$ for all $i \in [K]$
4: Pull arm $i_t \in \arg\max_{i\in[K]} \theta_i(t)$ and observe $X_{i_t}(t)$
5: Set $n_{i_t} \leftarrow n_{i_t} + 1$ and update the empirical mean $\hat{\mu}_{i_t,n_{i_t}}$ of the pulled arm i_t accordingly.
6: **end for**

Theorem 1 *The regret of Algorithm 1 is at most*

$$\sum_{i\in\mathcal{A}:\Delta_i>0} \frac{1252\ln\left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)}{\Delta_i} + \frac{18\ln\left(T\Delta_i^2\right)}{\Delta_i} + \frac{182.5}{\Delta_i} + \Delta_i \quad , \tag{4}$$

where $\Delta_i := \mu_1 - \mu_i$ denotes the single round performance loss when pulling a sub-optimal arm i.

The full proof for Theorem 1 is deferred to Appendix C. The improvement of our regret bound mainly comes from Lemma 2 below. It answers the question raised at the end of Section 1 and significantly improves the results shown in Lemma 2.13 in [1]. Note that in Lemma 2.13 of [1], the RHS of (5) below is e^{64}. Let $\mathcal{F}_t = \{i_\tau, X_{i_\tau}(\tau), \tau = 1, \ldots, t\}$ collect all the history information by the round of round t consisting of the pulled arms and their associated rewards.

Lemma 2 *Let $\tau_s^{(1)}$ be the round when the s-th pull of the optimal arm 1 occurs and $\theta_{1,s} \sim \mathcal{N}\left(\hat{\mu}_{1,s}, \frac{1}{s}\right)$. Then, for any integer $s \geq 1$, we have*

$$\mathbb{E}_{\mathcal{F}_{\tau_s^{(1)}}}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} | \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \leq 29 \quad . \tag{5}$$

Also, for any integer $s \geq L_{1,i} := \frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2 \ln\left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)}{\Delta_i^2}$, we have

$$\mathbb{E}_{\mathcal{F}_{\tau_s^{(1)}}}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} | \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \leq \frac{180}{T\Delta_i^2} \quad . \tag{6}$$

Discussion. If event $\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2}$ occurs, we can view the optimal arm has a good posterior sample as compared to the sub-optimal arm i. Informally, a good posterior sample of the optimal arm indicates the pull of the optimal arm. The pulls of the optimal arm contribute to reducing the variance of the posterior distribution and the width of the confidence intervals. The value of $\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{\tau_s^{(1)}} = F_{\tau_s^{(1)}}\right\}$ is the probability of the event that the optimal arm has a good posterior sample occurs given the history information. $\mathbb{E}\left[1/\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{\tau_s^{(1)}}\right\} - 1\right]$ quantifies the expected number of independent draws needed before the optimal arm has a good posterior sample. The first result shown in Lemma 2 says that even if the optimal arm has not been observed enough, i.e., $s < L_{1,i}$, it takes at most 29 draws in expectation before the optimal arm has a good posterior sample. That is also to say, it takes at most 29 rounds in expectation before the next pull of the optimal arm. The second result says that after the optimal arm has been observed enough, i.e., $s \geq L_{1,i}$, the expected number of draws before the optimal arm has a good posterior sample is in the order of $\frac{1}{T\Delta_i^2}$. That is also to say, after the optimal arm has been observed enough, it will be pulled very frequently.

4. Conclusion and Future Work

In this paper, we have revisited Vanilla Thompson Sampling and derived a new problem-dependent regret bound that significantly improves the existing regret bound. The next step is to further tighten the coefficient of the leading term using numerical optimization.

References

[1] Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson Sampling. http://www.columbia.edu/~sa3305/papers/j3-corrected.pdf, 2017.

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[4] Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pages 54–72. PMLR, 2022.

[5] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. *Advances in neural information processing systems*, 24, 2011.

[6] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[7] Bingshan Hu, Tianyue H Zhang, Nidhi Hegde, and Mark Schmidt. Optimistic Thompson sampling-based algorithms for episodic reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 890–899. PMLR, 2023.

[8] Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. MOTS: Minimax optimal Thompson Sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

[9] Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Anima Anandkumar. Finite-time regret of Thompson Sampling algorithms for exponential family multi-armed bandits. *Advances in Neural Information Processing Systems*, 35:38475–38487, 2022.

[10] Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson Sampling with less exploration is fast and optimal. *International Conference on Machine Learning*, 2023.

[11] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600. PMLR, 2012.

[12] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

[13] Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

Appendix A. Posterior Distribution Computation of Algorithm 1

Let $\mathcal{N}\left(x \mid \mu, \sigma^2\right) := \frac{1}{\sigma\sqrt{2\pi}} 2^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, where $x \in \mathbb{R}$. Our Algorithm 1 also strictly follows the template of Thompson Sampling described in [1, 5].

1. A set ψ of parameters $\tilde{\mu}_i$. Here, set ψ collects all the real numbers, and thus, the true parameter $\mu_i \in \psi$.
2. An assumed prior distribution $p(\tilde{\mu}_i)$ on these parameters collected in ψ. Here, conceptually, we assume the prior distribution is a zero-mean Gaussian distribution with an infinite variance. Indeed, it captures the fact that the player has no belief in the learning problems before seeing any evidence.
3. An assumed reward likelihood function $p(x_i \mid \tilde{\mu}_i) = \mathcal{N}\left(x_i \mid \tilde{\mu}_i, 1\right)$.
4. Past observations $\mathcal{D}_i$ consisting of all the observed rewards of this arm i. The likelihood function for all the data in $\mathcal{D}_i$ has the form $p(\mathcal{D}_i \mid \tilde{\mu}_i) \propto \mathcal{N}\left(\hat{\mu}_{i,n_i} \mid \tilde{\mu}_i, \frac{1}{n_i}\right)$, where n_i is the number of observations in $\mathcal{D}_i$ and $\hat{\mu}_{i,n_i}$ is the empirical average of these n_i observations.
5. A posterior distribution $p(\tilde{\mu}_i \mid \mathcal{D}_i) \propto p(\mathcal{D}_i \mid \tilde{\mu}_i) \cdot p(\tilde{\mu}_i) \propto \mathcal{N}\left(\tilde{\mu}_i \mid \hat{\mu}_{i,n_i}, \frac{1}{n_i}\right)$.

To implement Algorithm 1, we do not need to use a Gaussian distribution with an infinite variance as the prior distribution. Instead, we can pull each arm once to initialize $n_i = 1$ and the empirical mean $\hat{\mu}_{i,n_i}$. Note that $\mathcal{N}\left(\hat{\mu}_{i,n_i}, n_i\right)$ is the posterior distribution when arm i only has one observation.

Appendix B. Useful Facts

Fact 3 *Let $X_1, X_2, \ldots, X_n$ be independent random variables with support $[0,1]$. Let $\mu_{1:n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $a > 0$, we have*

$$\mathbb{P}\left\{\left|\mu_{1:n} - \mathbb{E}\left[\mu_{1:n}\right]\right| \geq a\right\} \leq 2e^{-2na^2} \quad . \tag{7}$$

Fact 4 *(Concentration and anti-concentration bounds of Gaussian distributions). For a Gaussian distributed random variable Z with mean μ and variance σ^2, for any $z > 0$, we have*

$$\mathbb{P}\left\{Z > \mu + z\sigma\right\} \leq \frac{1}{2}e^{-\frac{z^2}{2}}, \quad \mathbb{P}\left\{Z < \mu - z\sigma\right\} \leq \frac{1}{2}e^{-\frac{z^2}{2}} \quad , \tag{8}$$

and

$$\mathbb{P}\left\{Z > \mu + z\sigma\right\} \geq \frac{1}{\sqrt{2\pi}}\frac{z}{z^2+1}e^{-\frac{z^2}{2}} \quad . \tag{9}$$

Appendix C. Proofs

Proof of Theorem 1: For a sub-optimal arm i such that $\Delta_i \leq \sqrt{\frac{1}{T}}$, we have the total regret of pulling this sub-optimal arm i over T rounds is at most $T\Delta_i \leq \sqrt{T} \leq \frac{1}{\Delta_i}$. For a sub-optimal arm i such that $\Delta_i > \sqrt{\frac{1}{T}}$, we upper bound its expected number of pulls $\mathbb{E}\left[n_i(T)\right]$ by the end of round T.

Let $\bar{\mu}_{i,n_i(t-1)} := \hat{\mu}_{i,n_i(t-1)} + \sqrt{\frac{2\ln\left(T\Delta_i^2\right)}{n_i(t-1)}}$ denote the upper confidence bound.

Let $L_i := \left\lceil \frac{4\left(\sqrt{0.5}+\sqrt{2}\right)^2 \ln\left(T\Delta_i^2\right)}{\Delta_i^2} \right\rceil$.

To decompose the regret, we define $\mathcal{E}_i^{\mu}(t-1) := \left\{ \left|\hat{\mu}_{i,n_i(t-1)} - \mu_i\right| \leq \sqrt{\frac{0.5\ln\left(T\Delta_i^2\right)}{n_i(t-1)}} \right\}$ and $\mathcal{E}_i^{\theta}(t) := \left\{\theta_i(t) \leq \bar{\mu}_{i,n_i(t-1)}\right\}$. Then, we have

$$
\begin{aligned}
& \mathbb{E}\left[n_i(T)\right] \\
\leq\ & L_i + \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, n_i(t-1) \geq L_i\right\}\right] \\
\leq\ & L_i + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \mathcal{E}_i^{\theta}(t), \mathcal{E}_i^{\mu}(t-1), n_i(t-1) \geq L_i\right\}\right]}_{=:\omega_1} \\
+\ & \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \overline{\mathcal{E}_i^{\theta}(t)}, n_i(t-1) \geq L_i\right\}\right]}_{=:\omega_2} + \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \overline{\mathcal{E}_i^{\mu}(t-1)}, n_i(t-1) \geq L_i\right\}\right]}_{=:\omega_3}.
\end{aligned}
\tag{10}
$$

Term ω_1 is very similar to Lemma 2.14 in [1], which is challenging to derive a tighter upper bound. Terms ω_2 (similar to Lemma 2.16 in [1]) and ω_3 (similar to Lemma 2.15 in [1]) are not difficult to bound. We use Gaussian concentration inequality (Fact 4) for upper bounding term ω_2 and Hoeffding's inequality (Fact 3) for upper bounding terms ω_3. From Lemma 6 and Lemma 7 we have $\omega_2 \leq \frac{0.5}{\Delta_i^2}$ and $\omega_3 \leq \frac{2}{\Delta_i^2}$.

For term ω_1, similar to Lemma 2.8 in [1], we have our Lemma 5, a lemma that links the probability of pulling a sub-optimal arm i to the probability of pulling the optimal arm 1 in round t by introducing the upper confidence bound $\bar{\mu}_{i,n_i(t-1)}$. Note that both events $\mathcal{E}_i^{\mu}(t-1)$ and $n_i(t-1) \geq L_i$ are determined by the history information. The value of $\bar{\mu}_{i,n_i(t-1)}$ is determined by the history information. Also, the distributions for $\theta_j(t)$ for all $j \in [K]$ are determined by the history information.

Lemma 5 *For any instantiation F_{t-1} of $\mathcal{F}_{t-1}$, we have*

$$
\mathbb{E}\left[\mathbf{1}\left\{i_t = i, \mathcal{E}_i^{\theta}(t)\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right] \leq \frac{\mathbb{P}\left\{\theta_1(t) \leq \bar{\mu}_{i,n_i(t-1)} | \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \bar{\mu}_{i,n_i(t-1)} | \mathcal{F}_{t-1} = F_{t-1}\right\}} \mathbb{E}\left[\mathbf{1}\left\{i_t = 1, \mathcal{E}_i^{\theta}(t)\right\} \mid \mathcal{F}_{t-1} = F_{t-1}\right].
$$

With Lemma 5 in hand, now, we are ready to upper bound term ω_1. We have

$$
\begin{aligned}
\omega_1 &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \mathcal{E}_i^{\theta}(t), \mathcal{E}_i^{\mu}(t-1), n_i(t-1) \geq L_i\right\}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_i^{\mu}(t-1), n_i(t-1) \geq L_i\right\} \cdot \underbrace{\mathbb{E}\left[\mathbf{1}\left\{i_t = i, \mathcal{E}_i^{\theta}(t)\right\} \mid \mathcal{F}_{t-1}\right]}_{\text{LHS in Lemma 5}}\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\Bigg[\underbrace{\mathbf{1}\left\{\mathcal{E}_i^{\mu}(t-1), n_i(t-1) \geq L_i\right\} \frac{\mathbb{P}\left\{\theta_1(t) \leq \bar{\mu}_{i,n_i(t-1)} \mid \mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \bar{\mu}_{i,n_i(t-1)} \mid \mathcal{F}_{t-1}\right\}} \underbrace{\mathbb{E}\left[\mathbf{1}\left\{i_t = 1, \mathcal{E}_i^{\theta}(t)\right\} \mid \mathcal{F}_{t-1}\right]}_{\text{RHS in Lemma 5}}}_{\eta}\Bigg] \\
&\leq^{(a)} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1(t) \leq \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1}\right\}} \cdot \mathbf{1}\left\{i_t = 1\right\} \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1(t) \leq \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1}\right\}} \cdot \mathbf{1}\left\{i_t = 1\right\}\right] \quad .
\end{aligned}
\tag{11}
$$

Inequality (a) in (11) uses the argument that for a specific F_{t-1} of $\mathcal{F}_{t-1}$ such that either event $\mathcal{E}_i^{\mu}(t-1)$ or $n_i(t-1) \geq L_i$ does not occur, term η in (11) will be 0. Note that for any F_{t-1}, we have $1 < \frac{1}{\mathbb{P}\left\{\theta_1(t) > \bar{\mu}_{i,n_i(t-1)} \mid \mathcal{F}_{t-1} = F_{t-1}\right\}} < +\infty$. Recall $L_i = \left\lceil \frac{4\left(\sqrt{0.5}+\sqrt{2}\right)^2 \ln\left(T\Delta_i^2\right)}{\Delta_i^2} \right\rceil$. For any specific F_{t-1} such that both events $\mathcal{E}_i^{\mu}(t-1)$ and $n_i(t-1) \geq L_i$ occur, we have $\bar{\mu}_{i,n_i(t-1)} = \hat{\mu}_{i,n_i(t-1)} + \sqrt{\frac{2\ln\left(T\Delta_i^2\right)}{n_i(t-1)}} \leq \mu_i + \sqrt{\frac{0.5\ln\left(T\Delta_i^2\right)}{n_i(t-1)}} + \sqrt{\frac{2\ln\left(T\Delta_i^2\right)}{n_i(t-1)}} \leq \mu_i + \sqrt{\frac{0.5\ln\left(T\Delta_i^2\right)}{L_i}} + \sqrt{\frac{2\ln\left(T\Delta_i^2\right)}{L_i}} \leq \mu_i + \frac{\Delta_i}{2} = \mu_1 - \frac{\Delta_i}{2}$, which implies $\frac{\mathbb{P}\left\{\theta_1(t) \leq \bar{\mu}_{i,n_i(t-1)} \mid \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \bar{\mu}_{i,n_i(t-1)} \mid \mathcal{F}_{t-1} = F_{t-1}\right\}} \leq \frac{\mathbb{P}\left\{\theta_1(t) \leq \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1} = F_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t) > \mu_1 - \frac{\Delta_i}{2} \mid \mathcal{F}_{t-1} = F_{t-1}\right\}}$.

Now, we partition all T rounds into multiple intervals based on the arrivals of the observations of the optimal arm 1. Let $\tau_s^{(1)}$ be the round when arm 1 is pulled for the s-th time. Note that this partition in time horizon ensures that the posterior distribution of arm 1 stays the same among all the rounds when $t \in \left\{\tau_s^{(1)} + 1, \ldots, \tau_{s+1}^{(1)}\right\}$. Let $L_{1,i} := \frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2 \ln\left(T\Delta_i^2\right)}{\Delta_i^2}$ and $\theta_{1,s} \sim \mathcal{N}\left(\hat{\mu}_{1,s}, \frac{1}{s}\right)$. Then, we have

$$
\begin{aligned}
\omega_1 &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1(t)\leq \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t)> \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{t-1}\right\}}\cdot \mathbf{1}\left\{i_t=1\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{T}\sum_{t=\tau_s^{(1)}+1}^{\tau_{s+1}^{(1)}} \frac{\mathbb{P}\left\{\theta_1(t)\leq \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{t-1}\right\}}{\mathbb{P}\left\{\theta_1(t)> \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{t-1}\right\}}\cdot \mathbf{1}\left\{i_t=1\right\}\right] \\
&\leq \sum_{s=1}^{T}\mathbb{E}\left[\frac{\mathbb{P}\left\{\theta_1\left(\tau_{s+1}^{(1)}\right)\leq \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_{s+1}^{(1)}-1}\right\}}{\mathbb{P}\left\{\theta_1\left(\tau_{s+1}^{(1)}\right)> \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_{s+1}^{(1)}-1}\right\}}\right] \\
&\leq \sum_{s=1}^{L_{1,i}}\mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s}> \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_s^{(1)}}\right\}}-1\right]+\sum_{s=L_{1,i}+1}^{T}\mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s}> \mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_s^{(1)}}\right\}}-1\right] \quad .
\end{aligned} \tag{12}
$$

With Lemma 2 in hand, we have

$$
\omega_1 \leq 29\cdot L_{1,i}+\sum_{s=1}^{T}\frac{180}{T\Delta_i^2} \leq \frac{116\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i^2}+\frac{180}{\Delta_i^2} \quad . \tag{13}
$$

Plugging the upper bounds of ω_1, ω_2, and ω_3 together in (10), we have

$$
\begin{aligned}
\mathbb{E}\left[n_i(T)\right] &\leq L_i+\omega_1+\omega_2+\omega_3 \\
&\leq \frac{4\left(\sqrt{0.5}+\sqrt{2}\right)^2\ln\left(T\Delta_i^2\right)}{\Delta_i^2}+1+\frac{116\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i^2}+\frac{180}{\Delta_i^2}+\frac{0.5}{\Delta_i^2}+\frac{2}{\Delta_i^2} \\
&\leq \frac{18\ln\left(T\Delta_i^2\right)}{\Delta_i^2}+\frac{1252\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i^2}+\frac{182.5}{\Delta_i^2}+1 \quad .
\end{aligned} \tag{14}
$$

■

Proof of Lemma 5: The proof is very similar to the proof of Lemma 2.8 in [1]. Note that the value of $\bar{\mu}_{i,n_i(t-1)}$ is determined by the history information. We have the following two pieces of arguments. The first piece is

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left\{i_t=i,\mathcal{E}_i^{\theta}(t)\right\}\mid \mathcal{F}_{t-1}=F_{t-1}\right] \\
\leq\ &\mathbb{E}\left[\mathbf{1}\left\{\theta_j(t)\leq \bar{\mu}_{i,n_i(t-1)},\forall j\in[K]\right\}\mid \mathcal{F}_{t-1}=F_{t-1}\right] \\
=\ &\mathbb{P}\left\{\theta_1(t)\leq \bar{\mu}_{i,n_i(t-1)}\mid \mathcal{F}_{t-1}=F_{t-1}\right\}\underbrace{\mathbb{P}\left\{\theta_j(t)\leq \bar{\mu}_{i,n_i(t-1)},\forall j\in[K]\setminus\{1\}\mid \mathcal{F}_{t-1}=F_{t-1}\right\}}_{>0}.
\end{aligned} \tag{15}
$$

The second piece is

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left\{i_t=1,\mathcal{E}_i^{\theta}(t)\right\}\mid \mathcal{F}_{t-1}=F_{t-1}\right] \\
\geq\ &\mathbb{E}\left[\mathbf{1}\left\{\theta_1(t)>\bar{\mu}_{i,n_i(t-1)}\geq \theta_j(t),\forall j\in[K]\setminus\{1\}\right\}\mid \mathcal{F}_{t-1}=F_{t-1}\right] \\
=\ &\underbrace{\mathbb{P}\left\{\theta_1(t)>\bar{\mu}_{i,n_i(t-1)}\mid \mathcal{F}_{t-1}=F_{t-1}\right\}}_{>0}\underbrace{\mathbb{P}\left\{\theta_j(t)\leq \bar{\mu}_{i,n_i(t-1)},\forall j\in[K]\setminus\{1\}\mid \mathcal{F}_{t-1}=F_{t-1}\right\}}_{>0}.
\end{aligned} \tag{16}
$$

Combining (15) and (16) concludes the proof. ■

Proof of Lemma 2: We have two results stated in Lemma 2. The main purpose of the first result is to show that even if the optimal arm's posterior distribution has high variance, that is, it has not been observed enough, after a constant number of rounds in expectation, the optimal arm will have a good posterior sample. This result further implies that the total regret in expectation between two consecutive pulls of the optimal arm is also a constant. The second result states that after the optimal arm has been pulled enough, the expected number of rounds needed before the optimal arm having a good posterior sample is very small, in the order of $1/(T\Delta_i^2)$. This result implies that the total regret in expectation between two consecutive pulls of the optimal arm is in the order of $1/(T\Delta_i)$.

Similar to the proof of Lemma 2.13 in [1], we introduce a geometric random variable in the proof. For any integer $s \geq 1$, we let $\mathcal{G}_{1,s}$ be a geometric random variable denoting the number of consecutive independent trials (including the trial) until event $\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2}$ occurs. Then, we have

$$\mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s}>\mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_s^{(1)}}\right\}}\right] = \mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s}>\mu_1-\frac{\Delta_i}{2}|\hat{\mu}_{1,s}\right\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{G}_{1,s} \mid \hat{\mu}_{1,s}\right]\right] = \mathbb{E}\left[\mathcal{G}_{1,s}\right]. \tag{17}$$

To upper bound $\mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s}>\mu_1-\frac{\Delta_i}{2}|\mathcal{F}_{\tau_s^{(1)}}\right\}}\right]$, it is sufficient to upper bound $\mathbb{E}\left[\mathcal{G}_{1,s}\right]$. To upper bound $\mathbb{E}\left[\mathcal{G}_{1,s}\right]$, we use the definition of expectation when the random variable has non-negative integers as support. We have

$$\mathbb{E}\left[\mathcal{G}_{1,s}\right] = \sum_{r=0}^{\infty}\mathbb{P}\left\{\mathcal{G}_{1,s}>r\right\} = \sum_{r=0}^{\infty}\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s}>r \mid \hat{\mu}_{1,s}\right\}\right] \quad . \tag{18}$$

For any $s \geq 1$, we claim

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s}>r \mid \hat{\mu}_{1,s}\right\}\right] \leq \begin{cases} 1, & r \in [0,12] \quad , \\ e^{-\sqrt{\frac{r}{\pi}}\cdot\frac{\ln(13)}{\ln(13)+2}}+\frac{1}{r}, & r \in [13,100] \quad , \\ e^{-\sqrt{\frac{r}{3\pi}}}+r^{-\frac{4}{3}}, & r \geq 101 \quad . \end{cases} \tag{19}$$

For any $s \geq L_{1,i} = \frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i^2}$, we claim

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s}>r \mid \hat{\mu}_{1,s}\right\}\right] \leq \begin{cases} 1, & r=0 \quad , \\ \frac{1}{r^2\left(T\Delta_i^2\right)}+\frac{0.5^r}{\left(T\Delta_i^2\right)^r}, & r \in \left[1, \left\lfloor\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3\right\rfloor\right] \quad , \\ e^{-\sqrt{\frac{r}{3\pi}}}+r^{-\frac{4}{3}}, & r \geq \left\lfloor\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3\right\rfloor+1 \quad . \end{cases} \tag{20}$$

The proofs for the results shown in (19) and (20) are deferred to the end of this session.

With (19) in hand, for any fixed integer $s \geq 1$, we have

$$
\begin{aligned}
& \mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} | \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \\
= \;& \mathbb{E}\left[\mathcal{G}_{1,s}\right] - 1 \\
= \;& \sum_{r=0}^{\infty} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] - 1 \\
= \;& \sum_{r=0}^{12} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] + \sum_{r=13}^{100} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] + \sum_{r=101}^{\infty} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] - 1 \\
\leq \;& 13 + \sum_{r=13}^{100} \left(e^{-\sqrt{\frac{r}{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}} + \frac{1}{r}\right) + \sum_{r=101}^{\infty} \left(e^{-\sqrt{\frac{r}{3\pi}}} + \frac{1}{r^{\frac{4}{3}}}\right) - 1 \\
\leq \;& 12 + \int_{12}^{100} \left(e^{-\sqrt{\frac{r}{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}} + \frac{1}{r}\right) dr + \int_{100}^{\infty} \left(e^{-\sqrt{\frac{r}{3\pi}}} + \frac{1}{r^{\frac{4}{3}}}\right) dr \\
\leq \;& 12 + 10.44 + 2.13 + 3.1 + 0.65 \\
\leq \;& 29 \quad ,
\end{aligned}
\tag{21}
$$

which concludes the proof for the first stated result in Lemma 2.

With (20), for any fixed integer $s \geq L_{1,i}$, we have

$$
\begin{aligned}
& \mathbb{E}\left[\frac{1}{\mathbb{P}\left\{\theta_{1,s} > \mu_1 - \frac{\Delta_i}{2} | \mathcal{F}_{\tau_s^{(1)}}\right\}} - 1\right] \\
= \;& \sum_{r=0}^{\infty} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] - 1 \\
\leq \;& 1 + \sum_{r=1}^{\left\lfloor \left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)^3 \right\rfloor} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] + \sum_{r=\left\lfloor \left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)^3 \right\rfloor + 1}^{\infty} \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] - 1 \\
\leq \;& \sum_{r=1}^{\left\lfloor \left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)^3 \right\rfloor} \left(\frac{1}{r^2\left(T\Delta_i^2\right)} + \frac{0.5^r}{\left(T\Delta_i^2\right)^r}\right) + \sum_{r=\left\lfloor \left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)^3 \right\rfloor + 1}^{\infty} \left(e^{-\sqrt{\frac{r}{3\pi}}} + r^{-\frac{4}{3}}\right) \\
\leq \;& \frac{1}{T\Delta_i^2} + \frac{0.5}{T\Delta_i^2} + \int_1^{+\infty} \left(\frac{1}{r^2\left(T\Delta_i^2\right)} + \frac{1}{\left(2T\Delta_i^2\right)^r}\right) dr + \int_{(T\Delta_i^2)^2}^{\infty} e^{-\frac{\sqrt{r}}{\sqrt{3\pi}}} dr + \int_{(T\Delta_i^2)^3}^{\infty} r^{-\frac{4}{3}} dr \\
\leq^{(a)} \;& \frac{1}{T\Delta_i^2} + \frac{0.5}{T\Delta_i^2} + \frac{1}{T\Delta_i^2} + \frac{0.5}{T\Delta_i^2} + \frac{6 \cdot 3\pi \cdot \sqrt{3\pi}}{T\Delta_i^2} + \frac{3}{T\Delta_i^2} \\
\leq \;& \frac{180}{T\Delta_i^2} \quad ,
\end{aligned}
\tag{22}
$$

which concludes the proof for the second stated result in Lemma 2. Inequality (a) uses the fact that, for any $a, b > 0$, we have $\int_b^{+\infty} e^{-a\sqrt{x}} dx = \frac{2\sqrt{b}}{ae^{a\sqrt{b}}} + \frac{2}{a^2 e^{a\sqrt{b}}} \leq \frac{2\sqrt{b}}{a \cdot \left(1 + a\sqrt{b} + \frac{1}{2}a^2 b\right)} + \frac{2}{a^2 \cdot \left(1 + a\sqrt{b} + \frac{1}{2}a^2 b\right)} \leq \frac{4}{a^3\sqrt{b}} + \frac{2}{a^3\sqrt{b}} = \frac{6}{a^3\sqrt{b}}$, where the first inequality uses $e^x \geq 1 + x + \frac{1}{2}x^2$.

Now, we present the proofs for the results shown in (19) and (20).

Proofs for (19). We express the LHS in (19) as

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] = 1 - \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} \le r \mid \hat{\mu}_{1,s}\right\}\right] = 1 - \underbrace{\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{G}_{1,s} \le r\right\} \mid \hat{\mu}_{1,s}\right]\right]}_{=:\gamma} \quad . \tag{23}$$

Our goal is to construct a lower bound for γ. Let $\theta_{1,s}^h$ for all $h \in [r]$ be i.i.d. random variables according to $\mathcal{N}\left(\hat{\mu}_{1,s}, \frac{1}{s}\right)$.

When $r \in [0, 12]$**,** the proof is trivial as $\gamma \ge 0$. Then, we have $\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] \le 1$.

When $r \in [13, 100]$**,** we introduce $z = \sqrt{\frac{1}{2}\ln(r)} > 0$ and have

$$\begin{aligned}
\gamma &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{G}_{1,s} \le r\right\} \mid \hat{\mu}_{1,s}\right]\right] \\
&\ge \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \mu_1 - \tfrac{\Delta_i}{2}\right\} \mid \hat{\mu}_{1,s}\right]\right] \\
&\ge \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \ge \mu_1 - \tfrac{\Delta_i}{2}\right\} \mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}}\right\} \mid \hat{\mu}_{1,s}\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \ge \mu_1 - \tfrac{\Delta_i}{2}\right\} \cdot \underbrace{\mathbb{P}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + z\sqrt{\frac{1}{s}} \mid \hat{\mu}_{1,s}\right\}}_{=:\beta}\right] \quad .
\end{aligned} \tag{24}$$

We construct a lower bound for β and have

$$\begin{aligned}
\beta &= \mathbb{P}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \mid \hat{\mu}_{1,s}\right\} \\
&= 1 - \prod_{h\in[r]} \left(1 - \mathbb{P}\left\{\theta_{1,s}^h > \hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \mid \hat{\mu}_{1,s}\right\}\right) \\
&= 1 - \left(1 - \mathbb{P}\left\{\theta_{1,s} > \hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \mid \hat{\mu}_{1,s}\right\}\right)^r \\
&\ge^{(a)} 1 - \left(1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{1}{2}\ln(r)}}{\frac{1}{2}\ln(r)+1} e^{-\frac{1}{4}\ln(r)}\right)^r \\
&\ge^{(b)} 1 - e^{-r\cdot\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{1}{2}\ln(r)}}{\frac{1}{2}\ln(r)+1} \cdot r^{-\frac{1}{4}}} \\
&\ge^{(c)} 1 - e^{-r^{\frac{1}{2}}\cdot\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{1}{2}\ln(r)}}{\frac{1}{2}\ln(r)+\frac{1}{\ln(13)}\ln(r)} \cdot r^{\frac{1}{4}}} \\
&= 1 - e^{-r^{\frac{1}{2}}\cdot\frac{1}{\sqrt{\pi}}\cdot\frac{\ln(13)}{\ln(13)+2} \frac{\sqrt{r^{\frac{1}{2}}}}{\sqrt{\ln(r)}}} \\
&\ge^{(d)} 1 - e^{-r^{\frac{1}{2}}\cdot\frac{1}{\sqrt{\pi}}\cdot\frac{\ln(13)}{\ln(13)+2}} \quad .
\end{aligned} \tag{25}$$

Inequalities (a) uses anti-concentration bounds of Gaussian distributions (Fact 4) and (b) uses the fact that $e^{-x} \ge 1 - x$. Inequalities (c) and (d) use the facts that when $r \ge 13$, we have $1 \le \frac{\ln(r)}{\ln(13)}$ and $\sqrt{r^{\frac{1}{2}}} \ge \sqrt{\ln(r)}$.

Now, we have

$$\begin{aligned}
\gamma &\geq \left(1 - e^{-r^{\frac{1}{2}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}}\right) \cdot \mathbb{P}\left\{\hat{\mu}_{1,s} + \sqrt{\tfrac{1}{2}\ln(r)}\sqrt{\tfrac{1}{s}} \geq \mu_1 - \tfrac{\Delta_i}{2}\right\} \\
&\geq \left(1 - e^{-r^{\frac{1}{2}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}}\right) \cdot \mathbb{P}\left\{\hat{\mu}_{1,s} + \sqrt{\tfrac{1}{2}\ln(r)}\sqrt{\tfrac{1}{s}} \geq \mu_1\right\} \\
&\geq^{(a)} \left(1 - e^{-r^{\frac{1}{2}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}}\right) \cdot \left(1 - \tfrac{1}{r}\right) \\
&\geq 1 - \left(e^{-\sqrt{\frac{r}{\pi}} \cdot \frac{\ln(13)}{\ln(13)+2}} + \tfrac{1}{r}\right) \quad ,
\end{aligned} \tag{26}$$

where (a) uses Hoeffding's inequality (Fact 3).

Plugging the lower bound of γ into (23) gives $\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] \leq e^{-\sqrt{\frac{r}{\pi}} \frac{\ln(13)}{\ln(13)+2}} + \frac{1}{r}$.

When $r \geq 101$**,** we introduce $z = \sqrt{\frac{2}{3}\ln(r)} > 0$. We still construct the lower bound of γ as

$$\begin{aligned}
\gamma &\geq \mathbb{E}\left[\mathbf{1}\left\{\hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \geq \mu_1 - \tfrac{\Delta_i}{2}\right\} \cdot \mathbb{P}\left\{\max_{h\in[r]} \theta^h_{1,s} > \hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \mid \hat{\mu}_{1,s}\right\}\right] \\
&\geq \left(1 - \left(1 - \tfrac{1}{\sqrt{2\pi}} \tfrac{z}{z^2+1} e^{-0.5z^2}\right)^r\right) \cdot \mathbb{P}\left\{\hat{\mu}_{1,s} + z\sqrt{\tfrac{1}{s}} \geq \mu_1 - \tfrac{\Delta_i}{2}\right\} \\
&\geq \left(1 - \underbrace{e^{-r \cdot \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-0.5z^2}}}_{=:f(r)}\right) \cdot \mathbb{P}\left\{\hat{\mu}_{1,s} + \sqrt{\tfrac{2\ln(r)}{3}}\sqrt{\tfrac{1}{s}} \geq \mu_1\right\} \\
&\geq \left(1 - e^{-\sqrt{\frac{r}{3\pi}}}\right) \cdot \left(1 - \tfrac{1}{r^{\frac{4}{3}}}\right) \\
&\geq 1 - \left(e^{-\sqrt{\frac{r}{3\pi}}} + \tfrac{1}{r^{\frac{4}{3}}}\right) \quad .
\end{aligned} \tag{27}$$

The third inequality in (27) uses the fact that

$$\begin{aligned}
f(r) &= e^{-r \cdot \frac{1}{\sqrt{2\pi}} \frac{z}{z^2+1} e^{-0.5z^2}} \\
&= e^{-r \cdot \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\frac{\ln r}{1.5}}}{\frac{\ln r}{1.5}+1} e^{-0.5 \cdot \frac{\ln r}{1.5}}} \\
&\leq^{(a)} e^{-\frac{1}{\sqrt{2\pi}} \frac{\sqrt{1.5\ln r}}{\ln r + 0.5\ln r} \cdot r^{\frac{2}{3}}} \\
&= e^{-r^{\frac{1}{2}} \cdot \frac{1}{\sqrt{3\pi}} \sqrt{\frac{r^{\frac{1}{3}}}{\ln r}}} \\
&\leq^{(b)} e^{-\sqrt{\frac{r}{3\pi}}} \quad ,
\end{aligned} \tag{28}$$

where inequality (a) and (b) use the facts that when $r \geq 100$, we have $1.5 < 0.5\ln r$ and $\sqrt{\frac{r^{\frac{1}{3}}}{\ln r}} \geq 1$.

Plugging the lower bound of γ into (23) gives $\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] \leq e^{-\sqrt{\frac{r}{3\pi}}} + \frac{1}{r^{\frac{4}{3}}}$.

Proofs for (20). We still express the LHS in (20) as

$$\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] = 1 - \mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} \leq r \mid \hat{\mu}_{1,s}\right\}\right] = 1 - \underbrace{\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{G}_{1,s} \leq r\right\} \mid \hat{\mu}_{1,s}\right]\right]}_{=:\gamma} \quad . \tag{29}$$

When $r = 0$**,** we have $\gamma \geq 0$, which means $\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] \leq 1$.

When $r \in \left[1, \left\lfloor \left(T\Delta_i^2 + 100^{\frac{1}{3}}\right)^3 \right\rfloor\right]$, we define event $\mathcal{E}_{1,s}^{\mu} := \left\{ \mu_1 \leq \hat{\mu}_{1,s} + \sqrt{\frac{0.5 \ln\left(r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)\right)}{s}} \right\}$.

Then, we have

$$
\begin{aligned}
\gamma &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\mathcal{G}_{1,s} \leq r\right\} \mid \hat{\mu}_{1,s}\right]\right] \\
&\geq \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \mu_1 - \tfrac{\Delta_i}{2}\right\} \mid \hat{\mu}_{1,s}\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \mu_1 - \tfrac{\Delta_i}{2}\right\}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \mu_1 - \tfrac{\Delta_i}{2}\right\}\right] + \mathbb{E}\left[\mathbf{1}\left\{\overline{\mathcal{E}_{1,s}^{\mu}}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \mu_1 - \tfrac{\Delta_i}{2}\right\}\right] \\
&\geq^{(a)} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + \sqrt{\frac{0.5 \ln\left(r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)\right)}{s}} - \tfrac{\Delta_i}{2}\right\}\right] \\
&\geq^{(b)} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + \sqrt{\frac{0.5 \ln\left(r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)\right)}{s}} - \sqrt{\frac{\left(\sqrt{2}+\sqrt{3.5}\right)^2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}}\right\}\right] \\
&\geq^{(c)} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} + \sqrt{\frac{0.5 \ln\left(\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^7\right)}{s}} - \sqrt{\frac{\left(\sqrt{2}+\sqrt{3.5}\right)^2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}}\right\}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\mathbf{1}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} - \sqrt{\frac{2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}}\right\}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\} \cdot \mathbb{P}\left\{\max_{h\in[r]} \theta_{1,s}^h > \hat{\mu}_{1,s} - \sqrt{\frac{2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}} \mid \hat{\mu}_{1,s}\right\}\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\} \cdot \left(1 - \mathbb{P}\left\{\max_{h\in[r]} \theta_{1,s}^h \leq \hat{\mu}_{1,s} - \sqrt{\frac{2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}} \mid \hat{\mu}_{1,s}\right\}\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\} \cdot \left(1 - \prod_{h\in[r]} \mathbb{P}\left\{\theta_{1,s}^h \leq \hat{\mu}_{1,s} - \sqrt{\frac{2 \ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}} \mid \hat{\mu}_{1,s}\right\}\right)\right] \\
&\geq^{(d)} \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\} \cdot \left(1 - \frac{0.5^r}{\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^r}\right)\right] \\
&= \left(1 - \frac{0.5^r}{\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^r}\right) \cdot \mathbb{E}\left[\mathbf{1}\left\{\mathcal{E}_{1,s}^{\mu}\right\}\right] \\
&\geq^{(e)} \left(1 - \frac{0.5^r}{\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^r}\right) \cdot \left(1 - \frac{1}{r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}\right) \\
&\geq 1 - \frac{1}{r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)} - \frac{0.5^r}{\left(T\Delta_i^2+100^{\frac{1}{3}}\right)^r} \\
&\geq 1 - \frac{1}{r^2\left(T\Delta_i^2\right)} - \frac{0.5^r}{\left(T\Delta_i^2\right)^r} \quad .
\end{aligned}
\tag{30}
$$

We now provide detailed explanation for some key steps in (30). Inequality (a) uses the fact that if event $\mathcal{E}_{1,s}^{\mu}$ is true, we have $\mu_1 \leq \hat{\mu}_{1,s}+\sqrt{\frac{0.5\ln\left(r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)\right)}{s}}$. Recall $L_{1,i} = \left\lceil \frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{\Delta_i^2} \right\rceil$. Inequality (b) uses the fact that $\frac{\Delta_i}{2} \geq \sqrt{\frac{4\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{4L_1}} \geq \sqrt{\frac{\left(\sqrt{2}+\sqrt{3.5}\right)^2\ln\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}{s}}$, when $s \geq L_{1,i}$. Recall $1 \leq r \leq \left\lfloor \left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3 \right\rfloor$. Inequality (c) uses the fact that $r^2 \leq \left(T\Delta_i^2+100^{\frac{1}{3}}\right)^6$. Inequality (d) uses Gaussian concentration bounds (Fact 4) and inequality (e) uses Hoeffding's inequality (Fact 3) giving $\mathbb{P}\left\{\mathcal{E}_{1,s}^{\mu}\right\} \geq 1-\frac{1}{r^2\left(T\Delta_i^2+100^{\frac{1}{3}}\right)}$.

Plugging the lower bound of γ into (29) gives $\mathbb{E}\left[\mathbb{P}\left\{\mathcal{G}_{1,s} > r \mid \hat{\mu}_{1,s}\right\}\right] \leq \frac{1}{r^2\left(T\Delta_i^2\right)} + \frac{0.5^r}{\left(T\Delta_i^2\right)^r}$.

When $r \geq \left\lfloor \left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3 \right\rfloor + 1$**,** we reuse the result shown in (19) directly. Note that we have $r \geq \left\lfloor \left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3 \right\rfloor + 1 \geq \left(T\Delta_i^2+100^{\frac{1}{3}}\right)^3 \geq 101$. ■

Lemma 6 *We have*

$$\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{i_t=i,\overline{\mathcal{E}_i^{\theta}(t)},n_i(t-1)\geq L_i\right\}\right] \quad \leq \quad \frac{0.5}{\Delta_i^2} \quad . \tag{31}$$

Proof of Lemma 6: This lemma is very similar to Lemma 2.16 in [1]. Recall event $\mathcal{E}_i^{\theta}(t) = \left\{\theta_i(t) \leq \hat{\mu}_{i,n_i(t-1)}+\sqrt{\frac{2\ln\left(T\Delta_i^2\right)}{n_i(t-1)}}\right\}$. Let $\tau_s^{(i)}$ be the round when arm i is pulled for the s-th time. We have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}\left\{i_t=i,\overline{\mathcal{E}_i^{\theta}(t)},n_i(t-1)\geq L_i\right\}\right] \\
\leq\ &\sum_{s=L_i}^{T}\mathbb{E}\left[\sum_{t=\tau_s^{(i)}+1}^{\tau_{s+1}^{(i)}}\mathbf{1}\left\{i_t=i,\overline{\mathcal{E}_i^{\theta}(t)}\right\}\right] \\
\leq\ &\sum_{s=L_i}^{T}\mathbb{E}\left[\mathbf{1}\left\{\overline{\mathcal{E}_i^{\theta}\left(\tau_{s+1}^{(i)}\right)}\right\}\right] \\
=\ &\sum_{s=L_i}^{T}\mathbb{E}\left[\underbrace{\mathbb{P}\left\{\theta_{i,s}>\hat{\mu}_{i,s}+\sqrt{\frac{2\ln(T\Delta_i^2)}{s}}\mid\hat{\mu}_{1,s}\right\}}_{\text{Fact 4}}\right] \\
\leq\ &\sum_{s=L_i}^{T}\frac{1}{2}e^{-0.5\cdot 2T\Delta_i^2} \\
\leq\ &\frac{0.5}{\Delta_i^2} \quad .
\end{aligned} \tag{32}$$

■

Lemma 7 *We have*

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \overline{\mathcal{E}_i^{\mu}(t-1)}, n_i(t-1) \geq L_i\right\}\right] \quad \leq \quad \frac{2}{\Delta_i^2} \quad . \tag{33}$$

Proof of Lemma 7: This lemma is very similar to Lemma 2.15 in [1]. Recall event $\mathcal{E}_i^{\mu}(t-1) = \left\{\left|\hat{\mu}_{i,n_i(t-1)} - \mu_i\right| \leq \sqrt{\frac{0.5\ln\left(T\Delta_i^2\right)}{n_i(t-1)}}\right\}$. Let $\tau_s^{(i)}$ be the round when arm i is pulled for the s-th time. Now, we partition all T rounds based on the pulls of arm i. We have

$$\begin{aligned}
& \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\left\{i_t = i, \overline{\mathcal{E}_i^{\mu}(t-1)}, n_i(t-1) \geq L_i\right\}\right] \\
\leq & \sum_{s=L_i}^{T} \mathbb{E}\left[\sum_{t=\tau_s^{(i)}+1}^{\tau_{s+1}^{(i)}} \mathbf{1}\left\{i_t = i, \overline{\mathcal{E}_i^{\mu}(t-1)}\right\}\right] \\
\leq & \sum_{s=L_i}^{T} \mathbb{E}\left[\mathbf{1}\left\{\overline{\mathcal{E}_i^{\mu}\left(\tau_{s+1}^{(i)} - 1\right)}\right\}\right] \\
= & \sum_{s=L_i}^{T} \underbrace{\mathbb{P}\left\{\left|\hat{\mu}_{i,s} - \mu_i\right| > \sqrt{\frac{0.5\ln\left(T\Delta_i^2\right)}{s}}\right\}}_{\text{Hoeffding's inequality}} \\
\leq & \sum_{s=L_i}^{T} \frac{2}{(T\Delta_i^2)} \\
\leq & \frac{2}{\Delta_i^2} \quad ,
\end{aligned} \tag{34}$$

which concludes the proof. ■