Gaussian Randomized Exploration for Semi-bandits with Sleeping Arms

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper provides theoretical analyses of problem-independent upper and lower regret bounds for Gaussian randomized algorithms in semi-bandits with sleeping arms, where arms may be unavailable in certain rounds, and available arms satisfying combinatorial constraints can be played simultaneously. We first introduce the CTS-G algorithm, an adaptation of Thompson sampling with Gaussian priors, achieving an upper bound of $\tilde{O}(m\sqrt{NT})$ over T rounds with N arms and up to m arms played per round, where $\tilde{O}$ hides the logarithmic factors. Next, we present CL-SG, which improves upon CTS-G by using a single Gaussian sample per round, achieving a near-optimal upper regret bound of $\tilde{O}(\sqrt{mNT})$. We also establish that both algorithms have lower regret bounds of $\Omega(\sqrt{mNT \ln \frac{N}{m}})$ and $\Omega(\sqrt{mNT})$, respectively.

1 Introduction

We consider a sleeping semi-bandit problem with a fixed set $[N] = \{1, 2, \ldots, N\}$ of N base arms and each base arm $a \in [N]$ is associated with an unknown probability distribution p_a supported on $[0, 1]$ and mean r_a. Unlike standard combinatorial bandits (Kveton et al., 2015), where a learning agent, in each round $t = 1, \ldots, T$, plays a super arm (combinations of base arms) $A_t \in \Theta$, where $\Theta \subseteq 2^{[N]}$ is a *feasible set* that satisfy certain constraints, sleeping semi-bandits involve a time-varying feasible set $\Theta_t \subseteq \Theta$, revealed at each round t. After observing the feasible set Θ_t in round t, the learning agent selects a super arm $A_t \in \Theta_t$, observes rewards $r_{a,t} \sim p_a$ for each base arm $a \in A_t$, and aims to minimize the T-round (pseudo)-regret defined as follows.

$$\mathcal{R}(T) := \sum_{t=1}^{T} \mathbf{E}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} r_a\right], \tag{1}$$

where $A_t^* := \arg\max_{A\in\Theta_t} \sum_{a\in A} r_a$ denotes the optimal super arm in round t and the expectation is taken over Θ_t, A_t, and A_t^*. Note that A_t^* is determined by Θ_t. We further denote by $m := \max_{A\in\Theta} |A|$ the maximum number of base arms in any super arm.

The *upper confidence bound (UCB)* (Agrawal, 1995; Auer et al., 2002) and *Thompson sampling (TS)* (Thompson, 1933; Kaufmann et al., 2012; Agrawal & Goyal, 2012, 2017a) are two leading algorithmic families for addressing stochastic bandit problems. For semi-bandit settings, the minimax lower bound is established as $\Omega(\sqrt{mNT})$(Kveton et al., 2015; Merlis & Mannor, 2020), and UCB-based algorithms achieve an upper bound of $O(\sqrt{mNT \ln T})$ (Kveton et al., 2015). Although TS-based algorithms have been analyzed for problem-dependent bounds in semi-bandits (Wang & Chen, 2018; Perrault et al., 2020), their results cannot be simply extended to reasonable problem-independent bounds because their bounds contain constant terms that grow exponentially with m.

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

While a substantial body of literature has explored the setting of sleeping semi-bandits (Hu et al., 2019; Li et al., 2019; Wu & Li, 2024) using *upper confidence bound (UCB)*-based algorithms with an upper bound of $O(\sqrt{mNT \ln T})$, the upper and lower bounds for *Thompson sampling (TS)*-based algorithms for (sleeping) semi-bandits still remain an open problem. Since TS is highly competitive with advanced UCB-based algorithms and widely used in large-scale applications (Chapelle & Li, 2011), investigating the theoretical performance of TS-based algorithms is crucial.

This work addresses long-standing gaps in the literature by introducing two algorithms with provable theoretical guarantees. The first algorithm, CTS-G, is an adaptation of TS with Gaussian priors specifically designed for sleeping semi-bandits, achieving an upper bound of $\tilde{O}(m\sqrt{NT})$, where $\tilde{O}$ hides the logarithmic factors, and a lower bound of $\Omega(\sqrt{mNT \ln \frac{N}{m}})$. We further introduce CL-SG, which improves upon CTS-G both theoretically and practically by employing only a single Gaussian sample, resulting in tighter bounds: an upper bound of $\tilde{O}(\sqrt{mNT})$ and a lower bound of $\Omega(\sqrt{mNT})$. CL-SG is minimax-optimal up to logarithmic factors compared to the known lower bound for combinatorial bandits (Kveton et al., 2015; Merlis & Mannor, 2020).

2 Gaussian Randomized Algorithms

We first present some notations specific to this section. Let $n_{a,t} := \sum_{\tau=1}^{t-1} \mathbf{1}[a \in A_\tau]$ denote the total number of times that base arm $a \in [N]$ has been pulled at the beginning of round t. Let $\hat{r}_{a,n_{a,t}} := \frac{\sum_{\tau=1}^{t-1} \mathbf{1}[a \in A_\tau] \cdot r_{a,\tau}}{n_{a,t}}$ denote the empirical mean of base arm a at the beginning of round t, which is the average of $n_{a,t}$ i.i.d. random variables according to reward distribution p_a. Let $\mathcal{F}_t$ collect all the actions and observed rewards up to the end of round t

In Sec. 2.1, we present CTS-G, an algorithm enjoying $\tilde{O}(m\sqrt{NT})$ and $\Omega(\sqrt{mNT \ln \frac{N}{m}})$ upper and lower regret bounds. In Sec. 2.2, we present CL-SG, an algorithm enjoying $\tilde{O}(\sqrt{mNT})$ and $\Omega(\sqrt{mNT})$ upper and lower regret bounds. The practical performance of both algorithms is discussed in Appendix A, and all the detailed proofs can be found in Appendix C to D.

2.1 Combinatorial Thompson Sampling with Gaussian Priors (CTS-G)

CTS-G presented in Alg. 1 is a direct adaptation of TS with Gaussian priors (Agrawal & Goyal, 2017b) to the sleeping semi-bandit problems. The core idea is to use posterior distributions to model the mean reward r_a of each base arm $a \in [N]$. In each round t, CTS-G draws a Gaussian posterior sample $w_{a,t} \sim \mathcal{N}(\hat{r}_{a,n_{a,t}}, \frac{\gamma m \ln t}{n_{a,t}+1})$ for each $a \in [N]$, where $\gamma > 0$ is a constant parameter to control the exploration level.[1] We can view the collection $\boldsymbol{w}_t = \{w_{a,t}, \forall a \in [N]\}$ of all posterior samples as the "sampled problem instance" based on which the learning agent conducts learning in round t. Then, based on the revealed feasible set Θ_t, CTS-G plays the super arm $A_t \in \arg\max_{A \in \Theta_t} \sum_{a \in A} w_{a,t}$ with the highest aggregated value of posterior samples and observes each individual base arm's random reward.

Theorem 1. *(1) The regret of CTS-G is* $O\left(m \ln(T)\sqrt{NT}\right)$. *(2) There exists a problem instance such that CTS-G suffers* $\Omega\left(\sqrt{mNT \ln\left(\frac{N}{m}\right)}\right)$ *regret.*

Discussion. Theorem 1 states that CTS-G is worst-case optimal up to an extra $\ln(T)\sqrt{m}$ factor. Compared with UCB-based algorithms for sleeping semi-bandits, our upper bound has an extra factor of $\sqrt{m \ln T}$ with the ones by Hu et al. (2019); Li et al. (2019), which are $O(\sqrt{mNT \ln T})$. However, it is important to note a significant aspect of our model: unlike the assumptions in Hu et al. (2019); Li et al. (2019), our bound is derived without relying on stochastic assumptions regarding the availability of arms. Furthermore, the upper bound is minimax optimal up to an extra $\ln(T)\sqrt{m}$ factor as compared to the $\Omega\left(\sqrt{mNT}\right)$ minimax lower bound for combinatorial bandits shown in Merlis & Mannor (2020).

[1] In practice, we only need to draw posterior samples for available arms to improve efficiency.

Upper bound proof sketch. The theoretical analysis is non-trivial due to overlapping base arms among super arms, the dynamic nature of the optimal super arm A_t^*, and its unobservability, as only the played super arm A_t is observed in each round t To decompose the regret, we define a high probability event for the empirical estimates. Let $\mathcal{E}_t := \left\{ |r_a - \hat{r}_{a,n_{a,t}}| \leq \sqrt{\frac{3 \ln(Nt)}{n_{a,t}+1}}, \forall a \in [N] \right\}$ be the event that the empirical means are close to their true means by the beginning of round t. Let $t' = \max\{\sqrt{m}, 4\}$ and $\mathbf{E}_{\Theta_t}[\cdot] := \mathbf{E}[\cdot \mid \Theta_t]$. Then, we decompose the regret defined in (1) as

$$\mathcal{R}(T) \leq \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\sum_{a \in A_t^*} r_a - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]\right]}_{=:I_1, \text{ optimism term}} + \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} (w_{a,t} - r_a)\, \mathbf{1}[\mathcal{E}_t]\right]\right]}_{=:I_2, \text{ deviation term}} + mt' + O(1).$$

The deviation term I_2 is easy to analyze as we can observe A_t, and is upper bounded by $\tilde{O}(m\sqrt{NT})$ via using concentration bounds. The center question is how to upper bound the optimism term, which measures the gap between the *maximum amount of true reward* $\sum_{a \in A_t^*} r_a$ the learning agent could achieve and the *expected maximum amount of reward* $\sum_{a \in A_t} w_{a,t}$ the learning agent can observe in round t. Intuitively, if the learning agent is lucky, i.e., the history $\mathcal{F}_{t-1}$ gives $\sum_{a \in A_t^*} r_a \leq \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]$, there is no regret in round t for this term. Let $(\cdot)^+ := \max\{\cdot, 0\}$ be an activation function. Then, we have

$$\sum_{a \in A_t^*} r_a - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right] \leq \left(\sum_{a \in A_t^*} r_a - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]\right)^+ . \tag{2}$$

Let $c(\gamma)$ be a constant only depending on γ. In our novel technical Lemma 1, inspired by Russo (2019), we show

$$\left(\sum_{a \in A_t^*} r_a - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]\right)^+ \leq c(\gamma) \cdot \mathbf{E}_{\Theta_t}\left[\left(\sum_{a \in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]\right)^+\right], \tag{3}$$

which tackles the challenge brought by the unobservability of A_t^*.

Next, via introducing an independent "ghost" copy $\tilde{w}_{a,t} \sim \mathcal{N}(\hat{r}_{a,n_{a,t}}, \frac{\gamma m \ln t}{n_{a,t}+1})$ of $w_{a,t}$, we show

$$\mathbf{E}_{\Theta_t}\left[\left(\sum_{a \in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a \in A_t} w_{a,t}\right]\right)^+\right] \leq \mathbf{E}_{\Theta_t}\left[\left|\sum_{a \in A_t} (w_{a,t} - \tilde{w}_{a,t})\right|\right], \tag{4}$$

which gets rid of the introduced activation function.

Since $w_{a,t} - \tilde{w}_{a,t} \sim \mathcal{N}\left(0, \frac{2\gamma m \ln t}{n_{a,t}+1}\right)$, we only need to deal with Gaussian random variables and have

$$\sum_{t=t'}^{T} \mathbf{E}\left[\left|\sum_{a \in A_t} (w_{a,t} - \tilde{w}_{a,t})\right|\right] \leq O\left(m \ln T\sqrt{\gamma NT}\right). \tag{5}$$

Lower bound proof sketch. Inspired by Theorem 1.4 in Agrawal & Goyal (2017b), the lower bound is refined by constructing a path selection problem with N links (base arms) and K paths (super arms) of m links. This reduces the semi-bandits to K independent path selections, and the result follows using the anti-concentration inequality for Gaussian variables (Appendix C.7)

2.2 Combinatorial Learning with Single Gaussian Seed (CL-SG)

Since the upper bound of CTS-G still has an extra $\ln(T)\sqrt{m}$ factor from the minimax lower bound Merlis & Mannor (2020) for combinatorial bandits, we are motivated to improve the upper bound by controlling the amount of randomness injected within the learning algorithm.

Inspired by Xiong et al. (2022), we devise CL-SG which enjoys a $\tilde{O}(\sqrt{mNT})$ regret bound. The key idea behind the removal of the extra $\sqrt{m}$ factor as compared to the regret of CTS-G (Alg. 1) is CL-SG uses a single random seed $w_t \sim \mathcal{N}(0, 1)$ to perturb the empirical estimates of all the base arms, as shown in Alg. 2. After drawing w_t, we construct $\bar{r}_{a,t} = \hat{r}_{a,n_{a,t}} + w_t \cdot \sqrt{\frac{\gamma \ln t}{n_{a,t}+1}}$ for all the base arms $a \in [N]$, where constant $\gamma > 0$ controls the exploration level. Then, we play $A_t = \arg\max_{A \in \Theta_t} \sum_{a \in A} \bar{r}_{a,t}$ from the feasible set Θ_t in round t.

Theorem 2. *(1) The regret of CL-SG is $O\left(\ln T\sqrt{mNT}\right)$. (2) There exists a problem instance such that CL-SG suffers $\Omega\left(\sqrt{mNT}\right)$ regret.*

Discussion. The extra $\sqrt{m}$ in CTS-G comes from the m factor in the variance of the Gaussian posterior sample $w_{a,t}$, necessary to keep $c(\gamma)$ bounded by a constant. To bound $c(\gamma)$, we must lower bound $\Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} - \hat{r}_{a,n_{a,t}} \geq \sum_{a \in A_t^*} \sqrt{\frac{4 \ln t}{n_{a,t}+1}}\right)$, requiring the Cauchy-Schwarz inequality to bring the summation inside the square root for the RHS term in the probability, which scales with $\sqrt{m}$, i.e., $\sum_{a \in A_t^*} \sqrt{\frac{4 \ln t}{n_{a,t}+1}} \leq \sqrt{m \sum_{a \in A_t^*} \frac{4 \ln t}{n_{a,t}+1}}$. This fact further results in an extra m in the variance of CTS-G Gaussian samples for the probability to be lower bounded by a constant. On the other hand, with CL-SG, using a single w_t, we lower bound a similar probability, $\Pr\left(\sum_{a \in A_t^*} w_t \sqrt{\frac{\gamma \ln t}{n_{a,t}+1}} \geq \sum_{a \in A_t^*} \sqrt{\frac{4 \ln t}{n_{a,t}+1}}\right)$, allowing us to divide both sides by $\sum_{a \in A_t^*} \sqrt{\frac{4 \ln t}{n_{a,t}+1}}$ and avoid the extra m in the variance.

The lower-bound proof considers the same problem instance to Theorem 1 but differs in addressing the arms' dependency in CL-SG due to the common w_t. Let $\Delta := \sqrt{\frac{N}{mT}}$ be the reward gap between each suboptimal arm and the optimal super arm, and let $Q_A(t)$ be the number of times that super arm A has been played at the beginning of round t. Then, denote by $B_t^* := \{Q_{A_1}(t) > t - cT\}$ the event that the optimal super arm A_1 has been observed enough times at the beginning of round t, where $c \in (0, 1)$ is a constant. It is easy to prove that the regret is lower bounded by $cT \cdot m \cdot \Delta = \Omega(\sqrt{mNT})$ when B_t^* is false for some $t \in [T]$. The main challenge is to demonstrate that, conditioned on the past histories F_{t-1} that lead to the happening of event B_t^*, the probability of playing a suboptimal super arm is at least a constant probability p_0, i.e., $\Pr\left(\exists A \in \Theta \setminus A_1 : A_t = A \mid \mathcal{F}_{t-1} = F_{t-1}\right) \geq p_0$. This leads to lower bound a probability that the empirical estimates of suboptimal arms are larger than the optimal super arm, i.e.,

$$\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a \in A} \hat{r}_{a,Q_A(t)} + w_t\sqrt{\frac{\gamma \ln t}{Q_A(t) + 1}} > \sum_{b \in A_1} \hat{r}_{b,Q_{A_1}(t)} + w_t\sqrt{\frac{\gamma \ln t}{Q_{A_1}(t) + 1}} \mid \mathcal{F}_{t-1} = F_{t-1}\right)$$
$$\geq \Pr\left(\exists A \in \Theta \setminus A_1 : w_t\left(1 - \sqrt{\frac{Q_A(t) + 1}{Q_{A_1}(t) + 1}}\right) > \Delta\sqrt{Q_A(t) + 1} \mid \mathcal{F}_{t-1} = F_{t-1}\right).$$

This requires analyzing the play ratio between suboptimal and optimal super arms, while in the lower-bound analysis of Theorem 1, we can avoid this situation by independently considering the estimates of the optimal super arm is smaller than 0, and that of suboptimal arms is larger than 0. The trick to address this ratio is to only consider the regret from αT to T, with $\alpha \in (0, 1)$ such that $\frac{Q_A(t)+1}{Q_{A_1}(t)+1} \leq \frac{cT+1}{(\alpha-c)T+1}$ is a constant by tuning c and α. Then, by applying the anti-concentration bound for Gaussian variables and solving a non-trivial optimization problem, we can prove such a p_0 exists, and regret is lower bounded by $(1 - \alpha)T \cdot p_0 \cdot m\Delta = \Omega(\sqrt{mNT})$.

3 Conclusion

In this paper, we have studied the problem of sleeping semi-bandits and presented CTS-G and CL-SG with theoretical guarantees. Our results bridge the existing gap in the literature by providing upper and lower bounds for TS-based algorithms in sleeping semi-bandits. Future work will focus on narrowing the gap between these bounds, and studying the relationship between the number of random variables and their variances.

References

Rajeev Agrawal. Sample Mean Based Index Policies by $O(\log n)$ Regret for The Multi-armed Bandit Problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for The Multi-armed Bandit Problem. In *Proc. Conference on Learning Theory (COLT)*, volume 23, pp. 39.1–39.26. PMLR, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal Regret Bounds for Thompson Sampling. *Jounal of ACM*, 64(5):30:1–30:24, September 2017a. ISSN 0004-5411.

Shipra Agrawal and Navin Goyal. Near-optimal Regret Bounds for Thompson Sampling. `http://www.columbia.edu/~sa3305/papers/j3-corrected.pdf`, 2017b.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of The Multiarmed Bandit Problem. *Machine Learning*, 47(2–3):235–256, 2002.

Olivier Chapelle and Lihong Li. An Empirical Evaluation of Thompson Sampling. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2249–2257, 2011.

Bingshan Hu, Yunjin Chen, Zhiming Huang, Nishant A. Mehta, and Jianping Pan. Intelligent Caching Algorithms in Heterogeneous Wireless Networks with Uncertainty. In *Proc. IEEE Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite-time Analysis. In *Proc. International Conference on Algorithmic Learning Theory (ALT)*, pp. 199–213. Springer, 2012.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits. In *Proc. Artificial Intelligence and Statistics (AISTATS)*, pp. 535–543, 2015.

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial Sleeping Bandits with Fairness Constraints. *IEEE Transactions on Network Science and Engineering (TNSE)*, 7(3):1799–1813, 2019.

Nadav Merlis and Shie Mannor. Tight Lower Bounds for Combinatorial Multi-armed Bandits. In *Proc. Conference on Learning Theory (COLT)*, pp. 2830–2857. PMLR, 2020.

Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical Efficiency of Thompson Sampling for Combinatorial Semi-bandits. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33:5429–5440, 2020.

Daniel Russo. Worst-case Regret Bounds for Exploration via Randomized Value Functions. *Proc. Advances in Neural Information Processing Systems*, 32, 2019.

William R Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of The Evidence of Two Samples. *Biometrika*, 25(3-4):285–294, 1933.

Siwei Wang and Wei Chen. Thompson Sampling for Combinatorial Semi-Bandits. In *Proc. International Conference on Machine Learning (ICML)*, pp. 5101–5109, 2018.

Xiaoyi Wu and Bin Li. Achieving Regular and Fair Learning in Combinatorial Multi-Armed Bandit. In *Proc. IEEE Conference on Computer Communications (INFOCOM)*, pp. 361–370. IEEE, 2024.

Zhihan Xiong, Ruoqi Shen, Qiwen Cui, Maryam Fazel, and Simon Shaolei Du. Near-optimal Randomized Exploration for Tabular Markov Decision Processes. In *Proc. Advances in Neural Information Processing Systems*, 2022.

A Numerical Experiments

A.1 Algorithm Description for CTS-G and CL-SG

The algorithm descriptions for CTS-G and CL-SG are presented in Algs. 1 and 2, respectively.

Algorithm 1 Combinatorial Thompson Sampling with Gaussian Priors (CTS-G)

Require: arm set $[N]$, exploration rate γ
  Initialize $n_{a,1} = 0$ and $\hat{r}_{a,n_{a,1}} = 0$ for all base arms $a \in [N]$
  for $t = 1, 2, \ldots$ **do**
    Observe feasible set Θ_t
    Draw $w_{a,t} \sim \mathcal{N}(\hat{r}_{a,n_{a,t}}, \frac{\gamma m \ln t}{n_{a,t}+1})$ for each base arm $a \in [N]$
    Play super arm $A_t = \arg\max_{A \in \Theta_t} \sum_{a \in A} w_{a,t}$
    Observe $r_{a,t} \sim p_a$ for all base arms $a \in A_t$ and update $n_{a,t}$ and $\hat{r}_{a,n_{a,t}}$ for all $a \in A_t$.
  end for

Algorithm 2 Combinatorial Learning with Single Gaussian Seed (CL-SG)

Require: arm set $[N]$, exploration rate γ
  Initialize $n_{a,1} = 0$ and $\hat{r}_{a,0} = 0$ for all base arms $a \in [N]$
  for $t = 1, \ldots$ **do**
    Observe decision set Θ_t
    Draw $w_t \sim \mathcal{N}(0, 1)$
    Construct $\bar{r}_{a,t} = \hat{r}_{a,n_{a,t}} + w_t \cdot \sqrt{\frac{\gamma \ln t}{n_{a,t}+1}}$ for all base arms $a \in [N]$
    Play super arm $A_t = \arg\max_{A \in \Theta_t} \sum_{a \in A} \bar{r}_{a,t}$
    Observe $r_{a,t} \sim p_a$ for all base arms $a \in A_t$ and update $n_{a,t}$ and $\hat{r}_{a,n_{a,t}}$ for all $a \in A_t$.
  end for

A.2 Combinatorial Learning with Least Gaussian Seed (CL-LG)

We aim to explore whether further reducing the number of Gaussian samples in the algorithm can enhance the practical performance. To this end, we propose the *Combinatorial Learning with Least Gaussian Seed (CL-LG)* algorithm, as shown in Alg. 3. Different from CL-SG (see Alg. 2), which requires an independent Gaussian sample in each round, our approach only draws a single Gaussian sample $w \sim \mathcal{N}(0, 1)$ at the beginning of the game.

Algorithm 3 Combinatorial Learning with Least Gaussian Seed (CL-LG)

Require: arm set $[N]$, exploration rate γ
  Initialize $n_{a,1} = 0$ and $\hat{r}_{a,0} = 0$ for all base arms $a \in [N]$
  Draw $w \sim \mathcal{N}(0, 1)$
  for $t = 1, \ldots$ **do**
    Observe feasible set Θ_t
    Construct $\bar{r}_{a,t} = \hat{r}_{a,n_{a,t}} + w \cdot \sqrt{\frac{\gamma \ln t}{n_{a,t}+1}}$ for all base arms $a \in [N]$
    Play super arm $A_t = \arg\max_{A \in \Theta_t} \sum_{a \in A} \bar{r}_{a,t}$
    Observe $r_{a,t} \sim p_a$ for all base arms $a \in A_t$ and update $n_{a,t}$ and $\hat{r}_{a,n_{a,t}}$ for all $a \in A_t$.
  end for

A.3 Experiment Settings

We conduct experiments in two settings to show the performance of the proposed algorithms with $\gamma = 0.1$ to study the number of Gaussian seeds and the impact of different γ, which can be found in Appendix A.4 and A.5. All the experiment results are the average of 100 independent experiments conducted on a MacBook Pro with M1 Max and 32GB RAM using Numpy.

In Setting 1, we consider a simple environment with $N = 10$ arms, and at most $m = 3$ arms can be played in each round. The actual rewards for all the arms follow the Bernoulli distributions, while the first three arms have a mean reward of 0.9, and the rest of the arms have a mean reward of 0.8. In Setting 2, we consider a more complicated setting where $N = 50$ and $m = 15$. In this setting, rewards are again based on Bernoulli distributions, where the first five arms have mean rewards generated uniformly from $[0.725, 0.75]$, and the rest of the arms have mean rewards generated uniformly from $[0.7, 0.725]$. For both settings, the availability of each arm is determined by a Bernoulli distribution with a mean of 0.5. The reason we chose Bernoulli distributions for the rewards is that we want to compare with the following CTS-B (which requires Bernoulli rewards) and CombUCB algorithms. Both algorithms play arms $A_t := \arg\max_{A \in \Theta_t} \sum_{a \in A} \theta_{a,t}$, where $\theta_{a,t}$ is defined differently as follows.

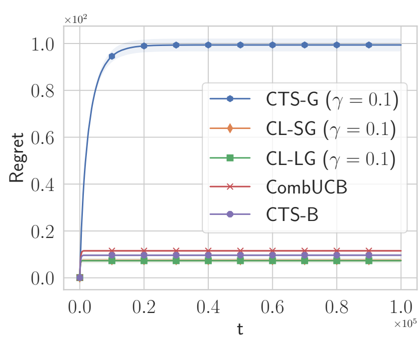
- CTS-B (Wang & Chen, 2018): In each round t, CTS-B draws random samples from Beta distributions for each available arm $\theta_{a,t} \sim \mathbf{Beta}(\hat{r}_{a,n_{a,t}} n_{a,t} + 1, n_{a,t} - n_{a,t}\hat{r}_{a,n_{a,t}} + 1)$, and plays arms $A_t := \arg\max_{A \in \Theta_t} \sum_{a \in A} \theta_{a,t}$.
- CombUCB (Kveton et al., 2015): In each rount t, CombUCB estimates the UCB values for each arm $\theta_{a,t} = \hat{r}_{a,n_{a,t}} + \sqrt{\frac{1.5 \ln t}{n_{a,t}}}$ in each round t.

A.4 Impact of Number of Gaussian Seed

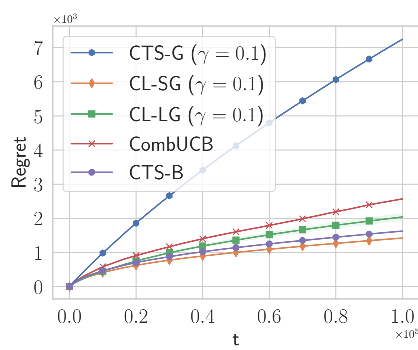
The regret results over $T = 10^5$ rounds are shown in Fig. 1 with 97.5% confidence intervals. In both settings, CTS-G performs worse than others, suffering the highest regret, because of the algorithm's reliance on Gaussian random samples, which are unbounded and result in an excessive exploration rate. This overemphasis on exploration, at the expense of exploiting known rewarding arms, fundamentally undermines the algorithm's efficiency.

On the other hand, CL-SG demonstrates comparable performance to CL-LG in Setting 1, both outperforming CTS-B. In Setting 2, CL-SG maintains its advantage, whereas CL-LG falls behind CTS-G. This highlights the effectiveness of CL-SG's design in optimizing the exploration-exploitation trade-off more efficiently than its counterparts.

Notably, in both settings, CL-LG with $\gamma = 0.1$ outperforms CombUCB, suggesting that the initial randomness incorporated in CL-LG helps balance the trade-off between exploration and exploitation. It remains an open question of how initial randomness helps.



(a) Setting 1 ($N = 10$ and $m = 3$)



(b) Setting 2 ($N = 50$ and $m = 15$)

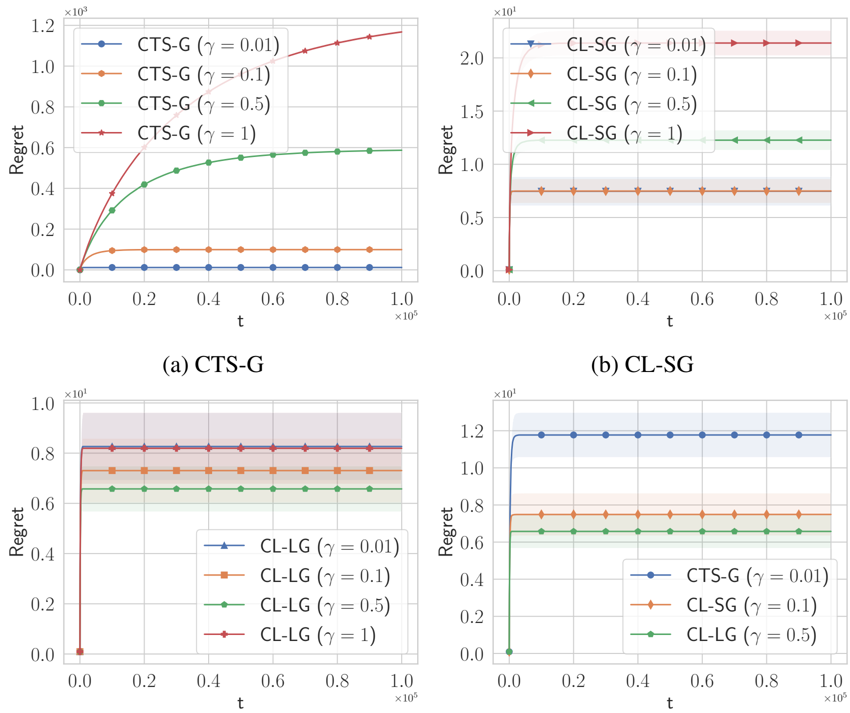
Figure 1: The comparison of regret for both settings with $\gamma = 0.1$.

A.5 Impact of Different γ

We performed experiments with the CTS-G, CL-SG, and CL-LG algorithms under Settings 1 and 2. The experiments utilized γ values of 0.01, 0.1, 0.5, and 1. The results are illustrated in Figs. 2 and 3.

Regarding Setting 1, CTS-G performs worse as γ increases, as shown in Fig. 2a, because a higher γ corresponds to a higher exploration rate, which will over-explore the simple scenario. For CL-SG, the performance with $\gamma = 0.1$ is better than that with other γ values. CL-LG achieves the best performance with 0.5, which indicates the performance of algorithms is not necessarily linear with γ.

When comparing the algorithms at their optimal γ values (see Fig. 2d), CTS-G shows the worst performance, whereas CL-SG performs comparably to CL-LG.



(a) CTS-G (b) CL-SG

(c) CL-LG (d) Best-case Comparison

Figure 2: The comparison of different γ for CTS-G, CL-SG and CL-LG in Setting 1.

Additionally, CTS-G is very sensitive to the change of γ, and the performance of CTS-G with $\gamma = 1$ is about 200 times worse than that of CTS-G with $\gamma = 0.01$. In contrast, CL-SG and CL-LG demonstrate greater robustness to changes in γ, showing that fewer Gaussian samples may prevent over-exploration.

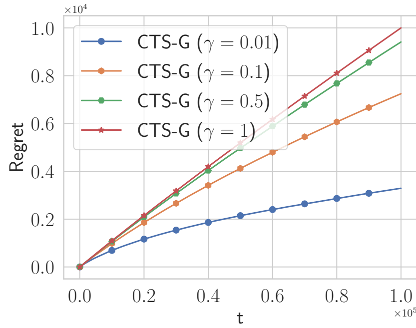
Regarding the more complicated Setting 2, we can observe a change in CL-LG, where $\gamma = 1$ leads to the worst performance, while $\gamma = 0.5$ achieves the best performance. This indicates that CL-LG with $\gamma = 1$ will over-explore. When comparing all the algorithms with their optimal γ values, we can see that CL-SG with $\gamma = 0.1$ achieves the best performance. More interestingly, we can observe that algorithms with fewer Gaussian samples require higher γ to achieve better performance.

From this experiment, we can see that different Gaussian samples react differently to different exploration rates. This observation raises an intriguing question for future research: what is the relationship between the number of random variables and their variance, and what is the optimal combination to achieve the best results?

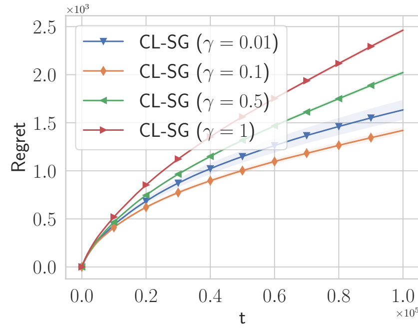
A.6 Tightness of regret bound

We consider a setting of 100 arms, and at most 10 arms can be played in each round. The mean rewards for the first 10 arms are 0.925, and the mean rewards for the rest suboptimal arms are 0.9. We compare the regret of CTS-G with the lower regret bound $0.1\sqrt{mNT\ \ln(\frac{N}{m})}$ in Fig. 4a. As we can see, there are still gaps between the actual performance and the theoretical lower bound, and the increasing rate of CTS-G is larger than the lower bound, which indicates that the lower bound may still have room to be improved.

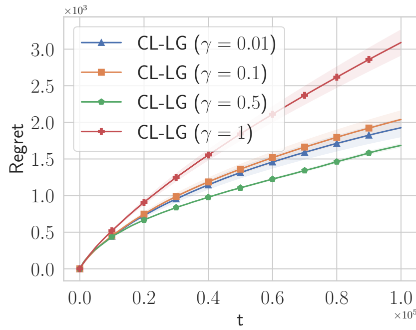
Similarly, we compared CT-SG with the lower bound of $0.1\sqrt{mNT}$, and we can see that the regret of CL-SG increases faster than the lower bound, indicating that the lower bound can be improved.



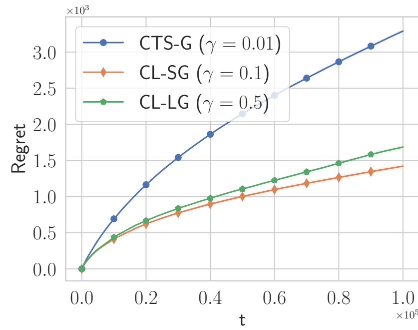
(a) CTS-G



(b) CL-SG

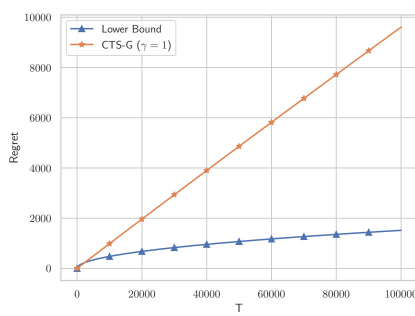


(c) CL-LG

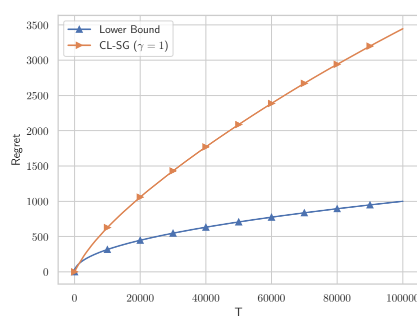


(d) Best-case Comparison

Figure 3: The comparison of different γ for CTS-G, CL-SG and CL-LG in Setting 2.



(a) Tightness of the lower bound for CTS-G



(b) Tightness of the lower bound for CS-SG

Figure 4: Tightness of Regret Bound for both CTS-G and CL-SG.

B Notations and Facts

Notations: Let $\mathcal{F}_{t-1}$ denote by the history of past actions and rewards until the end of round $t-1$. Recall that $\mathbf{E}_{\Theta_t}[\cdot] := \mathbf{E}[\cdot \mid \Theta_t]$ and $\Pr_{\Theta_t}(\cdot) := \Pr(\cdot \mid \Theta_t)$. Denote by $\mathcal{E}_t := \left\{\forall a \in [N] : |r_a - \hat{r}_{a,n_{a,t}}| \leq \sqrt{\frac{3 \ln Nt}{n_{a,t}+1}}\right\}$ the high-probability event that the empirical mean is close to the true mean reward for arm a, and by $\overline{\mathcal{E}_t}$ the complementary event of $\mathcal{E}_t$. Recall that $\tilde{w}_{a,t} \sim \mathcal{N}(\hat{r}_{a,n_{a,t}}, \frac{\gamma m \ln t}{n_{a,t}+1})$ is i.i.d. of $w_{a,t}$ for CTS-G, and $\tilde{r}_t := \hat{r}_{a,n_{a,t}} + \tilde{w}_t\sqrt{\frac{\gamma \ln t}{n_{a,t}+1}}$, where $\tilde{w}_t \sim \mathcal{N}(0, 1)$ is i.i.d. of w_t for CL-SG.

Fact 1. *For a Gaussian distributed random variable Z with mean μ and variance δ^2, for any z, we have that*

$$\frac{1}{4\sqrt{\pi}} \cdot e^{-7z^2/2} \leq \Pr(|Z - \mu| > z\sigma) \leq \frac{1}{2}e^{-z^2/2}, \tag{6}$$

and for any $z > 0$,

$$\Pr(Z - \mu > z\sigma) \geq \frac{1}{\sqrt{2\pi}} \frac{z}{z^2 + 1} e^{-\frac{z^2}{2}}. \tag{7}$$

Fact 2. *Let $X_1, \ldots, X_N$ be N real random variables with $X_i \sim \mathrm{subG}\left(\sigma^2\right), i = 1, \ldots, N$, not necessarily independent. Then,*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} |X_i|\right] \leq \sigma\sqrt{2\log(2N)}.$$

C Proofs for Theorem 1

C.1 Proof of Lemma 1

Lemma 1. *In any round $t \geq \max\{\sqrt{m}, 4\}$, the optimism part in CTS-G satisfies that*

$$\mathbf{E}\left[\sum_{t=\max\{\sqrt{m},4\}}^{T} \left(\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} w_{a,t}\right)\right] \leq 8\sqrt{3\gamma}\Phi(-\sqrt{4/\gamma})^{-1} m \ln T\sqrt{NT}. \tag{8}$$

Proof. For each $a \in [N]$, we let $\tilde{w}_{a,t} \sim \mathcal{N}\left(\hat{r}_{a,n_{a,t}}, \frac{m\gamma \ln t}{n_{a,t}+1}\right)$ be an independent copy of $w_{a,t}$. Let $(\cdot)^+ := \max\{\cdot, 0\}$. Let $\mathbf{w}$ collect all the Gaussian random variables $w_{a,t}$ for all $a \in [N]$. Recall that $\mathbf{E}_{\Theta_t}[\cdot] := \mathbf{E}[\cdot \mid \Theta_t]$. There are three steps for the proofs.

Step 1: we show that in each round $t \geq \max\{\sqrt{m}, 4\}$, we have

$$\mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} w_{a,t}\right] \leq 2\Phi(-\sqrt{4/\gamma})^{-1}\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+\right]. \tag{9}$$

Step 2: we further bound the expectation term in the RHS of (9) as follows.

$$\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+\right] \leq \mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right]. \tag{10}$$

Step 3: summing over T, we show that (10) is upper bounded as follows.

$$\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right] \leq 4m \ln T\sqrt{3\gamma NT} \tag{11}$$

Combining these three steps, we have

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=\max\{\sqrt{m},4\}}^{T} \left(\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} w_{a,t}\right)\right] \\
&\leq 2\Phi(-\sqrt{4/\gamma})^{-1}\mathbf{E}\left[\sum_{t=\max\{\sqrt{m},4\}}^{T}\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right] \\
&\leq 2\Phi(-\sqrt{4/\gamma})^{-1}\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right] \\
&\leq 8\sqrt{3\gamma}\Phi(-\sqrt{4/\gamma})^{-1} m \ln T\sqrt{NT}.
\end{aligned} \tag{12}$$

Now, we give the details for these three steps.

Let $\alpha := \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} w_{a,t}\right]$.

Step 1 proof. If $\alpha = \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} w_{a,t}\right] \le 0$, the proof is trivial as the RHS of (9) is non-negative. Note that $2\Phi(-\sqrt{4/\gamma})^{-1} < +\infty$

For the case where $\alpha > 0$, we view $\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+ \ge 0$ as a non-negative random variable and use Markov's inequality. We have

$$\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+\right] \ge \alpha \Pr_{\Theta_t}\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right] \ge \alpha\right), \tag{13}$$

which gives

$$\begin{aligned}
\alpha &\le \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right] \ge \alpha\right)} \\
&= \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right] \ge \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t^*} r_a - \sum\limits_{a\in A_t} w_{a,t}\right]\right)} \\
&= \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum\limits_{a\in A_t} w_{a,t} \ge \sum\limits_{a\in A_t^*} r_a\right)} \overset{(a)}{\le} \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum\limits_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum\limits_{a\in A_t} w_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum\limits_{a\in A_t^*} w_{a,t} \ge \sum\limits_{a\in A_t^*} r_a\right)} \\
&\overset{(b)}{\le} 2\Phi(-\sqrt{4/\gamma})^{-1} \cdot \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+\right],
\end{aligned} \tag{14}$$

where step (a) is due to that A_t is the optimal super arm, and thus, we have $\sum_{a\in A_t^*} w_{a,t} \le \sum_{a\in A_t} w_{a,t}$ and step (b) uses the result shown in Lemma 4.

Step 2 proof. Recall that $w_{a,t}$ and $\tilde{w}_{a,t}$ are i.i.d. according to $\mathcal{N}\left(\hat{r}_{a,n_{a,t}}, \frac{m\gamma \ln t}{n_{a,t}+1}\right)$, and A_t is the optimal super arm based on Θ_t and $\mathbf{w}$. We have $\mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right] = \mathbf{E}_{\Theta_t}\left[\max_{A\in\Theta_t}\sum_{a\in A} w_{a,t}\right] = \mathbf{E}_{\Theta_t}\left[\max_{A\in\Theta_t}\sum_{a\in A} \tilde{w}_{a,t}\right] \ge \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \tilde{w}_{a,t} \mid A_t\right] = \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \tilde{w}_{a,t} \mid A_t, \mathbf{w}\right]$. Then, we have

$$
\begin{aligned}
\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} w_{a,t}\right]\right)^+\right] &\le \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \tilde{w}_{a,t} \mid A_t\right]\right)^+\right] \\
&= \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \tilde{w}_{a,t} \mid A_t, \mathbf{w}\right]\right)^+\right] \\
&= \mathbf{E}_{\Theta_t}\left[\left(\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right) \mid A_t, \mathbf{w}\right]\right)^+\right] \\
&\le \mathbf{E}_{\Theta_t}\left[\left|\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right) \mid A_t, \mathbf{w}\right]\right|\right] \\
&\le \mathbf{E}_{\Theta_t}\left[\mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right| \mid A_t, \mathbf{w}\right]\right] \\
&= \mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right],
\end{aligned} \tag{15}
$$

where the last inequality is due to Jensen's inequality.

Step 3 proof. Since $w_{a,t} - \tilde{w}_{a,t} \sim \mathcal{N}\left(0, \frac{2\gamma m \ln t}{n_{a,t}+1}\right)$, we can express $w_{a,t} - \tilde{w}_{a,t}$ as $\sqrt{2}\zeta_{a,t}\delta_{a,t}$, where $\zeta_{a,t} \sim \mathcal{N}(0, 1)$ and $\delta_{a,t} = \sqrt{\frac{\gamma m \ln t}{n_{a,t}+1}}$. Thus, we have

$$
\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \tilde{w}_{a,t}\right|\right] &\le \sqrt{2}\mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t} |\zeta_{a,t}\delta_{a,t}|\right] \\
&\overset{(a)}{\le} \sqrt{2}\mathbf{E}\left[\max_{t\in[T], a\in[N]} |\zeta_{a,t}| \sum_{t=1}^{T}\sum_{a\in A_t} |\delta_{a,t}|\right] \\
&= \sqrt{2}\mathbf{E}\left[\max_{t\in[T], a\in[N]} |\zeta_{a,t}| \sum_{t=1}^{T}\sum_{a\in A_t} \sqrt{\frac{\gamma m \ln t}{n_{a,t} + 1}}\right] \\
&\overset{(b)}{\le} 2m\sqrt{2\gamma NT \ln T}\,\mathbf{E}\left[\max_{t\in[T], a\in[N]} \zeta_{a,t}\right] \\
&\overset{(c)}{\le} 2m\sqrt{2\gamma NT \ln T} \cdot \sqrt{6 \ln T} \\
&\le 4m \ln T\sqrt{3\gamma NT}.
\end{aligned} \tag{16}
$$

where step (a) is due to Hölder's inequality. Step (b) is due to Lemma 5 such that $\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}+1}} \le 2\sqrt{mNT}$. Step (c) is due to the maximal inequality for Gaussian variables (Fact 2) such that $\mathbf{E}\left[\max_{t\in[T], a\in[N]} \zeta_{a,t}\right] \le \sqrt{2 \ln 2NT} \le \sqrt{6 \ln T}$ because $2 \le N \le T$. $\square$

C.2 Proof of Lemma 2

Lemma 2. *Let* $\mathcal{E}_t := \left\{\forall a \in [N], \hat{r}_{a,n_{a,t}} - r_a \le \sqrt{\frac{3 \ln Nt}{n_{a,t}+1}}\right\}$. *In CTS-G, the regret of the deviation part is*

$$
\mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] \le 2m \ln T\sqrt{6\gamma NT} + 2\sqrt{6mNT \ln T}.
$$

Proof. We can do decomposition as follows.

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \hat{r}_{a,n_{a,t}} + \sum_{a\in A_t} \hat{r}_{a,n_{a,t}} - \sum_{a\in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] \\
&= \mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t} w_{a,t} - \sum_{a\in A_t} \hat{r}_{a,n_{a,t}}\right)\mathbf{1}[\mathcal{E}_t]\right] + \mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t} \hat{r}_{a,n_{a,t}} - \sum_{a\in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] \\
&\overset{(a)}{\leq} \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\left(w_{a,t} - \hat{r}_{a,n_{a,t}}\right)\right] + \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{6\ln T}{n_{a,t}+1}}\right] \\
&\overset{(b)}{\leq} \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\left(w_{a,t} - \hat{r}_{a,n_{a,t}}\right)\right] + 2\sqrt{6mNT\ln T},
\end{aligned} \tag{17}$$

where step (a) is because event $\mathcal{E}_t$ is true and $\ln NT \leq 2\ln T$ because of $N \leq T$, and step (b) is due to Lemma 5 such that $\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}+1}} \leq 2\sqrt{mNT}$.

We can represent each $w_{a,t} - \hat{r}_{a,n_{a,t}}$ by $\zeta_{a,t}\delta_{a,t}$, where $\zeta_{a,t} \sim \mathcal{N}(0,1)$ and $\delta_{a,t} = \sqrt{\frac{\gamma m \ln t}{n_{a,t}+1}}$. Then, we can bound the first term on the RHS of the above equation as follows:

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\left(w_{a,t} - \hat{r}_{a,n_{a,t}}\right)\right] &\leq \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\zeta_{a,t}\delta_{a,t}\right] \\
&\overset{(a)}{\leq} \mathbf{E}\left[\max_{t\in[T],a\in[N]}|\zeta_{a,t}|\cdot\sum_{t=1}^{T}\sum_{a\in A_t}|\delta_{a,t}|\right] \\
&= \mathbf{E}\left[\max_{t\in[T],a\in[N]}|\zeta_{a,t}|\cdot\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{\gamma m\ln t}{n_{a,t}+1}}\right],
\end{aligned} \tag{18}$$

where (a) is due to Hölder's inequality. By invoking Lemma 5 again, we have that

$$\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{\gamma m\ln t}{n_{a,t}+1}} \leq \sqrt{\gamma m\ln T}\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}+1}} \leq 2m\sqrt{\gamma NT\ln T}. \tag{19}$$

Then, using the maximal inequality (Fact 2), we have $\mathbf{E}\left[\max_{t\in[T],a\in[N]}|\zeta_{a,t}|\right] \leq \sqrt{2\ln 2NT} \leq \sqrt{6\ln T}$, where the last inequality is due to that $2 \leq N \leq T$. Thus, we have

$$\mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\left(w_{a,t} - \hat{r}_{a,n_{a,t}}\right)\right] \leq 2m\ln T\sqrt{6\gamma NT}. \tag{20}$$

Finally, by substituting (20) into (17), we complete the proof. □

C.3 Proof of Lemma 3

Lemma 3. *The probability that event $\overline{\mathcal{E}_t}$ to happen satisfies that*

$$\sum_{t=1}^{T}\Pr(\overline{\mathcal{E}_t}) \leq \frac{\pi^2}{3}.$$

Proof. By a union bound and Hoeffding's inequality, we have that

$$\begin{aligned}
&\sum_{t=1}^{T} \Pr\left(\exists a \in [N] : |r_a - \hat{r}_{a,n_{a,t}}| > \sqrt{\frac{3 \ln Nt}{n_{a,t} + 1}}\right) \\
\leq\; &\sum_{t=1}^{T} \sum_{a\in[N]} \sum_{s=0}^{t-1} \Pr\left(|\hat{r}_{a,s} - r_a| > \sqrt{\frac{3 \ln Nt}{s + 1}}\right) \\
=\; &\sum_{a\in[N]} \sum_{t=1}^{T} \left(\Pr\left(r_a > \sqrt{3 \ln Nt}\right) + \sum_{s=1}^{t-1} \Pr\left(|\hat{r}_{a,s} - r_a| > \sqrt{\frac{3 \ln Nt}{s + 1}}\right)\right) \qquad (21) \\
\overset{\text{(a)}}{\leq}\; &\sum_{a\in[N]} \left(0 + \sum_{t=1}^{T} \sum_{s=1}^{t-1} \Pr\left(|\hat{r}_{a,s} - r_a| > \sqrt{\frac{3 \ln Nt}{2s}}\right)\right) \\
\leq\; &N \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \frac{2}{(Nt)^3} = \frac{\pi^2}{3N^2},
\end{aligned}$$

where step (a) is due to $r_a \in [0, 1], \forall a \in [N]$ and $3 \ln Nt > 1$ because $N \geq 2$, and that $s + 1 \leq 2s$ for any $s \geq 1$.

□

C.4 Proof of Lemma 4

Lemma 4. *In each round* $t \geq \max\{\sqrt{m}, 4\}$*, given any* Θ_t*, we have*

$$\frac{1}{\Pr_{\Theta_t}\left(\sum_{a\in A_t^*} w_{a,t} \geq \sum_{a\in A_t^*} r_a\right)} \leq 2\Phi\left(-\sqrt{4/\gamma}\right)^{-1},$$

where $\Phi(\cdot)$ *is the cdf of the standard Gaussian distribution.*

Proof. Given Θ_t, A_t^* is determined. Define $\mathcal{H}_t := \left\{\forall a \in A_t^* : |r_a - \hat{r}_{a,n_{a,t}}| \leq \sqrt{\frac{4 \ln t}{n_{a,t}+1}}\right\}$. Since $t \geq \max\{\sqrt{m}, 4\}$, we have that

$$\begin{aligned}
\Pr_{\Theta_t}(\mathcal{H}_t) &\geq 1 - \sum_{a\in A_t^*} \sum_{s_a=0}^{t-1} \Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \geq \sqrt{\frac{4 \ln t}{s_a + 1}}\right) \\
&= 1 - \sum_{a\in A_t^*} \sum_{s_a=1}^{t-1} \Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \geq \sqrt{\frac{4 \ln t}{s_a + 1}}\right) \\
&\geq 1 - \sum_{a\in A_t^*} \sum_{s_a=1}^{t-1} \Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \geq \sqrt{\frac{4 \ln t}{2s_a}}\right) \qquad (22) \\
&\geq 1 - mt \cdot 2 \cdot e^{-2\cdot s_a \cdot 4 \ln t/(2s_a)} \\
&= 1 - \frac{2mt}{t^4} \\
&\geq 1 - \frac{2}{t} \\
&\geq 0.5 \quad .
\end{aligned}$$

We have

$$
\begin{aligned}
\Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} \;\geq\; \sum_{a \in A_t^*} r_a\right) \;&\geq\; \Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} \;\geq\; \sum_{a \in A_t^*} r_a,\, \mathcal{H}_t\right) \\
&=\; \Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} - \hat{r}_{a,n_{a,t}} \;\geq\; \sum_{a \in A_t^*} r_a - \hat{r}_{a,n_{a,t}},\, \mathcal{H}_t\right) \\
&=\; \Pr_{\Theta_t}(\mathcal{H}_t) \cdot \Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} - \hat{r}_{a,n_{a,t}} \;\geq\; \sum_{a \in A_t^*} r_a - \hat{r}_{a,n_{a,t}} \;\middle|\; \mathcal{H}_t\right) \\
&\overset{\text{(a)}}{\geq}\; 0.5 \cdot \Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} - \hat{r}_{a,n_{a,t}} \;\geq\; \sum_{a \in A_t^*} \sqrt{\frac{4\ln t}{n_{a,t}+1}}\right) \\
&\overset{\text{(b)}}{\geq}\; 0.5 \cdot \Pr_{\Theta_t}\left(\sum_{a \in A_t^*} w_{a,t} - \hat{r}_{a,n_{a,t}} \;\geq\; \sqrt{m \sum_{a \in A_t^*} \frac{4\ln t}{n_{a,t}+1}}\right) \\
&=\; 0.5 \cdot \Phi\left(-\sqrt{4/\gamma}\right) \quad ,
\end{aligned}
\tag{23}
$$

where step (a) is due to (22) and the fact that event $\mathcal{E}_t$ is true. Step (b) uses the Cauchy–Schwarz inequality, i.e., we have $\sum_{a \in A_t^*} \sqrt{\frac{4\ln t}{n_{a,t}+1}} \leq \sqrt{m \cdot \sum_{a \in A_t^*} \frac{4\ln t}{n_{a,t}+1}}$. The last equality is due to the standardization of Gaussian distribution. $\square$

C.5 Proof of Lemma 5

Lemma 5. *We have* $\sum_{t=1}^{T} \sum_{a \in A_t} \sqrt{\frac{1}{n_{a,t}+1}} \leq 2\sqrt{mNT}$.

Proof. Note that the LHS of the above inequality is a random variable. We provide an upper bound for this random variable.

Recall $n_{a,t} := \sum_{\tau=1}^{t-1} \mathbf{1}\left[a \in A_\tau\right]$ is the number of times that arm a has been played at the beginning of round t. Let $\tau_a(n)$ denote the round for arm a to be played for the n-th time, and thus $n_{a,\tau_a(n)} = n - 1$.

$$
\begin{aligned}
\sum_{t=1}^{T} \sum_{a \in A_t} \sqrt{\frac{1}{n_{a,t}+1}} &= \sum_{t=1}^{T} \sum_{a \in [N]} \sqrt{\frac{1}{n_{a,t}+1}}\mathbf{1}[a \in A_t] \\
&\overset{\text{(a)}}{=} \sum_{a \in [N]} \sum_{n=1}^{n_{a,T+1}} \sum_{t=\tau_a(n)}^{\tau_a(n+1)-1} \sqrt{\frac{1}{n_{a,t}+1}}\mathbf{1}[a \in A_t] \\
&\overset{\text{(b)}}{=} \sum_{a \in [N]} \sum_{n=1}^{n_{a,T+1}} \sqrt{\frac{1}{n}} \leq \sum_{a \in [N]} \int_0^{n_{a,T+1}} \sqrt{\frac{1}{n}} dn \\
&= 2 \sum_{a \in [N]} \sqrt{n_{a,T+1}} \overset{\text{(c)}}{\leq} 2\sqrt{N \sum_{a \in [N]} n_{a,T+1}} \\
&\overset{\text{(d)}}{=} 2\sqrt{mNT},
\end{aligned}
\tag{24}
$$

where step (a) partitions all T rounds into multiple intervals based on the arrivals of observations from arm a. Step (b) uses the fact that $\sum_{t=\tau_a(n)}^{\tau_a(n+1)-1} \mathbf{1}[a \in A_t] \cdot \sqrt{\frac{1}{n_{a,t}+1}} = \sqrt{\frac{1}{n-1+1}} = \sqrt{\frac{1}{n}}$, because $n_{a,\tau_a(n)} = n - 1$ and $\mathbf{1}[a \in A_t] = 0$ for all $t \in \{\tau_a(n) + 1, \ldots, \tau_a(n+1) - 1\}$. Step (c) uses Cauchy-Schwarz inequality. Step (d) uses the fact that $\sum_{a \in [N]} n_{a,T+1} \leq mT$.

$\square$

C.6 Proof of Upper bound

Upper Bound Proof of Theorem 1. Denote by $\mathcal{E}_t := \left\{\forall a \in [N] : |r_a - \hat{r}_{a,n_{a,t}}| \leq \sqrt{\frac{3 \ln Nt}{n_{a,t}+1}}\right\}$ the high-probability event that the empirical mean reward is close to the true mean reward for arm a, and by $\overline{\mathcal{E}_t}$ the complementary event of $\mathcal{E}_t$.

Let $t' = \max\{\sqrt{m}, 4\}$. We first decompose the regret as follows:

$$\begin{aligned}
\mathcal{R}(T) &= \sum_{t=1}^{t'-1} \mathbf{E}\left[\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} r_a\right] + \sum_{t=t'}^{T} \mathbf{E}\left[\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} r_a\right] \\
&\overset{\text{(a)}}{\leq} m \max\{\sqrt{m}, 4\} + \mathbf{E}\left[\sum_{t=t'}^{T}\left(\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] + \mathbf{E}\left[\sum_{t=t'}^{T}\left(\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} r_a\right)\mathbf{1}[\overline{\mathcal{E}_t}]\right] \\
&\overset{\text{(b)}}{\leq} m \max\{\sqrt{m}, 4\} + \mathbf{E}\left[\sum_{t=t'}^{T}\left(\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} w_{a,t} + \sum_{a \in A_t} w_{a,t} - \sum_{a \in A_t} r_a\right)\mathbf{1}[\mathcal{E}_t]\right] + m\frac{\pi^2}{3N^2} \\
&\leq \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\left(\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} w_{a,t}\right)\right]}_{=:I_1,\ \text{optimism part}} + \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\sum_{a \in A_t} (w_{a,t} - r_a)\,\mathbf{1}[\mathcal{E}_t]\right]}_{=:I_2,\ \text{deviation part}} + m \max\{\sqrt{m}, 4\} + \frac{\pi^2}{3},
\end{aligned} \tag{25}$$

where step (a) is due to the fact that $\sum_{a \in A_t^*} r_a - \sum_{a \in A_t} r_a \leq m$ by the definition of r_a and m and step (b) is due to Lemma 3.

Now, invoking Lemma 1 with proofs in Appendix C.1, we have term I_1 bounded as follows:

$$I_1 \leq 8\sqrt{3\gamma}\Phi(-\sqrt{4/\gamma})^{-1} m \ln T\sqrt{NT}, \tag{26}$$

and I_2 can be bounded by using Lemma 2 with proofs in Appendix C.2:

$$I_2 \leq 2m \ln T\sqrt{6\gamma NT} + 2\sqrt{6mNT \ln T}. \tag{27}$$

Thus, we have that

$$\begin{aligned}
\mathcal{R}(T) \leq& \left(2\sqrt{6\gamma} + 8\sqrt{3\gamma}\Phi(-\sqrt{4/\gamma})^{-1}\right) m \ln T\sqrt{NT} + 2\sqrt{6mNT \ln T} \\
&+ m\left(\max\{\sqrt{m}, 4\} + \frac{\pi^2}{3}\right).
\end{aligned} \tag{28}$$

Using numerical optimization methods searching from $\gamma = 0.0001$ to $\gamma = 100$, we can find that when $\gamma = 6.4$, the coefficient for the first item can achieve a minimum value of 175.74.

□

C.7 Proof of Lower Bound

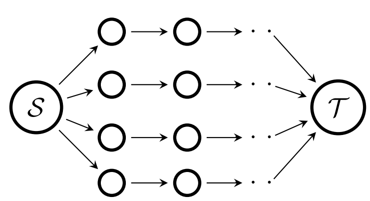


Figure 5: Problem instance for the lower-bound proof. Nodes $\mathcal{S}$ and $\mathcal{T}$ are the starting and ending points for each path.

Lower bound Proof in Theorem 1. Our proof uses similar ideas to the proofs of Theorem 1.4 in Agrawal & Goyal (2017b).

We construct a path selection problem involving N links (each link corresponds to a base arm) and K paths (each path corresponds to a super arm), as illustrated in Fig. 5. Each path consists of m links, and thus, the total number of base arms $N = mK$. We consider a fixed availability set throughout all T rounds, i.e., $\Theta_t = \Theta := \{A_1, A_2, \ldots, A_K\}$ for all rounds $t \in [T]$ with each super arm A_k being a feasible path. We assume the first path A_1 is the unique optimal one.

We construct the following Bernoulli reward distributions for each base arm. Let $\Delta := \sqrt{K \ln K/T}$. For any base arm in the optimal super arm A_1, we use a degenerate distribution putting mass 1 on a single point $\sqrt{\gamma}\Delta$, i.e., if A_1 is played, for any base arm in it, we always observe $\sqrt{\gamma}\Delta$ as the random reward. Similarly, for the remaining base arms in the sub-optimal super arms, we put mass 1 on a single point 0, i.e., the random reward is always 0 for any base arm in a sub-optimal super arm.

Let $Q_A(t)$ denote the number of times that super arm $A \in \Theta$ has been played at the beginning of round t. Since there are no overlapping base arms between two distinct super arms, we have $Q_A(t) = n_{a,t}$ for all $a \in A$. Let $c \in (0, 1)$ be some universal constant that will be tuned later. Define $B_t^* := \{Q_{A_1}(t) > t - cT\}$ as the event that the optimal super arm A_1 has been observed at least $(t - cT)$ times by the beginning of round t.

We lower bound the total regret from round 1 to the end of round T by analyzing two cases that are exhaustive and mutually exclusive based on events B_t^* for all rounds $t \in [T]$.

If B_t^* is not true for some $t \in [T]$, we have the total number of times of playing sub-optimal super arms by the beginning of round t is $\sum_{A \in \Theta \setminus A_1} Q_A(t) = t - Q_{A_1}(t) \geq t - (t - cT) \geq cT$, which implies the total regret by the end of round T is at least $cT \cdot m \cdot \sqrt{\gamma}\Delta = \Omega(m\sqrt{KT \ln K}) = \Omega(\sqrt{mNT \ln(N/m)})$. Note that the total regret from round 1 to round t is a lower bound for the total regret over all T rounds.

If B_t^* is true for all $t \in [T]$, we have the total number of times $\sum_{A \in \Theta \setminus A_1} Q_A(t)$ of playing sub-optimal super arms by the beginning of round t is upper bounded by

$$\textstyle\sum_{A \in \Theta \setminus A_1} Q_A(t) = t - Q_{A_1}(t) \leq t - (t - cT) = cT \quad . \tag{29}$$

Due to the spread of a sub-optimal super arm's posterior distribution, the learning agent will make mistakes when deciding which super arm to play. Formally, we show that with at least a constant probability, the learning agent will play a sub-optimal super arm. Note that whether event B_t^* is true or not is determined by the history information $\mathcal{F}_{t-1}$.

Recall $w_{a,t} \sim \mathcal{N}\left(\hat{r}_{a,n_{a,t}}, \frac{\gamma m \ln t}{n_{a,t}+1}\right)$. Now, we construct a lower bound for the probability of selecting a sub-optimal arm in round t conditioned on instantiations F_{t-1} of $\mathcal{F}_{t-1}$ such that B_t^* is true. We have

$$\begin{aligned}
&\Pr\left(\exists A \in \Theta \setminus A_1 : A_t = A \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
\geq\ &\Pr(\exists A \in \Theta \setminus A_1 : \sum_{a \in A} w_{a,t} > \sum_{a \in A_1} w_{a,t} \mid \mathcal{F}_{t-1} = F_{t-1}) \\
\overset{(a)}{\geq}\ &\Pr(\exists A \in \Theta \setminus A_1 : \sum_{a \in A} w_{a,t} \geq m\sqrt{\gamma}\Delta \mid \mathcal{F}_{t-1} = F_{t-1}) \\
&\cdot \Pr(\sum_{a \in A_1} w_{a,t} < m\sqrt{\gamma}\Delta \mid \mathcal{F}_{t-1} = F_{t-1}) \\
\overset{(b)}{=}\ &\Pr(\exists A \in \Theta \setminus A_1 : \sum_{a \in A} w_{a,t} \geq m\sqrt{\gamma}\Delta \mid \mathcal{F}_{t-1} = F_{t-1}) \cdot \frac{1}{2} \quad ,
\end{aligned} \tag{30}$$

where step (a) uses the fact that all super arms are independent based on our construction of the path selection problem. Step (b) uses the fact that the sum of multiple independent Gaussian random variables is still Gaussian and $\left(\sum_{a \in A_1} w_{a,t} - m\sqrt{\gamma}\Delta\right)$ is a zero-mean Gaussian distribution. Note that the empirical mean of each base arm in the optimal super arm is exactly $\sqrt{\gamma}\Delta$.

Now, we construct a lower bound for (30) by using Gaussian anti-concentration bounds. We have

$$
\begin{aligned}
&\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} w_{a,t} \geq m\sqrt{\gamma}\Delta \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
\geq\ & \Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} w_{a,t} \geq m\sqrt{\gamma}\Delta\sqrt{\ln t} \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
=\ & 1 - \Pr\left(\sum_{a\in A} w_{a,t} < m\Delta\sqrt{\gamma \ln t}, \forall A \in \Theta \setminus A_1 \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
=\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \Pr\left(\sum_{a\in A} w_{a,t}\sqrt{(Q_A(t)+1)} \geq m\Delta\sqrt{(Q_A(t)+1)\gamma \ln t} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\right) \\
=\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \underbrace{\Pr\left(\frac{\sum_{a\in A} w_{a,t}\sqrt{(Q_A(t)+1)}}{m\sqrt{\gamma \ln t}} \geq \Delta\sqrt{(Q_A(t)+1)} \mid \mathcal{F}_{t-1} = F_{t-1}\right)}_{\geq \frac{1}{8\sqrt{\pi}}e^{-\frac{7}{2}\Delta^2(Q_A(t)+1)}}\right) \\
\geq\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-\frac{7}{2}\Delta^2(Q_A(t)+1)}\right) \quad ,
\end{aligned}
\tag{31}
$$

where the last inequality uses the fact that $\frac{\sum_{a\in A} w_{a,t}\sqrt{Q_A(t)+1}}{m\sqrt{\gamma \ln t}} \sim \mathcal{N}(0,1)$ and then the one-sided anti-concentration inequality shown in (6).

Tune constant $c = 0.001$. Then, we use the upper bound constructed in (29) to continue lower bounding (31). We have

$$
\begin{aligned}
& 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-\frac{7}{2}\Delta^2(Q_A(t)+1)}\right) \\
\overset{(a)}{\geq}\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-\frac{7c}{2}\Delta^2\frac{\sqrt{KT \ln K}}{(K-1)\Delta} - \frac{7\Delta^2}{2}}\right) \\
=\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-\frac{7c}{2}\frac{K \ln K}{(K-1)} - \frac{7\frac{K \ln K}{T}}{2}}\right) \\
=\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-(\frac{7c}{2}\frac{K}{K-1} + \frac{7K}{2T}) \ln K}\right) \\
\overset{(b)}{\geq}\ & 1 - \prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{8\sqrt{\pi}}e^{-\ln K}\right) \\
=\ & 1 - \left(1 - \frac{1}{8\sqrt{\pi}K}\right)^{K-1} \\
\overset{(c)}{\geq}\ & 1 - \left(e^{-\frac{1}{8\sqrt{\pi}K}}\right)^{K-1} \\
\geq\ & 1 - e^{-\frac{1}{16\sqrt{\pi}}},
\end{aligned}
\tag{32}
$$

where step (a) is due to the fact that, constrained on (29), i.e., $\sum_{A\in\Theta\setminus A_1} Q_A(t) \leq cT = c\frac{\sqrt{KT \ln K}}{\Delta}$, the quantity $\prod_{A\in\Theta\setminus A_1}\left(1 - \frac{1}{4\sqrt{\pi}}e^{-\frac{7c}{2}\Delta^2\frac{\sqrt{KT \ln K}}{(K-1)\Delta} - \frac{7\Delta^2}{2}}\right)$ is maximized when $Q_A(t) = \frac{c\sqrt{KT \ln K}}{(K-1)\Delta}$ for all $A \in \Theta \setminus A_1$. Step (b) uses the fact that, when $c = 0.001$, we have $\frac{7c}{2}\frac{K}{K-1} + \frac{7K}{2T} \leq 1$ when T is sufficiently large, e.g., $T > 5K$. Step (c) uses $1 - x \leq e^{-x}$.

Now, we are ready to complete the proof. Let $p := \frac{1}{2}\left(1 - e^{-\frac{1}{16\sqrt{\pi}}}\right)$. By plugging the lower bound constructed in (32) into (30), we have $\Pr(\exists A \in \Theta \setminus A_1 : A_t = A \mid \mathcal{F}_{t-1} = F_{t-1}) \geq p$, which implies the total regret by the end of round T is at least $Tpm\sqrt{\gamma}\Delta = \Omega(\sqrt{mNT \ln(N/m)})$. $\square$

D Proofs for Theorem 2

D.1 Proof of Lemma 6

Lemma 6. *The optimism part in CL-SG satisfies that*

$$\mathbf{E}\left[\sum_{t=\max\{m,4\}}^{T}\left(\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right)\right] \leq 8\sqrt{2\gamma}\Phi(-\sqrt{4/\gamma})^{-1} \ln T\sqrt{mNT}. \tag{33}$$

Proof. Similar to the proof of Lemma 1. There are three steps for the proofs.

Step 1: Let $t' = \max\{\sqrt{m}, 4\}$ we show that the following inequality holds for each round t conditioned on Θ_t:

$$\mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right] \leq 2\Phi(-\sqrt{4/\gamma})^{-1} \cdot \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} \bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \bar{r}_{a,t}\right]\right)^+\right]. \tag{34}$$

Step 2: Let $\tilde{r}_{a,t} = \hat{r}_{a,t} + \tilde{w}_t\sqrt{\frac{\gamma \ln t}{n_{a,t}+1}}$, where $\tilde{w}_t \sim \mathcal{N}(0, 1)$ is an independent copy of w_t. With $\tilde{r}_{a,t}$, we can further bound the last term in (34) as follows.

$$\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t} \bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t} \bar{r}_{a,t}\right]\right)^+\right] \leq \mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t} \bar{r}_{a,t} - \sum_{a\in A_t} \tilde{r}_{a,t}\right|\right]. \tag{35}$$

Step 3: Summing over T rounds, we have that

$$\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t} \bar{r}_{a,t} - \sum_{a\in A_t} \tilde{r}_{a,t}\right|\right] \leq 4 \ln T\sqrt{2\gamma mNT} \tag{36}$$

Combining these three steps, we have

$$\begin{aligned}\mathbf{E}\left[\sum_{t=\max\{\sqrt{m},4\}}^{T}\left(\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right)\right] &\leq 2\Phi(-\sqrt{4/\gamma})^{-1}\mathbf{E}\left[\sum_{t=\max\{\sqrt{m},4\}}^{T}\left|\sum_{a\in A_t} \bar{r}_{a,t} - \sum_{a\in A_t} \tilde{r}_{a,t}\right|\right]\\ &\leq 2\Phi(-\sqrt{4/\gamma})^{-1}\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t} \bar{r}_{a,t} - \sum_{a\in A_t} \tilde{r}_{a,t}\right|\right]\\ &\leq 8\sqrt{2\gamma}\Phi(-\sqrt{4/\gamma})^{-1} \ln T\sqrt{mNT}.\end{aligned} \tag{37}$$

Now, we give the details for these three steps.

Step 1 proof. If $\mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right] \leq 0$, the proof is trivial as the RHS in (34) is non-negative.

Recall $(\cdot)^+ := \max\{\cdot, 0\}$. For the case where $\alpha := \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right] > 0$, we use Markov's inequality and have

$$\begin{aligned}\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+\right] &\geq \alpha\cdot\Pr_{\Theta_t}\left(\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+ \geq \alpha\right)\\ &\geq \alpha\cdot\Pr_{\Theta_t}\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right] \geq \alpha\right),\end{aligned} \tag{38}$$

which gives

$$\begin{aligned}\alpha &= \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t}\bar{r}_{a,t}\right]\\ &\leq \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right] \geq \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t^*} r_a - \sum_{a\in A_t}\bar{r}_{a,t}\right]\right)}\\ &= \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum_{a\in A_t}\bar{r}_{a,t} \geq \sum_{a\in A_t^*} r_a\right)} \leq \frac{\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+\right]}{\Pr_{\Theta_t}\left(\sum_{a\in A_t^*}\bar{r}_{a,t} \geq \sum_{a\in A_t^*} r_a\right)}\\ &= 2\Phi(-\sqrt{4/\gamma})^{-1}\cdot\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar{r}_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right]\right)^+\right].\end{aligned} \tag{39}$$

Step 2 proof. Since w_t and $\tilde{w}_t$ are i.i.d., we have $\mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar{r}_{a,t}\right] = \mathbf{E}_{\Theta_t}\left[\max_{A\in\Theta_t}\sum_{a\in A}\bar{r}_{a,t}\right] = \mathbf{E}_{\Theta_t}\left[\max_{A\in\Theta_t}\sum_{a\in A}\tilde{r}_{a,t}\right] \geq \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\tilde{r}_{a,t} \mid A_t\right] = \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\tilde{r}_{a,t} \mid A_t, w_t\right]$. Then, we have

$$
\begin{aligned}
\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar r_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\bar r_{a,t}\right]\right)^+\right] &\le \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar r_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\tilde r_{a,t} \mid A_t\right]\right)^+\right] \\
&= \mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar r_{a,t} - \mathbf{E}_{\Theta_t}\left[\sum_{a\in A_t}\tilde r_{a,t} \mid A_t, w_t\right]\right)^+\right] \\
&= \mathbf{E}_{\Theta_t}\left[\left(\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}\tilde r_{a,t}\right) \mid A_t, w_t\right]\right)^+\right] \\
&\le \mathbf{E}_{\Theta_t}\left[\left|\mathbf{E}_{\Theta_t}\left[\left(\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}\tilde r_{a,t}\right) \mid A_t, w_t\right]\right|\right] \\
&\le \mathbf{E}_{\Theta_t}\left[\mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}\tilde r_{a,t}\right| \mid A_t, w_t\right]\right] \\
&\le \mathbf{E}_{\Theta_t}\left[\left|\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}\tilde r_{a,t}\right|\right].
\end{aligned}
\tag{40}
$$

Step 3 proof. By Hölder's inequality, we have that

$$
\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T}\left|\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}\tilde r_{a,t}\right|\right] &\le \mathbf{E}\left[\sum_{t=1}^{T}|w_t - \tilde w_t|\sum_{a\in A_t}\sqrt{\frac{\gamma\ln t}{n_{a,t}+1}}\right] \\
&\le \mathbf{E}\left[\max_{t\in[T]}|w_t - \tilde w_t|\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{\gamma\ln t}{n_{a,t}+1}}\right] \\
&\overset{(a)}{\le} \mathbf{E}\left[\max_{t\in[T]}|w_t - \tilde w_t|\right]\cdot 2\sqrt{\gamma mNT\ln T} \\
&\overset{(b)}{\le} 2\sqrt{\ln 2T}\cdot 2\sqrt{\gamma mNT\ln T} \\
&\le 4\ln T\sqrt{2\gamma mNT},
\end{aligned}
\tag{41}
$$

where step (a) is due to Lemma 5, and step (b) is due to Fact 2 and $w_t - \tilde w_t$ is a Gaussian variable with variance 2. □

D.2 Proof of Lemma 7

Lemma 7. *In CL-SG, the regret of the deviation part is*

$$
\mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}r_a\right)\mathbf{1}[\mathcal{E}_t]\right] \le 4\ln T\sqrt{\gamma mNT} + 2\sqrt{6mNT\ln T}.
$$

Proof. Recall that $\bar r_{a,t} = \hat r_{a,n_{a,t}} + w_t\sqrt{\frac{\gamma\ln t}{n_{a,t}+1}}$. When $\mathcal{E}_t$ happens, we have that

$$
\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T}\left(\sum_{a\in A_t}\bar r_{a,t} - \sum_{a\in A_t}r_a\right)\mathbf{1}[\mathcal{E}_t]\right] &= \mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\left(\hat r_{a,n_{a,t}} + w_t\sqrt{\frac{\gamma\ln t}{n_{a,t}+1}} - \hat r_{a,n_{a,t}} + \sqrt{\frac{3\ln Nt}{n_{a,t}+1}}\right)\mathbf{1}[\mathcal{E}_t]\right] \\
&\le \sqrt{\gamma\ln T}\,\mathbf{E}\left[\sum_{t=1}^{T}w_t\left(\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}+1}}\right)\right] + \sqrt{6\ln T}\,\mathbf{E}\left[\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}+1}}\right],
\end{aligned}
\tag{42}
$$

where the last inequality is due to that $N \le T$. Regarding the first item in RHS of (42), we can apply Hölder's inequality to have that

$$\mathbf{E}\left[\sum_{t=1}^{T} w_t \left(\sum_{a \in A_t} \sqrt{\frac{1}{n_{a,t}\,+\,1}}\right)\right] \le \mathbf{E}\left[\max_{1\le t\le T} |w_t| \cdot \left|\sum_{t=1}^{T}\sum_{a\in A_t}\sqrt{\frac{1}{n_{a,t}\,+\,1}}\right|\right] \\ \le \mathbf{E}\left[\max_{1\le t\le T} |w_t| \cdot 2\sqrt{mNT}\right] \\ \le 4\sqrt{mNT\,\ln T}, \tag{43}$$

where the second inequality is due to Lemma 5, and the last inequality is due to the maximal inequality (Fact 2) for Gaussian variables such that $\mathbf{E}\,[\max_{1\le t\le T} |w_t|] \le \sqrt{2\,\ln 2T} \le 2\sqrt{T}$.

Regarding the second term in RHS of (42), we can invoke Lemma 5 again to give a bound of $2\sqrt{6mNT\,\ln T}$.

□

D.3 Proof of Lemma 8

Lemma 8. *In each round* $t > \max\{\sqrt{m}, 4\}$*, given any* Θ_t*, we have that for CL-SG:*

$$\frac{1}{\Pr_{\Theta_t}\left(\sum_{a\in A_t^*}\bar{r}_{a,t} \ge \sum_{a\in A_t^*} r_a\right)} \le 2\Phi\left(-\sqrt{4/\gamma}\right)^{-1}. \tag{44}$$

Proof of Lemma 8. Given Θ_t, A_t^* is determined. Define $\mathcal{H}_t := \left\{\forall a \in A_t^* : |r_a - \hat{r}_{a,n_{a,t}}| \le \sqrt{\frac{4\,\ln t}{n_{a,t}+1}}\right\}$. We have

$$\begin{aligned} \Pr_{\Theta_t}(\mathcal{H}_t) &\ge 1 - \sum_{a\in A_t^*}\sum_{s_a=0}^{t-1}\Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \ge \sqrt{\tfrac{4\,\ln t}{s_a+1}}\right) \\ &= 1 - \sum_{a\in A_t^*}\sum_{s_a=1}^{t-1}\Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \ge \sqrt{\tfrac{4\,\ln t}{s_a+1}}\right) \\ &\ge 1 - \sum_{a\in A_t^*}\sum_{s_a=1}^{t-1}\Pr_{\Theta_t}\left(|r_a - \hat{r}_{a,s_a}| \ge \sqrt{\tfrac{4\,\ln t}{2s_a}}\right) \\ &\ge 1 - mt \cdot 2 \cdot e^{-2\cdot s_a \cdot 4\,\ln t/(2s_a)} \\ &= 1 - \tfrac{2mt}{t^4} \\ &\ge 1 - \tfrac{2}{t} \\ &\ge 0.5\quad, \end{aligned} \tag{45}$$

where the last two inequalities are due to that $t > \max\{\sqrt{m}, 4\}$.

We have

$$
\begin{aligned}
\Pr_{\Theta_t}\left(\sum_{a\in A_t^*} \bar{r}_{a,t} \;\ge\; \sum_{a\in A_t^*} r_a\right) &\ge \Pr_{\Theta_t}\left(\sum_{a\in A_t^*} \bar{r}_{a,t} \;\ge\; \sum_{a\in A_t^*} r_a,\, \mathcal{H}_t\right)\\
&= \Pr_{\Theta_t}\left(\sum_{a\in A_t^*} \bar{r}_{a,t} - \hat{r}_{a,n_{a,t}} \;\ge\; \sum_{a\in A_t^*} r_a - \hat{r}_{a,n_{a,t}},\, \mathcal{H}_t\right)\\
&= \Pr_{\Theta_t}(\mathcal{H}_t)\cdot \Pr_{\Theta_t}\left(\sum_{a\in A_t^*} \bar{r}_{a,t} - \hat{r}_{a,n_{a,t}} \;\ge\; \sum_{a\in A_t^*} r_a - \hat{r}_{a,n_{a,t}} \;\middle|\; \mathcal{H}_t\right)\\
&\overset{(a)}{\ge} 0.5\cdot \Pr_{\Theta_t}\left(\sum_{a\in A_t^*} w_t\sqrt{\frac{\gamma\ln t}{n_{a,t}+1}} \;\ge\; \sum_{a\in A_t^*} \sqrt{\frac{4\ln t}{n_{a,t}+1}}\right)\\
&= 0.5\cdot \Pr_{\Theta_t}\left(w_t \;\ge\; \sqrt{4/\gamma}\right)\\
&= 0.5\cdot \Phi\left(-\sqrt{4/\gamma}\right) \quad ,
\end{aligned}
\tag{46}
$$

where step (a) is due to (45) and the fact that event $\mathcal{E}_t$ is true. □

D.4 Proof of Upper Bound

Proof. Recall that $\mathcal{E}_t := \left\{\forall a \in [N] : |r_a - \hat{r}_{a,n_{a,t}}| \le \sqrt{\frac{3\ln Nt}{n_{a,t}+1}}\right\}$ is the high-probability event that the empirical mean reward is close to the true mean reward for arm a, and $\overline{\mathcal{E}_t}$ is the complementary event of $\mathcal{E}_t$.

Similar to the proof of the upper bound for CTS-G, we first let $t' = \max\{\sqrt{m}, 4\}$, and then decompose the regret as follows:

$$
\begin{aligned}
\mathcal{R}(T) \;\le\; & \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\left(\sum_{a\in A_t^*} r_a - \sum_{a\in A_t} \bar{r}_{a,t}\right)\right]}_{=:I_1,\ \text{optimism part}} + \underbrace{\sum_{t=t'}^{T} \mathbf{E}\left[\sum_{a\in A_t} (\bar{r}_{a,t} - r_a)\, \mathbf{1}[\mathcal{E}_t]\right]}_{=:I_2,\ \text{deviation part}}\\
& + m\left(\max\{\sqrt{m}, 4\} + \frac{\pi^2}{3}\right),
\end{aligned}
\tag{47}
$$

Now, invoking Lemma 6 with proofs in Appendix D.1, we have term I_1 bounded as follows:

$$
I_1 \;\le\; 8\sqrt{2\gamma}\Phi(-\sqrt{4/\gamma})^{-1}\ln T\sqrt{mNT}, \tag{48}
$$

and I_2 can be bounded by using Lemma 7 with proofs in Appendix D.2:

$$
I_2 \;\le\; 4\ln T\sqrt{\gamma mNT} + 2\sqrt{6mNT\ln T}. \tag{49}
$$

Therefore, we have the regret bounded as follows:

$$
\begin{aligned}
\mathcal{R}(T) \;\le\; & \left(4\sqrt{\gamma} + 8\sqrt{2\gamma}\Phi(-\sqrt{4/\gamma})^{-1}\right)\ln T\sqrt{mNT} + 2\sqrt{6mNT\ln T}\\
& + m\left(\max\{\sqrt{m}, 4\} + \frac{\pi^2}{3}\right),
\end{aligned}
$$

where the coefficient of the first term can be minimized to 144.43 at $\gamma = 4.57$. □

D.5 Proof of Lower Bound

Lower bound Proof in Theorem 2. The main challenge in the proofs arises from the fact that all base arms share a single random Gaussian seed, creating dependencies between paths that are no longer independent. However, the lower-bound proof for Theorem 1 relies on the independence of each super arm. Therefore, this proof must manage these dependencies effectively.

Figure 6: The regret of CL-SG is lower bounded by the regret from rounds αT to T.

We construct a path selection problem involving N links (i.e., each link corresponds to a base arm) and K paths (i.e., each path corresponds to a super arm). Each path consists of m links as illustrated in Fig. 5 and the total number of base arms is $N = mK$. We use a fixed availability set throughout all T rounds, i.e., $\Theta_t = \Theta := \{A_1, A_2, \ldots, A_K\}$ for all rounds $t \in [T]$ with each A_k being a feasible path. We assume the first path is the unique optimal one.

We construct the following Bernoulli reward distributions for each base arm. Let $\Delta := \sqrt{K/T}$. For any base arm in the optimal super arm A_1, we use a degenerate distribution putting mass 1 on a single point $\sqrt{\gamma}\Delta$, i.e., if A_1 is played, for each base arm in it, the observed random reward is always $\sqrt{\gamma}\Delta$. Similarly, for the remaining base arms in the sub-optimal super arms, we put mass 1 on a single point 0, i.e., the random reward is always 0 for any base arm in a sub-optimal super arm.

Let $Q_A(t)$ denote the total number of times that super arm $A \in \Theta$ has been played at the beginning of round t. Since there are no overlapping base arms between two distinct super arms, we have $Q_A(t) = n_{a,t}$ for all $a \in A$, i.e., all base arms in a super arm have the same amount of observations.

Let $c := \frac{1}{6}$. Define $B_t^* := \{Q_{A_1}(t) > t - cT\}$ as the event that the optimal super arm A_1 has been observed enough times at the beginning of round t.

We lower bound the total regret from round 1 to the end of round T by analyzing two cases that are exhaustive and mutually exclusive based on events B_t^* for all rounds $t \in [T]$.

If B_t^* is not true for some round $t \in [T]$, we have the total number of times of playing sub-optimal super arms until the beginning of round t is $\sum_{A\in\Theta\setminus A_1} Q_A(t) = t - Q_{A_1}(t) \geq cT$. This lower bound implies the total regret from round 1 to the end of round $t-1$ is at least $cT \cdot m \cdot \sqrt{\gamma}\Delta = \Omega(Tm\sqrt{K/T}) = \Omega(Tm\sqrt{N/(mT)}) = \Omega(\sqrt{mNT})$. Note that this lower bound is also a regret lower bound for the total regret from round 1 to the end of round T.

Let $\alpha := \frac{5}{6}$.

If B_t^* is true for all rounds $t \in [T]$, the total regret from round 1 to the end of round T is lower bounded by the total regret from round $t = \alpha T$ to the end of round T, as shown in Fig. 6. In each round $t \geq \alpha T$, we have the following inequalities:

$$\textstyle\sum_{A\in\Theta\setminus A_1} Q_A(t) = t - Q_{A_1}(t) \leq t - (t - c\cdot T) = c \cdot T \quad , \tag{50}$$

$$Q_{A_1}(t) > t - c\cdot T \geq \alpha \cdot T - c \cdot T = (\alpha - c)\cdot T \quad , \tag{51}$$

and

$$Q_A(t) \leq c \cdot T, \quad \forall A \in \Theta \setminus A_1 \quad . \tag{52}$$

From (51) and (52), for each sub-optimal super arm $A \in \Theta \setminus A_1$, we have

$$\sqrt{\tfrac{Q_A(t)+1}{Q_{A_1}(t)+1}} \leq \sqrt{\tfrac{cT+1}{(\alpha-c)T+1}} = \sqrt{\tfrac{T/6+1}{4T/6+1}} \leq \tfrac{1}{2} \quad , \tag{53}$$

which gives

$$1 - \sqrt{\tfrac{Q_A(t)+1}{Q_{A_1}(t)+1}} \geq \tfrac{1}{2} \quad . \tag{54}$$

Let $p_0 := \frac{1}{8\sqrt{\pi}}e^{-\frac{28}{3}}$. In the following, we prove that, with at least a constant probability p_0, a sub-optimal super arm is played in each round $t \geq \alpha \cdot T$ conditioned on event B_t^* is true. Note that whether event B_t^* is true or not is determined by the history information $\mathcal{F}_{t-1}$.

Conditioned on any instantiation F_{t-1} of $\mathcal{F}_{t-1}$ such that event B_t^* is true, we have the probability of playing a sub-optimal super arm is

$$\begin{aligned}
&\Pr\left(\exists A \in \Theta \setminus A_1 : A_t = A \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
\geq\ &\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} \bar{r}_{a,t} > \sum_{b\in A_1} \bar{r}_{b,t} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
=\ &\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} \hat{r}_{a,n_{a,t}} + w_t\sqrt{\frac{\gamma \ln t}{n_{a,t}+1}} > \sum_{b\in A_1} \hat{r}_{b,n_{b,t}} + w_t\sqrt{\frac{\gamma \ln t}{n_{b,t}+1}} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
\overset{\text{(a)}}{=}\ &\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} w_t\sqrt{\frac{\gamma \ln t}{n_{a,t}+1}} > \sum_{b\in A_1} \left(\sqrt{\gamma}\Delta + w_t\sqrt{\frac{\gamma \ln t}{n_{b,t}+1}}\right) \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
\overset{\text{(b)}}{=}\ &\Pr\left(\exists A \in \Theta \setminus A_1 : \sum_{a\in A} w_t\sqrt{\frac{\gamma \ln t}{Q_A(t)+1}} > \sum_{b\in A_1} \left(\sqrt{\gamma}\Delta + w_t\sqrt{\frac{\gamma \ln t}{Q_{A_1}(t)+1}}\right) \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
=\ &\Pr\left(\exists A \in \Theta \setminus A_1 : w_t\sqrt{\frac{\ln t}{Q_A(t)+1}} > \Delta + w_t\sqrt{\frac{\ln t}{Q_{A_1}(t)+1}} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
\geq\ &\Pr\left(\exists A \in \Theta \setminus A_1 : w_t\left(1 - \sqrt{\frac{Q_A(t)+1}{Q_{A_1}(t)+1}}\right) > \Delta\sqrt{Q_A(t)+1} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
\overset{\text{(c)}}{\geq}\ &\Pr\left(\exists A \in \Theta \setminus A_1 : w_t \cdot \frac{1}{2} > \Delta\sqrt{Q_A(t)+1} \mid \mathcal{F}_{t-1} = F_{t-1}\right)\\
=\ &1 - \underbrace{\Pr\left(w_t \leq 2\Delta\sqrt{Q_A(t)+1}, \forall A \in \Theta \setminus A_1 \mid \mathcal{F}_{t-1} = F_{t-1}\right)}_{\lambda}\quad,
\end{aligned} \tag{55}$$

where step (a) uses the fact that for any base arm in a sub-optimal super arm, the empirical mean is 0, whereas for any base arm in the optimal super arm, the empirical mean is $\sqrt{\gamma}\Delta$ based on our reward distribution construction. Step (b) uses the fact that all base arms in a super arm have the same number of observations. Step (c) uses the lower bound constructed in (54), i.e., $1 - \sqrt{\frac{Q_A(t)+1}{Q_{A_1}(t)+1}} \geq \frac{1}{2}$. Note that for λ, the only randomness is $w \sim \mathcal{N}(0, 1)$ as all $Q_A(t)$ are determined by the history.

To construct an upper bound for λ above, we construct an optimization problem first using the constraint shown in (50). Recall (50) is $\sum_{A\in\Theta\setminus A_1} Q_A(t) \leq cT$. We construct the optimization problem with the objective function shown in the following (56)

$$\max_{x_1,x_2,\ldots,x_{K-1}} \Pr_{w\sim\mathcal{N}(0,1)}\left(w \leq 2\Delta\sqrt{x_a+1}, \forall a \in [K-1]\right) \quad , \tag{56}$$

and constraints shown in (57)

$$x_a \geq 0, \forall a \in [K-1] \quad \text{and} \quad \textstyle\sum_{a=1}^{K-1} x_a \leq c \cdot T \quad . \tag{57}$$

Note that the optimal solution to (56) is the same as the optimal solution to the following objective function (58):

$$\max_{x_1,x_2,\ldots,x_{K-1}} \Pr_{w\sim\mathcal{N}(0,1)}\left(w \leq \min_{a\in[K-1]} x_a\right) \quad . \tag{58}$$

It is not hard to verify that the objective function shown in (58) is maximized when $x_a = \frac{cT}{K-1} = \frac{c\sqrt{KT}}{(K-1)\Delta}$ for all $a \in [K-1]$. Therefore, $x_a = \frac{cT}{K-1} = \frac{c\sqrt{KT}}{(K-1)\Delta}$ for all $a \in [K-1]$ is also the optimal solution to (56) and the maximum value of the objective function shown in (56) is $\Pr_w\left(w \leq 2\Delta\sqrt{\frac{c\sqrt{KT}}{(K-1)\Delta}+1}\right)$.

Now, we are ready to construct an upper bound for λ and have

$$
\begin{aligned}
\lambda &= \Pr_w\left(w \le 2\Delta\sqrt{Q_A(t)+1}, \forall A \in \Theta \setminus A_1 \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
&\le \max_{x_1,x_2,\ldots,x_{K-1}} \Pr_w\left(w \le 2\Delta\sqrt{x_a+1}, \forall a \in [K-1] \mid \mathcal{F}_{t-1} = F_{t-1}\right) \\
&= \max_{x_1,x_2,\ldots,x_{K-1}} \Pr_w\left(w \le 2\Delta\sqrt{x_a+1}, \forall a \in [K-1]\right) \\
&\overset{\text{(a)}}{\le} \Pr_w\left(w \le 2\Delta\sqrt{\frac{c\sqrt{KT}}{(K-1)\Delta}+1}\right) \\
&\overset{\text{(a)}}{\le} 1 - \frac{1}{8\sqrt{\pi}} \cdot e^{-\frac{7}{2}\cdot 4\Delta^2\cdot\left(\frac{c\sqrt{KT}}{(K-1)\Delta}+1\right)} \\
&\overset{\text{(c)}}{\le} 1 - \frac{1}{8\sqrt{\pi}} \cdot e^{-\frac{7}{2}\cdot 4\Delta^2\cdot\frac{2c\sqrt{KT}}{0.5K\Delta}} \\
&= 1 - \frac{1}{8\sqrt{\pi}} \cdot e^{-\frac{7}{2}\cdot 8\sqrt{\frac{K}{T}}\cdot\frac{1}{6}\cdot\frac{\sqrt{KT}}{0.5K}} \\
&= 1 - \frac{1}{8\sqrt{\pi}} e^{-\frac{28}{3}} \\
&= 1 - p_0 \quad ,
\end{aligned}
\tag{59}
$$

where step (a) uses the fact that $\Pr_w\left(w \le 2\Delta\sqrt{\frac{c\sqrt{KT}}{(K-1)\Delta}+1}\right)$ is the maximum value of the objective function shown in (56). Step (b) uses the one-sided anti-concentration inequality shown in (6). Step (c) uses the fact that $K-1 > 0.5K$, $\Delta = \sqrt{K/T}$, and when T is large enough, we have $\frac{c\sqrt{KT}}{(K-1)\Delta} \ge 1$.

By plugging the upper bound for λ into (55), we have $\Pr\left(\exists A \in \Theta \setminus A_1 : A_t = A \mid \mathcal{F}_{t-1} = F_{t-1}\right) \ge p_0$, which concludes the proof for the statement that with at least a constant probability p_0, a sub-optimal super arm is played in round t.

To complete the proof, we use the fact that the total regret from round $t = \alpha T$ to round T is at least $(1-\alpha)T \cdot p_0 \cdot m \cdot \sqrt{\gamma}\Delta = \Omega(Tm\Delta) = \Omega(Tm\sqrt{K/T}) = \Omega(Tm\sqrt{N/(mT)}) = \Omega(\sqrt{mNT})$, which is also a regret lower bound for the total regret from round 1 to round T. $\square$