Multi-Armed Bandits With Machine Learning-Generated Surrogate Rewards

Wenlong Ji
Department of Statistics
Stanford University
Stanford, CA 94305
jwl2000@stanford.edu

Yihan Pan
Spiegel Research Center
Northwestern University
Evanston, IL 60208
yihanpan2027@u.northwestern.edu

Ruihao Zhu
SC Johnson College of Business
Cornell University
Ithaca, NY, 14853
ruihao.zhu@cornell.edu

Lihua Lei
Graduate School of Business
Stanford University
Stanford, CA 94305
lihualei@stanford.edu

Abstract

Multi-armed bandit (MAB) is a widely adopted framework for sequential decision-making under uncertainty. Traditional bandit algorithms rely solely on online data, which tends to be scarce as it must be gathered during the online phase when the arms are actively pulled. However, in many practical settings, rich auxiliary data, such as covariates of past users, is available prior to deploying any arms. We introduce a new setting for MAB where pre-trained machine learning (ML) models are applied to convert side information and historical data into *surrogate rewards*. A prominent feature of this setting is that the surrogate rewards may exhibit substantial bias, as true reward data is typically unavailable in the offline phase, forcing ML predictions to heavily rely on extrapolation. To address the issue, we propose the Machine Learning-Assisted Upper Confidence Bound (MLA-UCB) algorithm, which can be applied to any reward prediction model and any form of auxiliary data. When the predicted and true rewards are jointly Gaussian, it provably improves the cumulative regret, provided that the correlation is non-zero – even in cases where the mean surrogate reward completely misaligns with the true mean rewards. Notably, our method requires no prior knowledge of the covariance matrix between true and surrogate rewards.

1 Introduction

The multi-armed bandit (MAB) framework is a widely adopted framework for sequential and interactive decision-making under incomplete information. In the MAB setting, a decision maker interacts with an unknown environment by sequentially selecting from a pre-specified set of actions (a.k.a. pulling arms). Each selected action yields a random reward from an action-dependent distribution that is unknown to the decision maker. A common objective is to choose actions to maximize cumulative rewards over time or to minimize cumulative regret. It is widely used in real-world applications, such as identifying effective treatments in clinical trials [17, 5] and optimizing recommendations in real time for online platforms [12, 8, 15].

Despite the appealing theoretical guarantees, most existing algorithms, such as the upper confidence bound (UCB) algorithm [4] and the Thompson sampling algorithm [1, 13], operate solely in the online

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML x OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

phase where the decision maker is actively pulling arms, without taking advantage of offline/historical data that is prevalent in practice. Although rewards are not directly observable in the offline phase before an arm is pulled, other variables may be informative about the reward distributions. To extract information, we can leverage machine learning algorithms, including pre-trained AI models like large language models (LLMs), to convert these offline variables into *surrogate rewards*. If ML-generated surrogate rewards are correlated with unrealized true rewards through the input to the ML model, they have the potential to reduce regret even in the early stage by increasing the effective number of pulls.

However, directly using the ML-generated surrogate rewards as a replacement for the true rewards is often unreliable, as they may contain bias and not preserve the ranking of the true mean rewards. To safely exploit the information from surrogate rewards without being potentially hurt by the bias, we propose the Machine Learning-Assisted Upper Confidence Bound (MLA-UCB) algorithm, which combines the online true rewards with ML-generated surrogate rewards for both offline and online units. MLA-UCB integrates the prediction-powered inference (PPI) approach [2, 3] to improve estimation precision of mean rewards for each arm by leveraging the surrogate rewards. Algorithmically, unlike most UCB algorithms, our MLA-UCB algorithm does not require the knowledge of the variance, or the sub-Gaussian parameter, of the true or surrogate rewards.

Under the assumption that the true and surrogate rewards are bivariate Gaussian, with an arbitrary unknown mean and covariance matrix, we derive a theoretical upper bound on the cumulative regret. The bound is never worse than the regret lower bound for MAB problems without surrogate rewards and strictly improves upon it when the correlation between surrogate and true rewards is non-zero for any suboptimal arm. In particular, the surrogate rewards is allowed to have arbitrarily large bias – the regret reduction is achieved through variance reduction instead of direct data aggregation. Even when the correlations are all zero, our regret bound strictly outperforms the best available UCB algorithm [9] at the second order in MAB problems with unknown reward variance and no surrogate rewards.

2 Preliminaries

2.1 Standard MAB setting

In an MAB problem, there are K arms (each corresponds to an action), index by $k = 1, \cdots, K$. We denote by $n_{k,t}$ the total number of pulls for arm k right before any arm is pulled at time t. Pulling arm k at time t yields a random reward $R_{k,n_{k,t}+1}$, which is assumed to be drawn independently from some distribution $\mathcal{P}_{R_k}$ with mean μ_k. When no confusion can arise, we may use the notation R_t to represent the reward observed in time t without specifying the arm being pulled. For each time step, an algorithm specifies a decision $A_t \in [K]$ based on the history $(A_1, R_1, A_2, R_2, \cdots, A_{t-1}, R_{t-1})$. The objective is to maximize the cumulative reward or minimize the cumulative regret over a time horizon T, defined as $\mathrm{Reg}_T = \sum_{t=1}^{T} (\mu^\star - \mu_{A_t})$, where $\mu^\star = \max_k \mu_k$ is the mean reward of the optimal arm. In particular, we assume that the distribution of each arm is fixed ahead, and there is no ties between arms, meaning that $\mu_i \neq \mu_j$ if $i \neq j$. We also define $k^\star = \arg\max_k \mu_k$ as the optimal arm and $\Delta_k = \mu^\star - \mu_k$ as the sub-optimality gap of each arm k.

In this paper, we primarily focus on the Gaussian bandit setting, where the distribution $\mathcal{P}_{R_k}$ is Gaussian, and the mean and variance are unknown to the decision maker. For this setting, [9] proposed the asymptotically optimal algorithm based on the upper confidence bound (UCB) algorithm [11] that achieves the following regret bound for any $T \geq 3K$:

$$\mathbb{E}[\mathrm{Reg_T}] \leq \sum_{k \neq k^\star} \frac{2\Delta_k \log T}{\log\left(1 + \frac{\Delta_k^2}{\sigma_k^2}\right)} + O((\log T)^{3/4} \log\log T). \tag{1}$$

We set this optimal algorithm as the baseline to compare with, and refer as the *classical UCB* algorithm thereafter.

2.2 MAB with surrogate rewards

We formally introduce the setting of MAB with surrogate rewards. In the online phase, if the arm k is pulled for the s-th time, the decision maker can observe a surrogate reward $\hat{R}_{k,s}$ alongside a true reward $R_{k,s}$ as in the standard MAB setting. In the offline phase, the decision maker has access to a static pool of surrogate rewards $\{\hat{R}^{\mathrm{off}}_{k,s}\}_{s=1}^{N_k}$ for arm k. We assume that $\hat{R}^{\mathrm{off}}_{k,1}, \ldots, \hat{R}^{\mathrm{off}}_{k,N_k}, \hat{R}_{k,1}, \hat{R}_{k,2}, \ldots$ are

independent and identically distributed (i.i.d.). As discussed in Section 1, the surrogate rewards can be generated by ML models. For example, if a feature vector $X_{k,s}^{\text{off}}$ and $X_{k,s}$ is available for each unit during both the offline and online phases, a predictive model f_k can be applied to produce surrogate rewards $\hat{R}_{k,s}^{\text{off}} = f_k(X_{k,s}^{\text{off}})$ and $\hat{R}_{k,s} = f_k(X_{k,s})$. Here, f_k may be trained using historical data or be a pre-trained AI model.

For regret analysis, we assume that the true and surrogate rewards are bivariate Gaussian with unknown mean and covariance matrix:

$$\begin{pmatrix} R_{k,s} \\ \hat{R}_{k,s} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_k \\ \tilde{\mu}_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & \rho_k \sigma_k \tilde{\sigma}_k \\ \rho_k \sigma_k \tilde{\sigma}_k & \tilde{\sigma}_k^2 \end{pmatrix} \right), \quad \forall s = 1, 2, \cdots. \tag{2}$$

Here, σ_k^2 and $\tilde{\sigma}_k^2$ are the variances of the true reward and prediction, and ρ_k is the correlation coefficient measuring the quality of the ML model. Importantly, we allow the means μ_k and $\tilde{\mu}_k$ to be arbitrarily different. Moreover, unlike [7], we do not even require access to any partial knowledge of the bias $\tilde{\mu}_k - \mu_k$. Therefore, to effectively leverage the correlation between the true and surrogate rewards, it is crucial to jointly observe both of them in the online phase. This marks a departure from prior work on MAB with offline data [14, 6, 18, 7], where auxiliary information is available only during the offline phase.

3 MLA-UCB: Algorithm and Regret Analysis

3.1 Machine learning-assisted mean estimator

The core building block of the UCB algorithm is the estimation of the mean reward for each arm. With surrogate rewards, we can apply the debiasing technique to reduce the variance of the sample average. We define the machine learning-assisted mean estimator for the mean reward μ_k of arm k as

$$\hat{\mu}_{k,t}^{\text{MLA}} = \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} R_{k,s} - \hat{\lambda}_{k,t} \left(\frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} \hat{R}_{k,s} - \frac{1}{N_k} \sum_{s=1}^{N_k} \hat{R}_{k,s}^{\text{off}} \right). \tag{3}$$

where

$$\hat{\lambda}_{k,t} = \frac{N_k}{n_{k,t} + N_k} \frac{\widehat{\text{Cov}}(R_k, \hat{R}_k)}{\widehat{\text{Var}}(\hat{R}_k)}, \tag{4}$$

and $\widehat{\text{Cov}}(R_k, \hat{R}_k), \widehat{\text{Var}}(\hat{R}_k)$ are the corresponding sample covariance computed on $\{R_{k,s}, \hat{R}_{k,s}\}_{s=1}^{n_{k,t}}$.

The expression of (3) is motivated by the PPI++ method [3]. For any non-random $\hat{\lambda}_{k,t}$, $\hat{\mu}_{k,t}^{\text{MLA}}$ is an unbiased estimate of μ_k as surrogate rewards are i.i.d. across both the offline and online phases. When $\hat{\lambda}_{k,t} = 1$, the estimator can be written as $(\frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} R_{k,s} - \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} \hat{R}_{k,s}) + \frac{1}{N_k} \sum_{s=1}^{N_k} \hat{R}_{k,s}^{\text{off}}$, where the last term is the biased offline estimate of μ_k and the difference of the first two terms can be viewed as a bias estimate. The choice of $\hat{\lambda}_{k,t}$ in (4) is the plug-in estimate of the variance minimizer – it ensures that the $\hat{\mu}_{k,t}^{\text{MLA}}$ is no worse than the sample mean estimator.

3.2 MLA-UCB algorithm and regret analysis

We introduce our MLA-UCB algorithm in Algorithm 1. As with the standard UCB algorithm, the MLA-UCB algorithm pulls the arm with the largest upper confidence bound defined in (16). To kick off the process with an initial variance estimate, we need to pull each arm four times. After the initial pulls, the significance level is set at $\frac{1}{2t\sqrt{\log t}}$ at time t. Note that the algorithm of [16] can be viewed as a special case of MLA-UCB when $N_k = \infty$, except that it uses a smaller significance level $1/t^2$, which may lead to suboptimal performance.

Next, we analyze the regret of the MLA-UCB algorithm. Intuitively, reducing the variance of the mean reward estimate results in a sharper regret bound. The formal result is presented in the following theorem.

Theorem 1. *Under the Gaussian reward model* (2)*, for any $T \geq 4K$, if offline sample size satisfies $N_k \geq \frac{1}{\delta_k} \left(2 \log T / \log \left(1 + \frac{\Delta_k^2}{24\sigma_k^2} \right) + 3 \right), \forall k \in [K]$, then the expected regret of Algorithm 1 can be*

ALGORITHM 1: Machine Learning-Assisted Upper Confidence Bound (MLA-UCB)

Initialization: Pull each arm four times.
for $t = 4K + 1$ to T **do**
Compute the machine learning-assisted mean estimator $\hat{\mu}_{k,t}^{\mathrm{MLA}}$ for each arm.
Compute the variance estimate $Z_{k,t}, \hat{\sigma}_{R,k,t}^2, \hat{\sigma}_{\epsilon,k,t}^2$ for each arm.
Pull the arm

$$A_t = \arg\max_k \left\{ \hat{\mu}_{k,t}^{\mathrm{MLA}} + q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}}\right)\left(\sqrt{\frac{\hat{\sigma}_{R,k,t}^2}{n_{k,t}+N_k}} + \sqrt{\frac{Z_{k,t}\hat{\sigma}_{\epsilon,k,t}^2}{n_{k,t}}}\right)\right\}.$$

end for

bounded by:

$$\mathbb{E}[\mathrm{Reg_T}] \leq \sum_{k\neq k^\star} \frac{2\Delta_k \log T}{\log\left(1 + \frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)} + O\left((\log T)^{2/3}\right). \tag{5}$$

Compared to the regret bound of the classical UCB algorithm in (1), the regret bound in (5) has an extra factor of $(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2$ multiplied by the reward variance σ_k^2 on the first term. In the following corollary, we further show that the factor can be reduced to $(1-\rho_k^2)$ as long as N_k is poly-logarithmic in T.

Corollary 1. *Under the Gaussian data generation model* (2)*, if* $N_k = \Omega((\log T)^{5/3}), \forall k \in [K]$*, then the expected regret of Algorithm 1 can be bounded by:*

$$\mathbb{E}[\mathrm{Reg_T}] \leq \sum_{k\neq k^\star} \frac{2\Delta_k \log T}{\log\left(1 + \frac{\Delta_k^2}{\sigma_k^2(1-\rho_k^2)}\right)} + O\left((\log T)^{2/3}\right). \tag{6}$$

We highlight a few implications of Corollary 1:

1. If $\rho_k = 1, \forall k = 1, \cdots, K$, the surrogate rewards are unbiased, yielding effectively N_k additional reward observations for each arm. As N_k grows, the leading term in the regret vanishes, even though the MLA-UCB algorithm is agnostic to the high quality of the surrogates.

2. If $\rho_k = 0, \forall k = 1, \cdots, K$, the predictions are independent of the true rewards under the Gaussian model (2) and hence provide no information. In this scenario, the regret bound (6) of MLA-UCB matches the leading term of the regret bound (1), which is known to be asymptotically optimal [11].

3. In the general case, if ρ_k are neither all 1s nor all 0s, the predictions are informative but are not perfect to infer the true reward. In this case, the regret of our MLA-UCB algorithm interpolates the two extreme cases, without prior knowledge of the bias or the correlation.

4 Conclusion

In this paper, we introduce the setting of MAB with surrogate rewards and explain how ML and AI models can be applied to convert side information into surrogate rewards to assist online decision-making. Within this framework, we develop the MLA-UCB algorithm that provably outperforms the optimal regret bound achievable without surrogate rewards under Gaussian rewards – even when the surrogate rewards have arbitrarily large bias and the amount of offline data is limited. Furthermore, MLA-UCB does not require knowledge of the variance of and correlation between true and surrogate rewards.

References

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 39.1–39.26. PMLR, 2012.

[2] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[3] Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023.

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

[5] Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

[6] Siddhartha Banerjee, Sean R Sinclair, Milind Tambe, Lily Xu, and Christina Lee Yu. Artificial replay: a meta-algorithm for harnessing historical data in bandits. *arXiv preprint arXiv:2210.00025*, 2022.

[7] Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

[8] Anna Coenen. How the new york times is experimenting with recommendation algorithms. *NYT Open*, 2019. [Last accessed Feb 3, 2025].

[9] Wesley Cowan, Junya Honda, and Michael N Katehakis. Normal bandits of unknown means and variances. *Journal of Machine Learning Research*, 18(154):1–28, 2018.

[10] Bhaskar K Ghosh. Some monotonicity theorems for $\chi 2$, f and t distributions with applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(3):480–492, 1973.

[11] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[12] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[13] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

[14] Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Artificial Intelligence and Statistics*, pages 1046–1054. PMLR, 2012.

[15] Yi Su, Xiangyu Wang, Elaine Ya Le, Liang Liu, Yuening Li, Haokai Lu, Benjamin Lipshitz, Sriraj Badam, Lukasz Heldt, Shuchao Bi, Ed H Chi, Cristos Goodrow, Su-Lin Wu, Lexi Baugher, and Minmin Chen. Long-term value of exploration: Measurements, findings and algorithms. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024.

[16] Arun Verma and Manjesh Kumar Hanawal. Stochastic multi-armed bandits with control variates. *Advances in Neural Information Processing Systems*, 34:27592–27603, 2021.

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

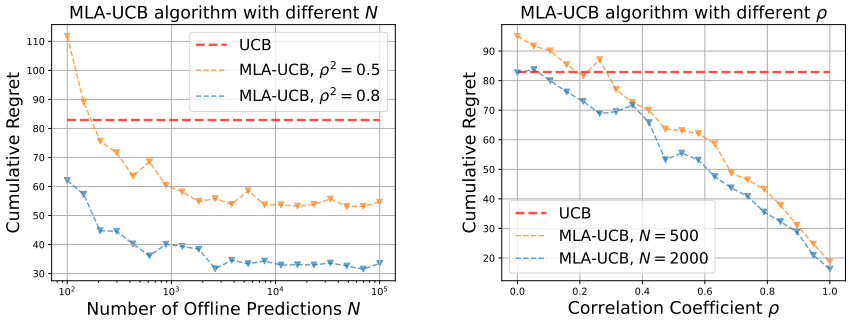
[18] Chicheng Zhang, Alekh Agarwal, Hal Daumé III, John Langford, and Sahand N Negahban. Warm-starting contextual bandits: Robustly combining supervised and bandit feedback. *arXiv preprint arXiv:1901.00301*, 2019.

A Numerical Simulations

In this section, we conduct numerical simulations to evaluate the performance of our MLA-UCB algorithm and compare it with the standard UCB algorithm without surrogate rewards. In Section A.1, we directly simulate surrogate rewards from the Gaussian model (2). In Section A.2, we simulate individual feature vectors and apply actual ML algorithms to produce surrogate rewards. While these surrogate rewards are typically non-Gaussian, meaning that our theory does not strictly apply, we nonetheless observe a sizable reduction in regret.

A.1 Non-ML generated Gaussian surrogate rewards

We simulate a multi-arm bandit model with $K = 5$ arms and time horizon $T = 1000$. The true and surrogate rewards from a bivariate Gaussian distribution (2) with the mean of true rewards $[\Delta, 0, 0, 0, 0]$ for some $\Delta > 0$ and the mean of surrogate rewards $[0, 0.25, 0.5, 0.75, 1]$. For simplicity, we set the correlation $\rho_k = \rho$ and offline sample size $N_k = N$ to be equal across arms. We vary the values of ρ, N, Δ in different ranges and compare the cumulative regret of MLA-UCB to the UCB algorithm [9]. Each experiment is repeated 100 times to report the average cumulative regrets. The experimental results are shown in Figure 1.



(a) Cumulative Regret under different N.

(b) Cumulative Regret under different ρ^2.

(c) Cumulative Regret under different Δ.

(d) Cumulative Regret over time.

Figure 1: **Cumulative regret of the MLA-UCB algorithm (Algorithm 1) and the classical UCB algorithm) under a Gaussian model.** For Figure 1a to 1c, we report the final cumulative regret at $T = 1000$ steps. For Figure 1d, we choose $\rho^2 = 0.5, \Delta = 0.5, N = 100$.

From Figure 1, we observe that the MLA-UCB algorithm can significantly improve upon the classical UCB algorithm under various settings. In particular, Figure 1a demonstrates that as the number of offline predictions N grows, the cumulative regret decreases and finally converges to a limit determined by the correlation ρ_k. For this experiment setting, collecting around $N = 1000$ offline predictions for each arm is enough to reach the optimal regret. Figure 1b illustrates that as ρ grows, the cumulative regret decays approximately linearly in ρ^2. In particular, we observe that as long as $N \geq 2000$, the MLA-UCB algorithm can improve upon the baseline UCB algorithm under any ρ_k^2 above 0.1. Figure 1c illustrates the behavior of regret w.r.t. the gap Δ, it shows that our algorithm can provide significant improvement when Δ is relatively large. In Figure 1d, we demonstrate that

the shape of the cumulative regret curve of the MLA-UCB algorithm is similar to the UCB algorithm, and it is shrunk towards 0 due to the variance reduction. Overall, it confirms that the MLA-UCB algorithm improves cumulative regret in various settings.

A.2 ML-generated non-Gaussian surrogate rewards

Next, we generate surrogate rewards from actual ML models. We consider the following reward generation process:

$$R_k = \sin(w_{k,1}x_1^2 + w_{k,2}x_2^2) + \epsilon,$$

where $\mathbf{x} = (x_1, x_2)^\top$ is the feature vector, $\mathbf{w}_k = (w_{k,1}, w_{k,2})$ is the arm-specific weight parameter, ϵ is the random noise. The features are assumed to be only visible when generating the predictions and we do not use them directly to make decisions. We generate $X \sim \mathcal{N}(0, I_2)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, due to the square and sine operation, the dependency of R_k on the $\mathbf{x}$ are highly nonlinear and hence its distribution is different from a Gaussian distribution.

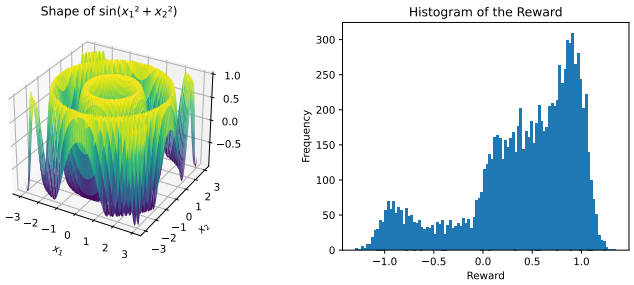


Figure 2: **The conditional expectation and distribution of the true rewards.** The true reward is generated using $w_{k,1} = w_{k,2} = 1$ and $\sigma = 0.1$. The reward function is highly nonlinear and the reward distribution is different from the Gaussian distribution.

We consider using four different ML algorithms in MLA-UCB to predict the reward for each arm based on the features: (1) linear regression, (2) support vector regression (SVR), (3) two-layer neural networks, and (4) decision trees. For each arm, we train an individual model to fit its reward. The correlation of predictions with the true reward of each model is summarized in Table 1, and the experimental results of the MLA-UCB algorithm using these predictions are shown in Figure 3.

Average ρ_k^2	Linear Regression	Support Vector Regression	Neural Nets	Decision Tree
$\sigma = 0$	0.002	0.662	0.675	0.849
$\sigma = 0.2$	0.002	0.571	0.572	0.655
$\sigma = 0.4$	0.002	0.390	0.402	0.314
$\sigma = 0.6$	0.002	0.248	0.276	0.157

Table 1: **Average ρ_k^2 between predictions and true rewards among all arms.**

From Figure 3, we observe that the MLA-UCB algorithm can effectively reduce the cumulative regret for predictions from various ML models, even if the Gaussianity assumption (2)does not hold. The regret reduction is more significant with better prediction models in terms of their ρ_k^2, while the exact relationship is more complicated than the experiments in Section A.1 since the correlation differs between arms. In particular, we found that although linear models fail to make meaningful predictions since the underlying function is highly nonlinear, the algorithm can still achieve comparable performance to the classical UCB algorithm. Overall, these experiments confirm that our algorithm can improve upon the classical UCB algorithm in various settings.

B Proof

In this section, we provide the proof for the regret analysis in Theorem 1 and Corollary 1.

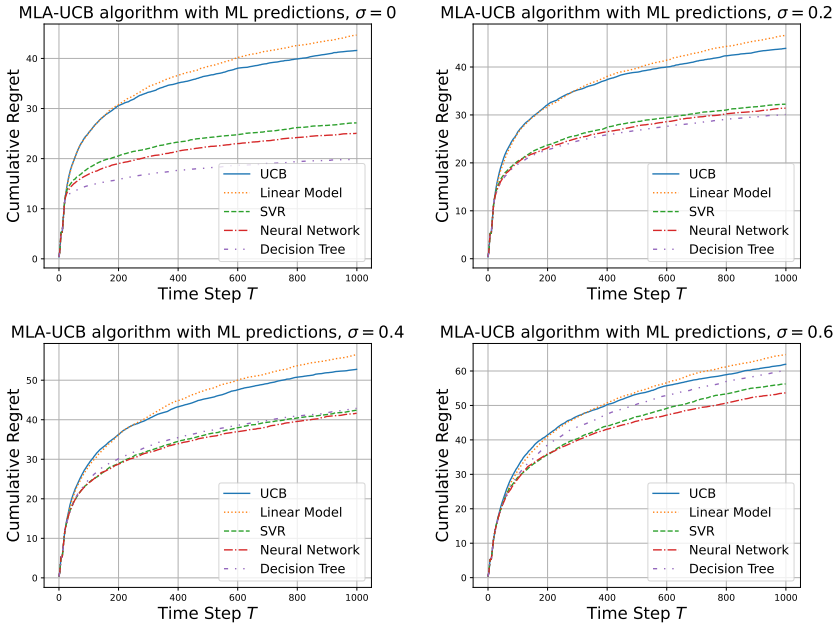


Figure 3: **Cumulative regret of the MLA-UCB algorithm and the classical UCB algorithm under different settings.** The true correlations of predictions of each setting are shown in Table 1. We repeat each experiment 100 times under the same data-generating model and machine learning model to report the average regret.

B.1 Analysis of the machine learning-assisted mean estimator

First, we analyze the property of the machine learning-assisted mean estimator (3). We begin by defining several quantities

$$\mathbb{E}_{k,t}[R] = \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} R_{k,s}, \quad \mathbb{E}_{k,t}[\hat{R}] = \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} \hat{R}_{k,s}, \quad \mathbb{E}_{k}^{\text{off}}[\hat{R}] = \frac{1}{N_k} \sum_{s=1}^{N_k} \hat{R}_{k,s}^{\text{off}},$$

and

$$\widehat{\text{Cov}}(R_k, \hat{R}_k) = \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} (R_{k,s} - \mathbb{E}_{k,t}[R])(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}]), \quad \widehat{\text{Var}}(\hat{R}_k) = \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2$$

Proposition 1. *Denote by the pooled sample mean of the online and offline predictions*

$$\mathbb{E}_{k,t}^{\text{all}}[\hat{R}] = \frac{1}{n_{k,t} + N_k} \left(\sum_{s=1}^{n_{k,t}} \hat{R}_{k,s} + \sum_{s=1}^{N_k} \hat{R}_{k,s}^{\text{off}} \right).$$

Then $\hat{\mu}_{k,t}^{\text{MLA}}$ is given by the intercept of the following ordinary least squares (OLS) estimator, i.e.,

$$\hat{\mu}_{k,t}^{\text{MLA}} = \arg\min_{\mu} \min_{\beta} \sum_{s=1}^{n_{k,t}} \left(R_{k,s} - \mu - \beta \left(\hat{R}_{k,s} - \mathbb{E}_{k,t}^{\text{all}}[\hat{R}] \right) \right)^2. \tag{7}$$

Proof. Recall that the definition of the machine learning-assisted mean estimator is

$$\hat{\mu}_{k,t}^{\text{MLA}} = \mathbb{E}_{k,t}[R] - \lambda_{k,t} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}_{k}^{\text{off}}[\hat{R}] \right),$$

where $\lambda_{k,t} = \frac{N_k}{n_{k,t}+N_k} \frac{\widehat{\mathrm{Cov}}(R_k, \hat{R}_k)}{\widehat{\mathrm{Var}}(\hat{R}_k)}$. By the standard regression result, we know that the solution of the ordinary least squares problem (7) is:

$$\begin{aligned}
\beta_{k,t} =& \frac{\widehat{\mathrm{Cov}}(R_k, \hat{R}_k - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}])}{\widehat{\mathrm{Var}}(\hat{R}_k - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}])} = \frac{\widehat{\mathrm{Cov}}(R_k, \hat{R}_k)}{\widehat{\mathrm{Var}}(\hat{R}_k)} \\
\alpha_{k,t} =& \mathbb{E}_{k,t}[R] - \beta_{k,t}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) \\
=& \mathbb{E}_{k,t}[R] - \frac{\widehat{\mathrm{Cov}}(R_k, \hat{R}_k)}{\widehat{\mathrm{Var}}(\hat{R}_k)} \left(\mathbb{E}_{k,t}[\hat{R}] - \frac{1}{n_{k,t}+N_k} \left(n_{k,t}\mathbb{E}_{k,t}[\hat{R}] + N_k \mathbb{E}^{\mathrm{off}}_{k}[\hat{R}] \right) \right) \\
=& \mathbb{E}_{k,t}[R] - \frac{N_k}{n_{k,t}+N_k} \frac{\widehat{\mathrm{Cov}}(R_k, \hat{R}_k)}{\widehat{\mathrm{Var}}(\hat{R}_k)} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{off}}_{k}[\hat{R}] \right) \\
=& \hat{\mu}^{\mathrm{MLA}}_{k,t},
\end{aligned} \tag{8}$$

which finishes the proof. □

Proposition 1 builds the equivalence with an OLS estimator, we can then use it to derive the following characterization of the distribution of $\hat{\mu}^{\mathrm{MLA}}_{k,t}$.

Proposition 2. *Define $\mathcal{F}_{k,t} = \sigma\left(\{\hat{R}^{\mathrm{off}}_{k,s}\}_{s=1}^{N_k}, \{\hat{R}_{k,s}\}_{s=1}^{n_{k,t}} \right)$ as the σ-field generated by all surrogate rewards for arm k up to time t. Then $\hat{\mu}^{\mathrm{MLA}}_{k,t}$ can be decomposed as*

$$\hat{\mu}^{\mathrm{MLA}}_{k,t} = \mu_k + S_1 + S_2, \tag{9}$$

where S_1 and S_2 are independent random variables, with $S_1 \in \mathcal{F}_{k,t}$,

$$S_1 \sim \mathcal{N}\left(0, \frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2\right), \quad S_2 \big| \mathcal{F}_{k,t} \sim \mathcal{N}\left(0, \frac{1}{n_{k,t}} Z_{k,t}(1-\rho_k^2)\sigma_k^2\right), \tag{10}$$

and $Z_{k,t} \in \mathcal{F}_{k,t}$ is defined as

$$Z_{k,t} = 1 + \frac{n_{k,t}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}])^2}{\sum_{s=1}^{n_{k,t}}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2}.$$

Proof. From (8), we know that

$$\hat{\mu}^{\mathrm{MLA}}_{k,t} = \mathbb{E}_{k,t}[R] - \beta_{k,t}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]). \tag{11}$$

From the data generation distribution 2, using the conditional distribution of multivariate Gaussian variables, we can decompose the reward as

$$R_{k,s} = \mu_k + \rho_k \frac{\sigma_k}{\tilde{\sigma}_k}(\mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}] - \tilde{\mu}_k) + \rho_k \frac{\sigma_k}{\tilde{\sigma}_k}(\hat{R}_{k,s} - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) + \epsilon_{k,s}, \tag{12}$$

where $\epsilon_{k,s} \sim \mathcal{N}(0, (1-\rho_k^2)\sigma_k^2)$ is independent of $\hat{R}_{k,s} - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]$. Then it gives us

$$\begin{aligned}
&\hat{\mu}^{\mathrm{MLA}}_{k,t} - \mu_k \\
=& \rho_k \frac{\sigma_k}{\tilde{\sigma}_k}(\mathbb{E}_{k,t}[\hat{R}] - \tilde{\mu}_k) + \frac{1}{n_{k,t}} \sum_{s=1}^{n_{k,t}} \epsilon_{k,s} - \left(\frac{\sum_{s=1}^{n_{k,t}} \epsilon_{k,s}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])}{\sum_{s=1}^{n_{k,t}}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} + \rho_k \frac{\sigma_k}{\tilde{\sigma}_k} \right) \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}] \right) \\
=& \rho_k \frac{\sigma_k}{\tilde{\sigma}_k}(\mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}] - \tilde{\mu}_k) + \sum_{s=1}^{n_{k,t}} \left(\frac{1}{n_{k,t}} - \frac{(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])}{\sum_{s=1}^{n_{k,t}}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}] \right) \right) \epsilon_{k,s}.
\end{aligned} \tag{13}$$

Notice that $\mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]$ is the complete sufficient statistics for the prediction mean $\tilde{\mu}_k$ with data $\{\hat{R}_{k,s}\}_{s=1}^{n_{k,t}} \cup \{\hat{R}^{\mathrm{off}}_{k,s}\}_{s=1}^{N_k}$, and $\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}]$, $\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]$ are ancillary statistics for the

prediction mean $\tilde{\mu}_k$. Therefore, by Basu's theorem, $\mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$ is independent of $\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}]$ and $\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$. Therefore, define

$$\begin{aligned} S_1 =& \rho_k \frac{\sigma_k}{\tilde{\sigma}_k}(\mathbb{E}^{\text{all}}_{k,t}[\hat{R}] - \tilde{\mu}_k) \\ S_2 =& \sum_{s=1}^{n_{k,t}} \left(\frac{1}{n_{k,t}} - \frac{(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])}{\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}] \right) \right) \epsilon_{k,s}, \end{aligned} \tag{14}$$

then we have S_1 and S_2 are independent. Then it is straightforward to verify that $S_1 \in \mathcal{F}_{k,t}$ and $S_1 \sim \mathcal{N}\left(0, \frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2\right)$. For S_2, conditional on $\mathcal{F}_{k,t}$, it is a linear combination of $\epsilon_{k,s}$, hence it is a Gaussian random variable with mean 0, and the variance is

$$\begin{aligned} \text{Var}(S_2|\mathcal{F}_{k,t}) =& (1-\rho_k)^2\sigma_k^2 \sum_{s=1}^{n_{k,t}} \left(\frac{1}{n_{k,t}} - \frac{(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])}{\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}] \right) \right)^2 \\ =& (1-\rho_k)^2\sigma_k^2 \sum_{s=1}^{n_{k,t}} \left(\left(\frac{1}{n_{k,t}} \right)^2 + \left(\frac{(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])}{\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \left(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}] \right) \right)^2 \right) \\ =& (1-\rho_k)^2\sigma_k^2 \left(\frac{1}{n_{k,t}} + \frac{(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}])^2}{\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \right) = \frac{1}{n_{k,t}} Z_{k,t}(1-\rho_k^2)\sigma_k^2, \end{aligned} \tag{15}$$

which finishes the proof. □

Proposition 2 demonstrates that the estimation error of $\hat{\mu}^{\text{MLA}}_{k,t}$ can be decomposed into two components: the first component S_1 represents the bias of using the empirical mean of surrogate rewards instead of the true mean in the regressor, and the second component represents the uncertainty of intercept estimation in the linear regression model. As a corollary, we can obtain the conditional distribution of $\hat{\mu}^{\text{MLA}}_{k,t}$ on $Z_{k,t}$, which will be useful in the proof of Theorem 1 later.

Corollary 2. *Define $\mathcal{S}_{k,t}$ as the sigma field generated by $\{\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}]\}_{s=1}^{n_{k,t}}$ and $\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$, then $Z_{k,t} \in \mathcal{S}_{k,t} \subset \mathcal{F}_{k,t}$, and*

$$\hat{\mu}^{\text{MLA}}_{k,t} - \mu | \mathcal{S}_{k,t} \sim \mathcal{N}\left(0, \frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2 + \frac{1}{n_{k,t}} Z_{k,t}(1-\rho_k^2)\sigma_k^2\right)$$

Proof. By definition, $Z_{k,t} \in \mathcal{S}_{k,t} \subset \mathcal{F}_{k,t}$ holds. And by Basu's theorem, $\mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$ is independent of $\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}]$ and $\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$. Hence, S_1 is independent of $\mathcal{S}_{k,t}$. Moreover, conditional on $\mathcal{S}_{k,t}$, S_2 is still a linear combination of $\epsilon_{k,s}$, it is independent of S_1 and the variance is the same as in (15). Hence, we know that S_1 and S_2 are two independent Gaussian random variable conditional on $\mathcal{S}_{k,t}$, and the variance is $\frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2 + \frac{1}{n_{k,t}} Z_{k,t}(1-\rho_k^2)$, which finishes the proof. □

Despite the clean distributional characterization, it does not directly imply an upper confidence bound on μ_k since the variance σ_k^2 and the correlation ρ_k are unknown. While both can be estimated consistently via plug-in methods, as previously noted, standard consistency alone is not enough. Following the logic of Student's t-tests, if we can construct an unbiased estimator $\hat{\sigma}_1^2$ of $\rho_k^2\sigma_k^2$ that is independent of S_1 and $k_1\hat{\sigma}_1^2$ is χ^2-distributed with k_1 degrees of freedom, and an unbiased estimator $\hat{\sigma}_2^2$ of $(1-\rho_k^2)\sigma_k^2$ that is independent of S_2 and $k_2\hat{\sigma}_2^2$ is χ^2-distributed with k_2 degrees of freedom, then

$$\mathbb{P}\left(-\frac{S_1}{\hat{\sigma}_1} \le q_{k_1}(\delta) \sqrt{\frac{1}{n_{k,t}+N_k}} \right) \ge 1-\delta, \text{ and } \mathbb{P}\left(-\frac{S_2}{\hat{\sigma}_2} \le q_{k_2}(\delta) \sqrt{\frac{Z_{k,t}}{n_{k,t}}} \right) \ge 1-\delta.$$

This can yield an upper confidence bound on μ_k by (10) and a union bound. We prove that the empirical variance of true rewards can be used as a conservative version of $\hat{\sigma}_1$, since $\text{Var}(R_k) = \sigma_k^2 \ge \rho_k^2\sigma_k^2$, and the residual mean square error of the regression (7) can be used as $\hat{\sigma}_2^2$, since $\text{Var}(R_k|\hat{R}_k) = (1-\rho_k^2)\sigma_k^2$. We are unable to find an unbiased estimator of $\rho_k^2\sigma_k^2$ that is independent of S_1, though we show it has a negligible effect when $N_k \gg n_{k,t}$ for suboptimal arms.

Proposition 3. *Let*

$$\hat{\sigma}^2_{R,k,t} = \frac{1}{n_{k,t}-1}\sum_{s=1}^{n_{k,t}} \left(R_{k,s} - \mathbb{E}_{k,t}[R]\right)^2,$$
$$\hat{\sigma}^2_{\epsilon,k,t} = \frac{1}{n_{k,t}-2}\sum_{s=1}^{n_{k,t}} \left(R_{k,s} - \hat{\mu}^{\mathrm{MLA}}_{k,t} - \beta_{k,t}\left(\hat{R}_{k,s} - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]\right)\right)^2.$$

Then, for any $\delta \in (\frac{1}{2}, 1)$ we have

$$\mathbb{P}\left(\mu_k \le \hat{\mu}^{\mathrm{MLA}}_{k,t} + q_{n_{k,t}-2}(\delta)\left(\sqrt{\frac{\hat{\sigma}^2_{R,k,t}}{n_{k,t}+N_k}} + \sqrt{\frac{Z_{k,t}\hat{\sigma}^2_{\epsilon,k,t}}{n_{k,t}}}\right)\right) \ge 1-2\delta, \tag{16}$$

where $q_d(\delta)$ is the $1-\delta$ quantile of the Stundet's t-distribution of d degrees of freedom.

Proof. To prove the concentration bound (16), we handle the two quantities S_1 and S_2 separately. First, for S_1, we have proven that $S_1 \sim \mathcal{N}(0, \frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2)$, and by definition,

$$\sigma^2_{R,k,t} = \frac{1}{n_{k,t}-1}\sum_{s=1}^{n_{k,t}} \left(R_{k,s} - \mathbb{E}_{k,t}[R]\right)^2 \sim \sigma_k^2(n_{k,t}-1)\chi^2_{n_{k,t}-1}. \tag{17}$$

Again, by Basu's theorem, S_1 is independent of $\sigma^2_{R,k,t}$, hence

$$\frac{S_1}{\sqrt{\sigma^2_{R,k,t}}} \sim \sqrt{\frac{1}{n_{k,t}+N_k}}\rho_k t_{n_{k,t}-1}.$$

Therefore, we can use the quantile of Student's t-distribution and obtain that

$$\mathbb{P}\left(S_1 \le -q_{n_{k,t}-1}(\delta)\sqrt{\frac{\hat{\sigma}^2_{R,k,t}}{n_{k,t}+N_k}}|\rho_k|\right) \le \delta. \tag{18}$$

Next, for S_2, from (15) we know that

$$S_2|\mathcal{F}_{k,t} \sim \mathcal{N}\left(0, \frac{1}{n_{k,t}}Z_{k,t}(1-\rho_k^2)\sigma_k^2\right).$$

Moreover, using Proposition 1, we know that $\hat{\sigma}^2_{\epsilon,k,t}$ is the average sum of residuals for the ordinary least squares problem (7). Therefore, using the linear representation (12), the classical results in OLS suggest that

$$\hat{\sigma}^2_{\epsilon,k,t}|\mathcal{F}_{k,t} \sim (1-\rho_k^2)\sigma_k^2(n_{k,t}-2)\chi^2_{n_{k,t}-2}, \tag{19}$$

and $\hat{\sigma}^2_{\epsilon,k,t}$ is independent of $\alpha_{k,t}, \beta_{k,t}$ conditional on $\mathcal{F}_{k,t}$. Notice that $S_2 = \hat{\mu}^{\mathrm{MLA}}_{k,t} - \mu_k - S_1$, $S_1 \in \mathcal{F}_{k,t}$, and $\hat{\mu}^{\mathrm{MLA}}_{k,t} = \alpha_{k,t}$ by Proposition 1, thus $\hat{\sigma}^2_{\epsilon,k,t}$ and S_2 are independent conditional on $\mathcal{F}_{k,t}$. Therefore, we know that

$$\frac{S_2}{\sqrt{\hat{\sigma}^2_{\epsilon,k,t}}}\Big|\mathcal{F}_{k,t} \sim \sqrt{\frac{Z_{k,t}}{n_{k,t}}}t_{n_{k,t}-2},$$

Again, we can use the quantile of Student's t-distribution and obtain that

$$\mathbb{P}\left(S_2 \le -q_{n_{k,t}-2}(\delta)\sqrt{\frac{Z_{k,t}\hat{\sigma}^2_{\epsilon,k,t}}{n_{k,t}}}\Big|\mathcal{F}_{k,t}\right) \le \delta. \tag{20}$$

In particular, it is well known that for any fixed $c > 0$, the tail probability $\mathbb{P}(t_d > c)$ is decreasing with respect to d (a proof can be found in Corollary 4.3 in [10]). Hence, we have $q_{n_{k,t}-2}(\delta) \ge q_{n_{k,t}-1}(\delta)$

for any $\delta \in (\frac{1}{2}, 1)$. Combining (18) and (20) together, we can obtain that

$$\begin{aligned}
&\mathbb{P}\left(\mu_k > \hat{\mu}_{k,t}^{\mathrm{MLA}} + q_{n_{k,t}-2}(\delta)\left(\sqrt{\frac{\hat{\sigma}_{R,k,t}^2}{n_{k,t}+N_k}} + \sqrt{\frac{Z_{k,t}\hat{\sigma}_{\epsilon,k,t}^2}{n_{k,t}}}\right)\right) \\
=&\mathbb{P}\left(S_1 + S_2 < -q_{n_{k,t}-2}(\delta)\left(\sqrt{\frac{\hat{\sigma}_{R,k,t}^2}{n_{k,t}+N_k}} + \sqrt{\frac{Z_{k,t}\hat{\sigma}_{\epsilon,k,t}^2}{n_{k,t}}}\right)\right) \\
\leq&\mathbb{P}\left(S_1 < -q_{n_{k,t}-2}(\delta)\sqrt{\frac{\hat{\sigma}_{R,k,t}^2}{n_{k,t}+N_k}}\right) + \mathbb{P}\left(S_2 < -q_{n_{k,t}-2}(\delta)\sqrt{\frac{Z_{k,t}\hat{\sigma}_{\epsilon,k,t}^2}{n_{k,t}}}\right) \\
\leq&\mathbb{P}\left(S_1 < -q_{n_{k,t}-1}(\delta)\sqrt{\frac{\hat{\sigma}_{R,k,t}^2}{n_{k,t}+N_k}}|\rho_k|\right) + \mathbb{E}\left[\mathbb{P}\left(S_2 < -q_{n_{k,t}-2}(\delta)\sqrt{\frac{Z_{k,t}\hat{\sigma}_{\epsilon,k,t}^2}{n_{k,t}}}\middle|\mathcal{F}_{k,t}\right)\right] \leq 2\delta,
\end{aligned}$$

which finishes the proof. □

By the law of large numbers, the empirical variance should converge to the true variance, i.e., as $n_{k,t} \to \infty$ we have $\hat{\sigma}_{R,k,t}^2 \to \sigma_k^2, \hat{\sigma}_{\epsilon,k,t}^2 \to (1-\rho_k^2)\sigma_k^2$, and similarly $Z_{k,t} \to 1$. Therefore, as long as $N_k \gg n_{k,t}$, the confidence bound in (16) will be approximately $\hat{\mu}_{k,t}^{\mathrm{MLA}} + q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}}\right)\sqrt{\frac{(1-\rho_k^2)\sigma_k^2}{n_{k,t}}}$ when $n_{k,t}$ is large. This demonstrates that the surrogate rewards effectively reduce the variance by ρ_k^2.

Compared to the upper confidence bound constructed in [9] for standard normal bandits, we use the exact quantile of t-distribution $q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}}\right)$ to scale the standard deviation, while they use a scaling parameter of $\sqrt{n_{k,t}(t^{\frac{2}{n_{k,t}-2}}-1)}$. As we shall see in the following proposition, their scaling parameter can be viewed as an upper bound for the exact quantile.

Proposition 4. *For any real number $s > 0$ and integer $d \geq 2$ we have*

$$q_d\left(\frac{1}{2s\sqrt{\log s}}\right) \leq \sqrt{d(s^{\frac{2}{d-1}}-1)}$$

Proof of Proposition 4. It suffices to prove that for a random variable $T_d \sim t_d$, the following inequality holds:

$$\mathbb{P}\left(T_d \geq \sqrt{d(s^{\frac{2}{d-1}}-1)}\right) \leq \frac{1}{2s\sqrt{\log s}}. \tag{21}$$

By the definition of the t-distribution, let $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi_d^2$ be two independent random variables, then $\frac{X}{\sqrt{Y/d}} \sim t_d$, and then by the Mill's inequality $\mathbb{P}(X > t) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}\exp(-\frac{t^2}{2})$,

$$\mathbb{P}\left(T_d \geq \sqrt{d(s^{\frac{2}{d-1}}-1)}\right) = \mathbb{P}\left(X \geq \sqrt{Y(s^{\frac{2}{d-1}}-1)}\right) \leq \frac{1}{\sqrt{2\pi(s^{\frac{2}{d-1}}-1)}}\mathbb{E}\left[Y^{-\frac{1}{2}}e^{-\frac{1}{2}Y(s^{\frac{2}{d-1}}-1)}\right], \tag{22}$$

for the expectation on the right hand side, notice that for any $k > 0$

$$\begin{aligned}
\mathbb{E}[Y^{-1/2}e^{-\frac{1}{2}kY}] &= \int_0^\infty \frac{1}{2^{d/2}\Gamma(d/2)}x^{(d-1)/2-1}e^{-(1+k)x/2}dx \\
&= \frac{1}{(1+k)^{\frac{d-1}{2}}}\int_0^\infty \frac{1}{2^{d/2}\Gamma(d/2)}x^{(d-1)/2-1}e^{-x/2}dx \\
&= \frac{1}{\sqrt{2}(1+k)^{\frac{d-1}{2}}}\frac{\Gamma((d-1)/2)}{\Gamma(d/2)},
\end{aligned} \tag{23}$$

take $k = s^{\frac{2}{d-1}} - 1$, we have $\mathbb{E}\left[Y^{-\frac{1}{2}} e^{-\frac{1}{2}Y(s^{\frac{2}{d-1}}-1)}\right] = \frac{1}{\sqrt{2}s}\frac{\Gamma((d-1)/2)}{\Gamma(d/2)}$. Moreover, it is easy to prove by induction that

$$\frac{\Gamma((d-1)/2)}{\Gamma(d/2)} \leq \sqrt{\frac{2\pi}{d}}, \quad \forall d \geq 2, d \in \mathbb{N}^+,$$

therefore we can derive from (22) that

$$\mathbb{P}\left(T_d \geq \sqrt{d(s^{\frac{2}{d-1}} - 1)}\right) \leq \frac{1}{\sqrt{2d}s\sqrt{s^{\frac{2}{d-1}} - 1}} \leq \frac{1}{\sqrt{2d}s\sqrt{\frac{2\log s}{d-1}}} \leq \frac{1}{2s\sqrt{\log s}}$$

□

B.2 Regret bound

After establishing the confidence bound for the machine learning-assisted mean estimator, now we are prepared to provide the regret analysis. We will use the following technical lemma on the tail bound of χ^2 distribution from Lemma 1.

Lemma 1 (Proposition 8 from [9]). *If a random variable $X \sim \chi^2_d$, then for any $\delta > 0$, we have*

$$\mathbb{P}(X > d(1+\delta)) \leq \left(e^{-\delta}(1+\delta)\right)^{k/2}$$

Using Lemma 1, we can prove the tail bound for $\hat{\sigma}^2_{R,k,t}$ and $\hat{\sigma}^2_{\epsilon,k,t}$.

Lemma 2. *For any $\delta > 0$, the following inequalities holds:*

$$\mathbb{P}(\sigma^2_{R,k,t} > (1+\delta)\sigma^2_k) \leq \left(e^{-\delta}(1+\delta)\right)^{\frac{n_{k,t}-1}{2}},$$
$$\mathbb{P}(\sigma^2_{\epsilon,k,t} > (1+\delta)(1-\rho^2_k)\sigma^2_k) \leq \left(e^{-\delta}(1+\delta)\right)^{\frac{n_{k,t}-2}{2}}.$$

Proof. Since we have proven that $\sigma^2_{R,k,t} \sim \sigma^2_k \chi^2_{n_{k,t}-1}$ and $\hat{\sigma}^2_{\epsilon,k,t}|\mathcal{F}_{k,t} \sim (1-\rho^2_k)\sigma^2_k \chi^2_{n_{k,t}-2}$ in (17) and (19), directly apply Lemma 1 on $\hat{\sigma}^2_{R,k,t}$ and $\hat{\sigma}^2_{\epsilon,k,t}$ will finish the proof. □

Similarly, we can prove the tail bound for $Z_{k,t}$.

Lemma 3. *For any $\delta > 0$, if $n_{k,t} \geq 6$, the following inequality holds:*

$$\mathbb{P}(Z_{k,t} \geq 1 + \delta) \leq \frac{3}{\delta^2}\frac{1}{(n_{k,t}-3)(n_{k,t}-5)}$$

Proof. Recall that by definition (2),

$$Z_{k,t} = 1 + \frac{n_{k,t}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}])^2}{\sum_{s=1}^{n_{k,t}}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2}.$$

Notice that

$$\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}] = \left(\frac{1}{n_{k,t}} - \frac{1}{n_{k,t}+N_k}\right)\sum_{s=1}^{n_{k,t}}\hat{R}_{k,s} - \frac{1}{n_{k,t}+N_k}\sum_{s=1}^{N_k}\hat{R}^{\text{off}}_{k,s},$$

thus $\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}]$ is a Gaussian random variable with mean 0 and variance

$$\text{Var}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\text{all}}_{k,t}[\hat{R}]) = \left(\frac{N_k^2}{n_{k,t}(n_{k,t}+N_k)^2} + \frac{N_k}{(n_{k,t}+N_k)^2}\right)\sigma^2_k = \frac{N_k\sigma^2_k}{n_{k,t}(n_{k,t}+N_k)}.$$

On the other hand, we know that

$$\sum_{s=1}^{n_{k,t}}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2 \sim \sigma^2_k \chi^2_{n_{k,t}-1}.$$

Moreover,

$$\begin{aligned}
&\mathrm{Cov}(\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}], \mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) \\
=&\mathrm{Cov}(\hat{R}_{k,s}, \mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) - \mathrm{Cov}(\mathbb{E}_{k,t}[\hat{R}], \mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) \\
=&\mathrm{Cov}(\hat{R}_{k,s}, \mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) - \frac{1}{n_{k,t}} \sum_{j=1}^{n_{k,t}} \mathrm{Cov}(\hat{R}_{k,j}, \mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}]) = 0.
\end{aligned}$$

Therefore, $(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}])^2$ and $\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2$ are independent random variables, and thus

$$Z_{k,t} - 1 = \frac{n_{k,t}(\mathbb{E}_{k,t}[\hat{R}] - \mathbb{E}^{\mathrm{all}}_{k,t}[\hat{R}])^2}{\sum_{s=1}^{n_{k,t}} (\hat{R}_{k,s} - \mathbb{E}_{k,t}[\hat{R}])^2} \sim \frac{N_k}{(n_{k,t} + N_k)(n_{k,t} - 1)} t^2_{n_{k,t}-1}.$$

Using Markov's inequality, we can conclude that

$$\begin{aligned}
&\mathbb{P}(Z_{k,t} \geq 1 + \delta) = \mathbb{P}\left(\frac{N_k}{(n_{k,t} + N_k)(n_{k,t} - 1)} t^2_{n_{k,t}-1} \geq \delta \right) \\
\leq &\frac{1}{\delta^2} \left(\frac{N_k}{(n_{k,t} + N_k)(n_{k,t} - 1)} \right)^2 \mathbb{E}[t^4_{n_{k,t}-1}] = \frac{3}{\delta^2} \left(\frac{N_k}{n_{k,t} + N_k} \right)^2 \frac{1}{(n_{k,t} - 3)(n_{k,t} - 5)} \\
\leq &\frac{3}{\delta^2} \frac{1}{(n_{k,t} - 3)(n_{k,t} - 5)}
\end{aligned}$$

which finishes the proof. □

Now we are ready to prove the regret bound of the MLA-UCB algorithm in Theorem 1. We will start with the following generalized version of regret bound.

Theorem 2. *Under the Gaussian data generation model* (2)*, for any $\epsilon \in (0, 1)$ and $T \geq 4K$, if the sample size of offline predictions satisfies*

$$N_k \geq \frac{1}{\delta_k} \left(\frac{2 \log T}{\log\left(1 + \frac{\Delta_k^2}{24\sigma_k^2} \frac{(1-\epsilon)^2}{1+\epsilon}\right)} + 3 \right), \forall k \in [K] \tag{24}$$

then the expected regret of Algorithm 1 can be bounded by:

$$\mathbb{E}[\mathrm{Reg_T}] \leq \sum_{k \neq k^\star} \left(\frac{2 \log T}{\log\left(1 + \frac{\Delta_k^2}{\sigma_k^2} \frac{(1-\epsilon)^2}{(1+\epsilon)} \frac{1}{(\sqrt{1-\rho_k^2} + \sqrt{\delta_k})^2}\right)} + 2\sqrt{\log T} + \frac{2(1+\epsilon)\sigma_k^2}{\epsilon^2 \Delta_k^2} + \frac{125}{\epsilon^2} + 4 \right) \Delta_k \tag{25}$$

Proof of Theorem 2. For any $\epsilon \in (0, 1)$, define $\tilde{\epsilon}_k = \Delta_k \epsilon$. Define the confidence bound in Algorithm 1 as

$$\hat{B}_{k,t} = q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}} \right) \left(\sqrt{\frac{\hat{\sigma}^2_{R,k,t}}{n_{k,t} + N_k}} + \sqrt{\frac{Z_{k,t} \hat{\sigma}^2_{\epsilon,k,t}}{n_{k,t}}} \right), \tag{26}$$

and define the following events

$$\begin{aligned}
\mathcal{A}_{k,t} =&\{A_t = k\} = \{\hat{\mu}^{\mathrm{MLA}}_{k,t} + \hat{B}_{k,t} \geq \hat{\mu}_{k^\star,t} + \hat{B}_{k^\star,t}\}, \\
\mathcal{B}_{k,t} =&\{\mu_k + \hat{B}_{k,t} + \tilde{\epsilon}_k \geq \mu^\star\}, \\
\mathcal{D}_{k,t} =&\{\hat{\mu}^{\mathrm{MLA}}_{k,t} \leq \mu_k + \tilde{\epsilon}_k\}, \\
\mathcal{E}^1_{k,t} =&\{\sigma^2_{R,k,t} \leq (1+\epsilon)\sigma_k^2\}, \\
\mathcal{E}^2_{k,t} =&\{\sigma^2_{\epsilon,k,t} \leq (1+\epsilon/3)(1-\rho_k^2)\sigma_k^2\}, \\
\mathcal{E}^3_{k,t} =&\{Z_{k,t} \leq (1+\epsilon/3)\}, \\
\mathcal{E}_{k,t} =&\cap_{i=1}^3 \mathcal{E}^i_{k,t}.
\end{aligned} \tag{27}$$

Then we can express the expected regret as

$$\mathbb{E}[\mathrm{Reg_T}] = \mathbb{E}\sum_{k\neq k^\star} n_{k,T}\Delta_k = \sum_{k\neq k^\star}\Delta_k \mathbb{E}\sum_{t=1}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}. \tag{28}$$

Furthermore, we can decompose the expectation as follows

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}\mathbb{I}_{\mathcal{A}_{k,t}} =& \mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}\cap\mathcal{D}_{k,t}\cap\mathcal{E}_{k,t}} + \mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}^C\cap\mathcal{D}_{k,t}\cap\mathcal{E}_{k,t}} \\ &+ \mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{D}_{k,t}^C\cap\mathcal{E}_{k,t}} + \mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{E}_{k,t}^C} + 4,\end{aligned} \tag{29}$$

and we are going to bound these terms separately in the following analysis. For the first term of (29),

$$\begin{aligned}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}\cap\mathcal{D}_{k,t}\cap\mathcal{E}_{k,t}} \overset{(1)}{\leq} \mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}_{k,t}}\mathbb{I}\{\Delta_k(1-\epsilon)\leq\hat{B}_{k,t}\} \\ \overset{(2)}{=}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}_{k,t}}\mathbb{I}\left\{\Delta_k(1-\epsilon)\leq q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}}\right)\left(\sqrt{\frac{\sigma_{\epsilon,k,t}^2 Z_{k,t}}{n_{k,t}}}+\sqrt{\frac{\sigma_{R,k,t}^2}{n_{k,t}+N_k}}\right)\right\} \\ \overset{(3)}{\leq}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}_{k,t}}\mathbb{I}\left\{\Delta_k(1-\epsilon)\leq q_{n_{k,t}-2}\left(\frac{1}{2t\sqrt{\log t}}\right)\sqrt{1+\epsilon}\left(\sqrt{\frac{(1-\rho_k^2)\sigma_k^2}{n_{k,t}}}+\sqrt{\frac{\sigma_k^2}{n_{k,t}+N_k}}\right)\right\} \\ \overset{(4)}{\leq}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}_{k,t}}\mathbb{I}\left\{\Delta_k(1-\epsilon)\leq\sqrt{(1+\epsilon)(t^{\frac{2}{n_{k,t}-3}}-1)}\left(\sqrt{(1-\rho_k^2)\sigma_k^2}+\sqrt{\frac{\sigma_k^2 n_{k,t}}{n_{k,t}+N_k}}\right)\right\} \\ \overset{(5)}{\leq}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}\left\{\Delta_k(1-\epsilon)\leq\sqrt{(1+\epsilon)(t^{\frac{2}{n_{k,t}-3}}-1)}\left(\sqrt{(1-\rho_k^2)\sigma_k^2}+\sqrt{\frac{\sigma_k^2 n_{k,t}}{n_{k,t}+N_k}}\right)\right\}.\end{aligned} \tag{30}$$

Here, the inequality (1) is from the definition of $\mathcal{B}_{k,t}$ and the fact that $\mathbb{I}_{\mathcal{D}_{k,t}}\leq 1$; the equality (2) is from the definition of $\hat{B}_{k,t}$ in (26); the inequality (3) is from the definition of $\mathcal{E}_{k,t}$ and the fact that $(1+\epsilon/3)^2\leq 1+\epsilon, \forall\epsilon\in(0,1)$; the inequality (4) is from Proposition 4; and the inequality (5) uses the fact that $\mathbb{I}_{\mathcal{E}_{k,t}}\leq 1$.

Define the event

$$\mathcal{G}_{k,t} = \left\{\Delta_k(1-\epsilon)\leq\sqrt{(1+\epsilon)(t^{\frac{2}{n_{k,t}-3}}-1)}\left(\sqrt{(1-\rho_k^2)\sigma_k^2}+\sqrt{\frac{\sigma_k^2 n_{k,t}}{n_{k,t}+N_k}}\right)\right\}.$$

Then under the sample size condition

$$N_k \geq \frac{1}{\delta_k}\left(\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{4\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\right)}+3\right),$$

we can control the ratio of online and offline samples on the event $\mathcal{G}_{k,t}$ by

$$\begin{aligned}\mathcal{G}_{k,t} &\subseteq \left\{\Delta_k(1-\epsilon)\leq\sqrt{(1+\epsilon)(t^{\frac{2}{n_{k,t}-3}}-1)}2\sigma_k\right\} \subseteq \left\{\log\left(1+\frac{\Delta_k^2}{4\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\right)\leq\frac{2\log t}{n_{k,t}-3}\right\} \\ &\subseteq \left\{n_{k,t}\leq\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{4\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\right)}+3\right\} \subseteq \left\{\frac{n_{k,t}}{n_{k,t}+N_k}\leq\delta_k\right\}.\end{aligned} \tag{31}$$

Combine it with (30), we have

$$\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}\cap\mathcal{D}_{k,t}\cap\mathcal{E}_{k,t}}\overset{(1)}{\leq}\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{G}_{k,t}}\overset{(2)}{=}\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{G}_{k,t}}\mathbb{I}\left\{\frac{n_{k,t}}{n_{k,t}+N_k}\leq\delta_k\right\}$$
$$\overset{(3)}{\leq}\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}\left\{\frac{n_{k,t}}{n_{k,t}+N_k}\leq\delta_k\right\}\mathbb{I}\left\{\Delta_k(1-\epsilon)\leq\sqrt{(1+\epsilon)(t^{\frac{2}{n_{k,t}-3}}-1)}\left(\sqrt{(1-\rho_k^2)\sigma_k^2}+\sqrt{\sigma_k^2\delta_k}\right)\right\}$$
$$\overset{(4)}{\leq}\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}\left\{n_{k,t}\leq\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+3\right\}$$
$$\overset{(5)}{\leq}\mathbb{E}\sum_{n_{k,t}=4}^{\infty}\mathbb{I}\left\{n_{k,t}\leq\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+3\right\}=\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+3 \tag{32}$$

Here, the inequality (1) is by (30) and the definition of $\mathcal{G}_{k,t}$; equality (2) is by (30); inequality (3) and (4) are straightforward algebra; inequality (5) uses the fact that $\mathcal{A}_{k,t}=\{n_{k,t+1}=n_{k,t}+1\}$ and $\mathbb{I}\left\{n_{k,t}\leq\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{1+\epsilon}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+3\right\}$ depends on t only through $n_{k,t}$, thus we can transform the sum over t into the sum over $n_{k,t}$.

For the second term of (29),

$$\begin{aligned}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}^C\cap\mathcal{D}_{k,t}\cap\mathcal{E}_{k,t}}\leq\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{B}_{k,t}^C\cap\mathcal{D}_{k,t}}\\&\leq\sum_{t=4K}^{T}\mathbb{P}\left(\hat{\mu}_{k,t}^{\mathrm{MLA}}+\hat{B}_{k,t}\geq\hat{\mu}_{k^\star,t}+\hat{B}_{k^\star,t},\quad\mu_k+\hat{B}_{k,t}+\tilde{\epsilon}_k<\mu^\star,\quad\hat{\mu}_{k,t}^{\mathrm{MLA}}\leq\mu_k+\tilde{\epsilon}_k\right)\\&\leq\sum_{t=4K}^{T}\mathbb{P}\left(\mu^\star>\hat{\mu}_{k^\star}+\hat{B}_{k^\star,t}\right)\overset{(1)}{\leq}\sum_{t=4K}^{T}\frac{1}{t\sqrt{\log t}}\leq\int_{t=3}^{T}\frac{dt}{t\sqrt{\log t}}\leq2\sqrt{\log T}-2.\end{aligned}\tag{33}$$

Here, the inequality (1) is from the definition of confidence band $\hat{B}_{k,t}$ in (26) and Proposition 3.

For the third term of (29),

$$\begin{aligned}&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}\cap\mathcal{D}_{k,t}^C\cap\mathcal{E}_{k,t}}\leq\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}_{k,t}^3}\mathbb{I}_{\mathcal{D}_{k,t}^C}\overset{(1)}{\leq}\mathbb{E}\sum_{n_{k,t}=4}^{\infty}\mathbb{I}_{\mathcal{E}_{k,t}^3}\mathbb{I}_{\mathcal{D}_{k,t}^C}\\&\overset{(2)}{\leq}\sum_{n_{k,t}=4}^{\infty}\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_{k,t}^3}\mathbb{P}\left(\hat{\mu}_{k,t}^{\mathrm{MLA}}>\mu_k+\tilde{\epsilon}_k|\mathcal{S}_{k,t}\right)\right]\\&\overset{(3)}{\leq}\sum_{n_{k,t}=4}^{\infty}\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_{k,t}^3}\exp\left(-\frac{\Delta_k^2\epsilon^2}{2(\frac{1}{n_{k,t}+N_k}\rho_k^2\sigma_k^2+\frac{1}{n_{k,t}}Z_{k,t}(1-\rho_k^2)\sigma_k^2)}\right)\right]\\&\overset{(4)}{\leq}\sum_{n_{k,t}=4}^{\infty}\mathbb{E}\left[\mathbb{I}_{\mathcal{E}_{k,t}^3}\exp\left(-\frac{n_{k,t}\Delta_k^2\epsilon^2}{2(1+\epsilon)\sigma_k^2}\right)\right]\leq\sum_{n_{k,t}=4}^{\infty}\exp\left(-\frac{n_{k,t}\Delta_k^2\epsilon^2}{2(1+\epsilon)\sigma_k^2}\right)\\&\leq\frac{1}{\exp\left(\frac{\Delta_k^2\epsilon^2}{2(1+\epsilon)\sigma_k^2}\right)-1}\leq\frac{2(1+\epsilon)\sigma_k^2}{\epsilon^2\Delta_k^2}<\infty.\end{aligned}\tag{34}$$

Here, inequality (1) is from the fact that $\mathcal{E}_{k,t}=\mathcal{E}_{k,t'},\mathcal{D}_{k,t}=\mathcal{D}_{k,t'}$ for any time step t,t' such that $n_{k,t}=n_{k,t'}$; inequality (2) is from the fact that $Z_{k,t}\in\mathcal{S}_{k,t}$ in Corollary 2; inequality (3) is from

Corollary 2 and Hoeffding's inequality of Gaussian random variable; inequality (4) is from the definition of $\mathcal{E}^3_{k,t}$.

For the fourth term of (29),

$$\begin{aligned}
&\mathbb{E}\sum_{t=4K}^{T}\mathbb{I}_{\mathcal{A}_{k,t}}\mathbb{I}_{\mathcal{E}^C_{k,t}} \overset{(1)}{\leq} \mathbb{E}\sum_{n_{k,t}=4}^{\infty}\mathbb{I}_{(\mathcal{E}^1_{k,t})^C}+\mathbb{I}_{(\mathcal{E}^2_{k,t})^C}+\mathbb{I}_{(\mathcal{E}^3_{k,t})^C}\\
&\overset{(2)}{\leq}\sum_{n_{k,t}=4}^{\infty}\left(\mathbb{P}(\sigma^2_{R,k,t}>(1+\epsilon)\sigma^2_R)+\mathbb{P}(\sigma^2_{\epsilon,k,t}>(1+\epsilon/3)(1-\rho^2_k)\sigma^2_R)+\mathbb{P}(Z_{k,t}>(1+\epsilon/3))\right)\\
&\overset{(3)}{\leq}\sum_{n_{k,t}=4}^{\infty}\left(e^{-\epsilon/3}(1+\epsilon/3)\right)^{\frac{n_{k,t}-2}{2}}+\left(e^{-\epsilon}(1+\epsilon)\right)^{\frac{n_{k,t}-1}{2}}+2+\sum_{n_{k,t}=6}^{\infty}\frac{27}{\epsilon^2}\frac{1}{(n_{k,t}-3)(n_{k,t}-5)}\\
&\overset{(4)}{\leq}\frac{1}{\sqrt{\frac{e^{\epsilon/3}}{1+\epsilon/3}}-1}+\frac{1}{\sqrt{\frac{e^{\epsilon}}{1+\epsilon}}-1}+2+\frac{27}{\epsilon^2}\frac{\pi^2}{6}\overset{(5)}{\leq}\frac{125}{\epsilon^2}+2<\infty.
\end{aligned} \tag{35}$$

Here, inequality (1) is from the fact that $\mathcal{E}_{k,t}=\mathcal{E}_{k,t'}$ for any time step t,t' such that $n_{k,t}=n_{k,t'}$; inequality (2) is by the definition of $\mathcal{E}^1_{k,t},\mathcal{E}^2_{k,t},\mathcal{E}^3_{k,t}$; inequality (3) is from Lemma 2 and 3; inequality (4) is straightforward algebra; inequality (5) uses the fact that $\frac{e^x}{1+x}\geq(1+\frac{x^2}{8})^2$.

Combining all four terms together, we obtain that

$$\mathbb{E}[\mathrm{Reg_T}]\leq\sum_{k\neq k^\star}\left(\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{(1+\epsilon)}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+2\sqrt{\log T}+\frac{2(1+\epsilon)\sigma_k^2}{\epsilon^2\Delta_k^2}+\frac{125}{\epsilon^2}+4\right)\Delta_k, \tag{36}$$

which finishes the proof. □

Furthermore, we can remove the dependency on ϵ in Theorem 2 using the following technical lemma:

Lemma 4 (Proposition 10 from [9])**.** *For any $G>0,\epsilon\in[0,\frac{1}{2}]$, the following holds:*

$$\frac{1}{\log\left(1+G\frac{(1-\epsilon)^2}{1+\epsilon}\right)}\leq\frac{1}{\log(1+G)}+\frac{10G}{(1+G)(\log(1+G))^2}\epsilon \tag{37}$$

Proof of Theorem 1. Using Lemma 4, denote $G_k=\frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}$ we can derive from (36) that

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg_T}]&\leq\sum_{k\neq k^\star}\left(\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{(1-\epsilon)^2}{(1+\epsilon)}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)}+2\sqrt{\log T}+\frac{2(1+\epsilon)\sigma_k^2}{\epsilon^2\Delta_k^2}+\frac{125}{\epsilon^2}+4\right)\Delta_k\\
&\leq\sum_{k\neq k^\star}\left(\frac{2\log T}{\log(1+G_k)}+\frac{10G_k\log T}{(1+G_k)(\log(1+G_k))^2}\epsilon+2\sqrt{\log T}+\frac{2(1+\epsilon)\sigma_k^2}{\epsilon^2\Delta_k^2}+\frac{125}{\epsilon^2}+4\right)\Delta_k.
\end{aligned} \tag{38}$$

Take $\epsilon = \frac{1}{2(\log T)^{1/3}}$, since $T \geq 4K \geq 8$, we know that $\epsilon = \frac{1}{2(\log T)^{1/3}} < \frac{1}{2}$, therefore the sample size condition in Theorem 1 is satisfied, and we have

$$\begin{aligned}\mathbb{E}[\mathrm{Reg_T}] &\leq \sum_{k\neq k^\star}\left(\frac{2\log T}{\log\left(1+G_k\right)} + \frac{5G_k(\log T)^{2/3}}{(1+G_k)(\log(1+G_k))^2} + 2\sqrt{\log T}\right.\\ &\qquad\left. + \frac{\sigma_k^2}{\Delta_k^2}(8(\log T)^{2/3} + 4(\log T)^{1/3}) + 500(\log T)^{2/3} + 4\right)\Delta_k\\ &\leq \sum_{k\neq k^\star}\left(\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)} + O\left((\log T)^{2/3}\right)\right)\Delta_k\end{aligned} \tag{39}$$

which concludes the proof. □

Proof of Corollary 1. Take $a_k = 1 + \frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}$, and $b_k = 1 + \frac{\Delta_k^2}{\sigma_k^2(1-\rho_k^2)}$, then $a_k \leq b_k$. Using the convexity of the function $\log(x)$, we have $\log(a_k) \geq \log(b_k) - \frac{b_k - a_k}{a_k}$, and hence

$$\frac{1}{\log(a_k)} \leq \frac{1}{\log(b_k)} + \frac{b_k - a_k}{a_k\log(a_k)\log(b_k)} \leq \frac{1}{\log(b_k)} + \frac{b_k - a_k}{a_k\log(a_k)^2}.$$

Notice that as long as $\delta_k \leq 1$ and $\delta_k = O((\log T)^{-2/3})$, we have

$$a_k = 1 + \frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2} \geq 1 + \frac{\Delta_k^2}{4\sigma_k^2}$$

$$b_k - a_k = \frac{\Delta_k^2}{\sigma_k^2}\frac{\delta_k + 2\sqrt{\delta_k}\sqrt{1-\rho_k^2}}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2(1-\rho_k^2)} \leq \frac{\Delta_k^2}{\sigma_k^2}\frac{\delta_k + 2\sqrt{\delta_k}\sqrt{1-\rho_k^2}}{(1-\rho_k^2)^2} = O((\log T)^{-1/3}).$$

Therefore, combining it with Theorem 1, we know that under the condition $N_k = \Omega((\log T)^{5/3})$, we have $\delta_k = O((\log T)^{-2/3})$ and

$$\begin{aligned}\mathbb{E}[\mathrm{Reg_T}] &\leq \sum_{k\neq k^\star}\left(\frac{2\log T}{\log\left(1+\frac{\Delta_k^2}{\sigma_k^2}\frac{1}{(\sqrt{1-\rho_k^2}+\sqrt{\delta_k})^2}\right)} + O\left((\log T)^{2/3}\right)\right)\Delta_k\\ &= \sum_{k\neq k^\star}\left(\frac{2\log T}{\log(a_k)} + O\left((\log T)^{2/3}\right)\right)\Delta_k\\ &\leq \sum_{k\neq k^\star}\left(\frac{2\log T}{\log(b_k)} + \frac{(b_k-a_k)\log T}{a_k\log(a_k)^2} + O\left((\log T)^{2/3}\right)\right)\Delta_k\\ &= \sum_{k\neq k^\star}\left(\frac{2\log T}{\log(1+\frac{\Delta_k^2}{\sigma_k^2(1-\rho_k^2)})} + O\left((\log T)^{2/3}\right)\right)\Delta_k,\end{aligned} \tag{40}$$

which finishes the proof. □