Leveraging (Biased) Information: Multi-armed Bandits with Offline Data

Wang Chi Cheung [1] **Lixing Lyu** [2]

Abstract

We leverage offline data to facilitate online learning in stochastic multi-armed bandits. The probability distributions that govern the offline data and the online rewards can be different. Without any non-trivial upper bound on their difference, we show that no non-anticipatory policy can outperform the UCB policy by (Auer et al. 2002), even in the presence of offline data. In complement, we propose an online policy MIN-UCB, which outperforms UCB when a non-trivial upper bound is given. MIN-UCB adaptively chooses to utilize the offline data when they are deemed informative, and to ignore them otherwise. MIN-UCB is shown to be tight in terms of both instance independent and dependent regret bounds. Finally, we corroborate the theoretical results with numerical experiments.

Keywords: multi-armed bandits, distributional shift, warm start

1. Introduction

Multi-armed bandit problem (MAB) is a classical model in sequential decision-making. In traditional models, online learning is initialized with no historical dataset. However, in many real world scenarios, historical datasets related to the underlying model exist before online learning begins. Those datasets could be beneficial to online learning. For instance, when a company launch a new product in Singapore, past sales data in the US could be of relevance. This prompts an intriguing question: Can we develop an approach that effectively utilizes offline data, and subsequently learns under bandit feedback? Such approaches carry significant allure, as they potentially reduce the amount of exploration in the learning process. Nevertheless, it is too strong and impractical to assume that the historical dataset and the online model follow the same probability distribution, given the ubiquity of distributional shifts. Therefore, it is desirable to design an adaptive policy that reaps the benefit of offline data if they closely match the online model, while judiciously ignores the offline data and learns from scratch otherwise.

Motivated by the above discussions, we consider a stochastic multi-armed bandit model with possibly biased offline dataset. The entire event horizon consists of a warm-start phase, followed by an online phase whereby decision making occurs. During the warm-start phase, the DM receives an offline dataset, consisting of samples governed by a latent probabiltiy distribution $P^{(\text{off})}$. The subsequent online phase is the same as standard multi-armed bandit model, except that the offline dataset can be incorporated in decision making. The random online rewards are governed by another latent probabiltiy distribution $P^{(\text{on})}$. The DM aims to maximize the total cumulative reward earned, or equivalently minimize the regret, during the online phase.

Intuitively, when $P^{(\text{off})}$, $P^{(\text{on})}$ are "far apart", the DM should conduct online learning from scratch and ignore the offline dataset. For example, the vanilla UCB policy by (Auer et al., 2002) incurs expected regret at most

$$O\left(\sum_{a:\Delta(a)>0} \frac{\log T}{\Delta(a)}\right), \tag{1}$$

where T is the length of online phase, and $\Delta(a) \geq 0$ is the difference between the expected reward of an optimal arm and arm a (also see Section 2). The bound (1) holds for all $P^{(\text{off})}$, $P^{(\text{on})}$. In contrary, when $P^{(\text{off})}$, $P^{(\text{on})}$ are "sufficiently close", the DM should incorporate the offline data into online learning and avoid unnecessary exploration. For example, when $P^{(\text{off})} = P^{(\text{on})}$, HUCB1 (Shivaswamy & Joachims, 2012) and MonUCB (Banerjee et al., 2022) incur expected regret at most

$$O\left(\sum_{a:\Delta(a)>0} \max\left\{\frac{\log T}{\Delta(a)} - T_{\text{S}}(a)\Delta(a), 0\right\}\right), \tag{2}$$

where $T_{\text{S}}(a)$ is the number of offline samples on arm a. (2) is a better regret bound than (1). However, (2) only

[*]Equal contribution [1]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore [2]Institute of Operations Research and Analytics, National University of Singapore, Singapore. Correspondence to: Wang Chi Cheung <isecwc@nus.edu.sg>, Lixing Lyu <lyulixing@u.nus.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

holds when $P^{(\text{off})} = P^{(\text{on})}$. The above inspires the following question:

(Q) Can the DM *outperform* the vanilla UCB, i.e. achieves a better regret bound than (1) when $P^{(\text{off})} = P^{(\text{on})}$, while retains the regret bound (1) for general $P^{(\text{off})}, P^{(\text{on})}$?

The answer to (Q) turns out to be somewhat mixed. Our novel contributions shed light on (Q) as follows:

Impossibility result. We show in Section 3 that no non-anticipatory policy can achieve the aim in (Q). Even with an offline dataset, no non-anticipatory policy can outperform the vanilla UCB without any additional knowledge or restriction on the difference between $P^{(\text{off})}, P^{(\text{on})}$.

Algorithm design and analysis. To bypass the impossibility result, we endow the DM with an auxiliary information dubbed *valid bias bound* V, which is an information additional to the offline dataset. The bound V serves as an upper bound on the difference between $P^{(\text{off})}$ and $P^{(\text{on})}$. We propose the MIN-UCB policy, which achieves a strictly better regret bound than (1) when $P^{(\text{off})}, P^{(\text{on})}$ are "sufficiently close". In particular, our regret bound reduces to (2) if the DM knows $P^{(\text{off})} = P^{(\text{on})}$. We provide both instance dependent and independent regret upper bounds on MIN-UCB, and we show their tightness by providing the corresponding regret lower bounds. Our analysis also pins down the precise meaning of "far apart" and "sufficientlly close" in the above discussion. The design and analysis of MIN-UCB are provided in Section 4.

In particular, when $P^{(\text{off})} = P^{(\text{on})}$, our analysis provide two novel contributions. Firstly, we establish the tightness of the instance dependent bound (2). Secondly, we establish a pair of instance independent regret upper and lower bounds, which match each other up to a logarithmic factor. MIN-UCB is proved to achieve the upper bound when $P^{(\text{off})} = P^{(\text{on})}$. Both the upper and lower bounds are novel in the literature, and we show that the optimal regret bound involves the optimum of a novel linear program.

1.1. Related Works

Multi-armed bandits with offline data has been studied in multiple works. (Shivaswamy & Joachims, 2012; Banerjee et al., 2022) are the most relevant works, focusing on the special case of $P^{(\text{on})} = P^{(\text{off})}$. They only provide instance dependent regret upper bound, while we both instance dependent and independent bounds. Online learning with offline data under $P^{(\text{on})} = P^{(\text{off})}$ is also studied in dynamic pricing settings (Bu et al., 2020) and reinforcement learning settings (Hao et al., 2023; Wagenmaker & Pacchiano, 2023), as well as when historical data are sequentially provided (Gur & Momeni, 2022), or provided as clustered data (Bouneffouf et al., 2019; Ye et al., 2020; Tennenholtz et al., 2021). (Zhang et al., 2019) is another closely related work on contextual bandits that allows $P^{(\text{on})} \neq P^{(\text{off})}$. We provide more discussions on it in Appendix A.1.

Bayesian policies, such as the Thompson sampling policy (TS), are popular for incorporating offline data into online learning. When the online and offline reward distributions are identical, offline data can be used to construct well-specified priors, which enables better regret bounds than (Russo & Van Roy, 2014; Russo & Roy, 2016; Liu & Li, 2016) when the offline data size increases. By contrast, (Liu & Li, 2016; Simchowitz et al., 2021) show that the TS initialized with mis-specified priors generally incurs worse regret bounds than state-of-the-art online policies without any offline data. Our work focuses on an orthogonal issue of deciding whether to incorporate the possibly biased offline data in online learning.

Our problem model closely relates to machine learning under domain shift. (Si et al., 2023) investigates a offline policy learning problem using historical data on contextual bandits under changing environment from a distributionally robust optimization perspective. (Chen et al., 2022) studies a variant in RL context, and their policy is also endowed with an upper bound on the distributional drift between offline and online models. We provide more discussion in Appendix A.3. Numerous research works consider the issue of distributional drift in supervised learning (Crammer et al., 2008; Mansour et al., 2009; Ben-David et al., 2010) and stochastic optimization (Besbes et al., 2022).

Our work is related to research on online learning with advice, which focuses on improving performance guarantees with hints or predictions. Indeed, our offline dataset could serve as a hint to an online learner. A variety of results are derived in multi-armed bandits (Wei & Luo, 2018), linear bandits (Cutkosky et al., 2022), contextual bandits (Wei et al., 2020b), and full feedback settings (Steinhardt & Liang, 2014; Rakhlin & Sridharan, 2013a;b). These works do not apply in our setting, since they do not consider refining instance dependent bounds. We provide more discussions in Appendix A.2. Our theme of whether to utilize the potentially biased offline data is related to online model selection, studied in (Agarwal et al., 2017; Pacchiano et al., 2020; Lee et al., 2021; Cutkosky et al., 2021; Pacchiano et al., 2023). While these works take a black-box approach, we tailor-make our decision making on whether to utilize offline data with a novel UCB algorithm and the auxilliary input V, and the latter is not studied in the abovemoentioned works.

Notation. We denote $\mathcal{N}(\mu, 1)$ as the Gaussian disribution with mean μ and variance 1. We abbreviate identically and independently distributed as "i.i.d.". The relationship $R \sim P$ means that the random variable R follows the probability distribution P.

2. Model

We consider a stochastic K-armed bandit model with possibly biased offline data. Denote $\mathcal{A} = \{1, \dots, K\}$ as the set of K arms. The event horizon consists of a warm-start phase, then an online phase. During the warm-start phase, the DM receive a collection of $T_{\mathrm{S}}(a) \in \mathbb{Z}_{\geq 0}$ i.i.d. samples, denoted $S(a) = \{X_s(a)\}_{s=1}^{T_{\mathrm{S}}(a)}$, for each $a \in \mathcal{A}$. We postulate that $X_1(a), \dots, X_{T_{\mathrm{S}}(a)}(a) \sim P_a^{(\mathrm{off})}$, the *offline reward distribution* on arm a.

The subsequent online phase consists of T time steps. For time steps $t = 1, \dots, T$, the DM chooses an arm $A_t \in \mathcal{A}$ to pull, which gives the DM a random reward R_t. Pulling arm a in the online phase generates a random reward $R(a) \sim P_a^{(\mathrm{on})}$, dubbed the *online reward distribution* on arm a. We emphasize that $P^{(\mathrm{on})} = \{P_a^{(\mathrm{on})}\}_{a \in \mathcal{A}}$ needs not be the same as $P^{(\mathrm{off})} = \{P_a^{(\mathrm{off})}\}_{a \in \mathcal{A}}$. Finally, the DM proceeds to time step $t+1$, or terminates when $t = T$.

The DM chooses arms $A_1, \dots, A_T$ with a *non-anticipatory policy* $\pi = \{\pi_t\}_{t=1}^{\infty}$. The function π_t maps the observations $H_{t-1} = (S = \{S(a)\}_{a \in \mathcal{A}}, \{(A_s.R_s)\}_{s=1}^{t-1})$ collected at the end of $t-1$, to an element in $\{x \in \mathbb{R}_{\geq 0}^K : \sum_{a \in \mathcal{A}} x_a = 1\}$. The quantity $\pi_t(a \mid H_{t-1}) \in [0, 1]$ is the probability of $A_t = a$, conditioned on H_{t-1} under policy π. While our proposed algorithms are deterministic, i.e. $\pi_t(a \mid H_{t-1}) \in \{0, 1\}$ always, we allow randomized policies in our regret lower bound results. The policy π could utilize S, which provides a potentially biased estimate to the online model since we allow $P^{(\mathrm{on})} \neq P^{(\mathrm{off})}$.

Altogether, the underlying instance I is specified by the tuple $(\mathcal{A}, \{T_{\mathrm{S}}(a)\}_{a \in \mathcal{A}}, P, T)$, where $P = (P^{(\mathrm{off})}, P^{(\mathrm{on})})$. The DM only knows $\mathcal{A}$ and S, but not P and T, at the start of the online phase. For every $a \in \mathcal{A}$, we assume that both $P_a^{(\mathrm{off})}, P_a^{(\mathrm{on})}$ are 1-subGaussian, which is also known to the DM. For $a \in \mathcal{A}$, denote $\mu^{(\mathrm{on})}(a) = \mathbb{E}_{R(a) \sim P_a^{(\mathrm{on})}}[R(a)]$, and $\mu^{(\mathrm{off})}(a) = \mathbb{E}_{R(a) \sim P_a^{(\mathrm{off})}}[R(a)]$. We do not have any boundedness assumption on $\mu^{(\mathrm{on})}(a), \mu^{(\mathrm{off})}(a)$.

The DM aims to maximize the total reward earned in the online phase. We quantify the performance guarantee of the DM's non-anticipatory policy by its regret. Define $\mu_*^{(\mathrm{on})} = \max_{a \in \mathcal{A}} \mu^{(\mathrm{on})}(a)$, and denote $\Delta(a) = \mu_*^{(\mathrm{on})} - \mu^{(\mathrm{on})}(a)$ for each $a \in \mathcal{A}$. The DM aims to design a non-anticipatory policy π minimize the regret

$$\mathrm{Reg}_T(\pi, P) = T\mu_*^{(\mathrm{on})} - \sum_{t=1}^{T} \mu^{(\mathrm{on})}(A_t) = \sum_{t=1}^{T} \Delta(A_t) \quad (3)$$

despite the uncertainty on P. While (3) involves only $\mu^{(\mathrm{on})}$ but not $\mu^{(\mathrm{off})}$, the choice of A_t is influenced by both the offline data S and the online data $\{A_s, R_s\}_{s=1}^{t-1}$.

As mentioned in the introduction and detailed in the forthcoming Section 3, the DM cannot outperform the vanilla UCB policy with a non-anticipatory policy. Rather, the DM requires information additional to the offline dataset and a carefully designed policy. We consider the following auxiliary input in our algorithm design.

Auxiliary input: valid bias bounds. We say that $V = \{V(a)\}_{a \in K} \in (\mathbb{R} \cup \{\infty\})^{|\mathcal{A}|}$ is a valid *bias bound* on an instance I if

$$V(a) \geq |\mu^{(\mathrm{off})}(a) - \mu^{(\mathrm{on})}(a)| \quad \text{for each } a \in \mathcal{A}. \quad (4)$$

In our forthcoming policy MIN-UCB, the DM is endowed the auxiliary input V in addition to the offline dataset S before the online phase begins. The quantity $V(a)$ serves as an upper bound on the amount of distributional shift from $P_a^{(\mathrm{off})}$ to $P_a^{(\mathrm{on})}$. The condition (4) always holds in the trivial case of $V(a) = \infty$, which is when the DM has no additional knowledge. By contrast, when $V(a) \neq \infty$, the DM has a non-trivial knowledge on the difference $|\mu^{(\mathrm{off})}(a) - \mu^{(\mathrm{on})}(a)|$.

The knowledge on an upper bound such as V is in line with the model assumptions in research works on learning under distributional shift. Similar upper bounds are assumed in the contexts of supervised learning (Crammer et al., 2008), stochastic optimization (Besbes et al., 2022), offline policy learning on contextual bandits (Si et al., 2023), multi-task bandit learning (Wang et al., 2021), for example. Methodologies for constructing such upper bounds are provided in some literature. For instance, (Blanchet et al., 2019) designs an approach based on several machine learning estimators, such as LASSO. (Si et al., 2023) provides some managerial insights on how to construct them empirically. (Chen et al., 2022) constructs them through cross-validation, and implement it in a case study.

3. An Impossibility Result

We illustrate the impossibility result with two instances I_P, I_Q. The instances I_P, I_Q share the same $\mathcal{A} = \{1, 2\}, \{T_{\mathrm{S}}(a)\}_{a \in \mathcal{A}}$ and T. However, I_P, I_Q differ in their respective reward distributions P, Q, and I_P, I_Q have different optimal arms. Consider a fixed but arbitrary $\beta \in (0, 1/2)$. Instance I_P has well-aligned reward distributions $P^{(\mathrm{off})} = P^{(\mathrm{on})}$, where

$$P_1^{(\mathrm{on})} = \mathcal{N}(0, 1), \quad P_2^{(\mathrm{on})} = \mathcal{N}(-T^{-\beta}, 1).$$

Note that $\Delta(2) = T^{-\beta}$. In I_P, an offline dataset provides useful hints on arm 1 being optimal. For example, given $T_{\mathrm{S}}(1) = T_{\mathrm{S}}(2) \geq 128(T^{2\beta} - T^{2\beta-\epsilon}) \log T$ where $\epsilon \in (0, \beta)$, the existing policies H-UCB and MonUCB achieve an expected regret (see Equation (2)) at most

$$O\left(\frac{\log T}{\Delta(2)} - T_{\mathrm{S}}(2)\Delta(2)\right) = O(\log T(T^{\beta} - T^{\beta} + T^{\beta-\epsilon})) = O(T^{\beta-\epsilon} \log T),$$

which strictly outperforms the vanilla UCB policy that incurs expected regret $O(\log T/\Delta(2)) = O(T^\beta \log T)$. Despite the apparent improvement, the following claim shows that any non-anticipatory policy (not just H-UCB or MonUCB) that outperforms the vanilla UCB on I_P would incur a larger regret than the vanilla UCB on a suitably chosen I_Q.

Proposition 3.1. *Let $T_S(1), T_S(2)$ be arbitrary. Consider an arbitrary non-anticipatory policy π (which only possesses the offline dataset S but not the auxiliary input V) satisfies $\mathbb{E}[Reg_T(\pi, P)] \le CT^{\beta-\epsilon}\log T$ on instance I_P, where $\epsilon \in (0, \beta)$, $C > 0$ is an absolute constant, and the horizon T is so large that $C < T^\epsilon/(4\log T)$. Set $Q = (Q^{(off)}, Q^{(on)})$ in the instance I_Q as*

$$Q^{(off)} = P^{(off)}, \quad Q_1^{(on)} = P_1^{(on)}, \text{ but}$$
$$Q_2^{(on)} = \mathcal{N}\left(\frac{1}{\sqrt{C\log T}T^{\beta-(\epsilon/2)}} - \frac{1}{T^\beta}, 1\right).$$

The following inequality holds:

$$\begin{aligned}\mathbb{E}[Reg(\pi, Q)] &\ge \frac{T^{1-\beta}}{4}\cdot e^{-2} - CT^{\beta-\epsilon}\log T \\ &= \Omega(T^{1-\beta}) > \Omega(\sqrt{T}).\end{aligned} \tag{5}$$

For the proof of Proposition 3.1, we set up some notation. Consider a non-anticipatory policy π and an instance I with reward distribution P. We denote $\rho_{P,\pi}$ as the joint probability distribution function on $S, A_1, R_1, \ldots, A_T, R_T$, the concatenation of the offline dataset S and the online trajectory under policy π on instance I. For a $\sigma(S, A_1, R_1, \ldots, A_T, R_T)$-measurable event E, we denote $\Pr_{P,\pi}(E)$ as the probability of E holds under $\rho_{P,\pi}$. For a $\sigma(S, A_1, R_1, \ldots, A_T, R_T)$-measurable random variable Y, we denote $\mathbb{E}_{P,\pi}[Y]$ as the expectation of Y under the joint probability distribution function $\rho_{P,\pi}$. In our analysis, we make use of Y being $\text{Reg}_T(\pi, P)$ or $N_T(a)$ for an arm a. To lighten our notational burden, we abbreviate $\mathbb{E}_{P,\pi}[\text{Reg}_T(\pi, P)]$ as $\mathbb{E}[\text{Reg}_T(\pi, P)]$.

Proof of Proposition 3.1. By the condition $C < T^\epsilon/(4\log T)$, we know that $\frac{1}{\sqrt{C\log T}T^{\beta-(\epsilon/2)}} - \frac{1}{T^\beta} > \frac{1}{T^\beta} > 0$, so in the instance with Q, arm 2 is the unqiue optimal arm. Consider the event $E = \{N_T(2) > T/2\}$.

$$\begin{aligned}&\mathbb{E}[\text{Reg}(\pi, P)] + \mathbb{E}[\text{Reg}(\pi, Q)] \\ \ge&\frac{1}{T^\beta}\cdot\frac{T}{2}\cdot\Pr_{P,\pi}(E) \\ &+\left[\frac{1}{\sqrt{C\log T}T^{\beta-(\epsilon/2)}} - \frac{1}{T^\beta}\right]\cdot\frac{T}{2}\Pr_{Q,\pi}(E^C) \\ \ge&\frac{T^{1-\beta}}{2}\cdot\left[\Pr_{P,\pi}(E) + \Pr_{Q,\pi}(E^C)\right] \\ \ge&\frac{T^{1-\beta}}{4}\cdot\exp\Big[-\mathbb{E}_{\pi,P}[N_T(2)]\cdot\text{KL}(P_2^{(\text{on})}, Q_2^{(\text{on})}) \\ &-T_S(2)\cdot\text{KL}(P_2^{(\text{off})}, Q_2^{(\text{off})})\Big] \qquad (6) \\ =&\frac{T^{1-\beta}}{4}\cdot\exp\left[-\mathbb{E}_{\pi,P}[N_T(2)]\cdot\text{KL}(P_2^{(\text{on})}, Q_2^{(\text{on})})\right]. \qquad (7)\end{aligned}$$

Step (6) is again by the chain rule in Theorem C.2 in the Appendix. Step (7) is because $P^{(\text{off})} = Q^{(\text{off})}$. It is evident that

$$\text{KL}(P_2^{(\text{on})}, Q_2^{(\text{on})}) = \frac{1}{2CT^{2\beta-\epsilon}\log T},$$

and by the Claim assumption that $\text{Reg}(\pi, P) = (1/T^\beta)\mathbb{E}_{\pi,P}[N_T(2)] \le CT^{\beta-\epsilon}\log T$, we have

$$\mathbb{E}_{\pi,P}[N_T(2)] \le CT^{2\beta-\epsilon}\log T,$$

which leads to

$$\mathbb{E}[\text{Reg}(\pi, P)] + \mathbb{E}[\text{Reg}(\pi, Q)] \ge \frac{T^{1-\beta}}{4}\cdot e^{-2}.$$

Again by the Proposition assumption $\mathbb{E}[\text{Reg}(\pi, P)] \le CT^{\beta-\epsilon}\log T$, we arrive at the claimed inequality, and the Proposition is proved. □

We first remark on I_Q. Different from I_P, arm 2 is the optimal arm in I_Q. Thus, the offline data from $Q^{(\text{off})}$ provides the misleading information that arm 1 has a higher expected reward. In addition, note that $Q^{(\text{off})} = P^{(\text{off})}$, so the offline datasets in instances I_P, I_Q are identically distributed. The proposition suggests that no non-anticipatory policy can simultaneously (i) establish that S is useful for online learning in I_P to improve upon the UCB policy, (ii) establish that S is to be disregarded for online learning in I_Q to match the UCB policy, negating the short-lived conjecture in (Q).

Lastly, we allow $T_S(1), T_S(2)$ to be arbitrary. In the most informative case of $T_S(1) = T_S(2) = \infty$, meaning that the DM knows $P^{(\text{off})}, Q^{(\text{off})}$ in I_P, I_Q respectively, the DM still cannot achieve (i, ii) simultaneously. Indeed, the assumed improvement $\mathbb{E}[\text{Reg}_T(\pi, P)] \le CT^{\beta-\epsilon}\log T$ limits the number of pulls on arm 2 in I_P during the online phase. We harness this property to construct an upper bound on the KL divergence between on online dynamics on I_P, I_Q,

Algorithm 1 Policy MIN-UCB

Input: Valid bias bound V on the instance, confidence parameter $\{\delta_t\}_{t=1}^{\infty}$, offline samples S.

For each $a \in \mathcal{A}$, compute $\hat{X}(a) = \frac{\sum_{s=1}^{T_{\mathrm{S}}(a)} X_s(a)}{T_{\mathrm{S}}(a)}$, and initialize $\hat{R}_1(a) = 0$.

At $t = 1, \ldots, K$, pull each arm once, then set $N_{K+1}(a) = 1$ for all a.

for $t = K+1, \ldots, T$ **do**

For all $a \in \mathcal{A}$, compute the vanilla UCB

$$\mathrm{UCB}_t(a) = \hat{R}_t(a) + \mathrm{rad}_t(a),$$

where $\mathrm{rad}_t(a) = \sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)}}$.

For all $a \in \mathcal{A}$, compute the warm-start $\mathrm{UCB}_t^{\mathrm{S}}(a) =$

$$\frac{N_t(a) \cdot \hat{R}_t(a) + T_{\mathrm{S}}(a) \cdot \hat{X}(a)}{N_t(a) + T_{\mathrm{S}}(a)} + \mathrm{rad}_t^{\mathrm{S}}(a), \quad (8)$$

where $\mathrm{rad}_t^{\mathrm{S}}(a) =$

$$\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a) + T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)}{N_t(a) + T_{\mathrm{S}}(a)} \cdot V(a). \quad (9)$$

Select

$$A_t \in \mathrm{argmax}_{a \in \mathcal{A}} \left\{ \min \left\{ \mathrm{UCB}_t(a), \mathrm{UCB}_t^{\mathrm{S}}(a) \right\} \right\}.$$

Observe $R_t \sim P_{A_t}^{(\mathrm{on})}$. For each $a \in \mathcal{A}$, update $N_{t+1}(a) = N_t(a) + \mathbf{1}(A_t = a)$,

$$\hat{R}_{t+1}(a) = \begin{cases} \frac{N_t(a) \cdot \hat{R}_t(a)}{N_t(a)+1} + \frac{R_t}{N_t(a)+1} & a = A_t \\ \hat{R}_t(a) & a \neq A_t \end{cases}.$$

end for

which in turns implies the DM cannot differentiate between I_P, I_Q under policy π. The KL divergence argument utilizes a chain rule (see Appendix C.1) on KL divergence adapted to our setting.

4. Design and Analysis of MIN-UCB

After the impossibility result, we present our MIN-UCB policy in Algorithm 1. MIN-UCB follows the optimism-in-face-of-uncertainty principle. At time $t > K$, MIN-UCB sets $\min\{\mathrm{UCB}_t(a), \mathrm{UCB}_t^{\mathrm{S}}(a)\}$ as the UCB on $\mu^{(\mathrm{on})}(a)$. While $\mathrm{UCB}_t(a)$ follows (Auer et al., 2002), the construction of $\mathrm{UCB}_t^{\mathrm{S}}(a)$ in (8) adheres to the following ideas that incorporates a valid bias bound V, the auxiliary input.

In (8), the sample mean $\frac{N_t(a) \cdot \hat{R}_t(a) + T_{\mathrm{S}}(a) \cdot \hat{X}(a)}{N_t(a) + T_{\mathrm{S}}(a)}$ incorporates both online and offline data. Then, we inject optimism by adding $\mathrm{rad}^{\mathrm{S}}(a)$, set in (9). The square-root term in (9) reflects the optimism in view of the randomness of the sample mean. The term $\frac{T_{\mathrm{S}}(a)}{N_t(a) + T_{\mathrm{S}}(a)} \cdot V(a)$ in (9) injects optimism to correct the potential bias from the offline dataset. The correction requires the auxiliary input V.

Finally, we adopt $\mathrm{UCB}_t^{\mathrm{S}}(a)$ as the UCB when the bias bound $V(a)$ is so small and the sample size $T_{\mathrm{S}}(a)$ is so large that $\mathrm{UCB}_t(a) > \mathrm{UCB}_t^{\mathrm{S}}(a)$. Otherwise, $P^{(\mathrm{off})}$ is deemed "far away" $P^{(\mathrm{on})}$, and we follow the vanilla UCB. Our adaptive decisions on whether to incorporate offline data at each time step is different from the vanilla UCB policy that always ignores offline data, or from HUCB and MonUCB that always incorporate all offline data.

Lastly, we remark on V. Suppose that the DM possesses no knowledge on $P^{(\mathrm{off})}, P^{(\mathrm{on})}$ additional to the offline dataset S at the start of the online phase. Then, it is advisable to set $V(a) = \infty$ for all a, which specializes MIN-UCB to the vanilla UCB policy. Indeed, the theoretical impossibility result in Section 3 dashes out any hope for adaptively tuning V in MIN-UCB with online data, in order to outperform the vanilla UCB. By contrast, when the DM has a non-trivial upper bound $V(a)$ on $\left|\mu^{(\mathrm{off})}(a) - \mu^{(\mathrm{on})}(a)\right|$, we show that can uniformly out-perform the vanilla UCB that direclty ignores the offline data.

We next embark on the analysis on MIN-UCB, by establishing instance dependent regret bounds in Section 4.1 and instance independent regret bounds in Section 4.3. Both sections relies on considering the following events of accurate estimations by $\mathrm{UCB}_t(a), \mathrm{UCB}_t^{\mathrm{S}}(a)$. For each t, define $\mathcal{E}_t = \cap_{a \in \mathcal{A}} (\mathcal{E}_t(a) \cap \mathcal{E}_t^{\mathrm{S}}(a))$, where

$$\mathcal{E}_t(a) = \left\{ \mu^{(\mathrm{on})}(a) \leq \mathrm{UCB}_t(a) \leq \mu^{(\mathrm{on})}(a) + 2\mathrm{rad}_t(a) \right\},$$

$$\mathcal{E}_t^{\mathrm{S}}(a) = \left\{ \mu^{(\mathrm{on})}(a) \leq \mathrm{UCB}_t^{\mathrm{S}}(a) \leq \mu^{(\mathrm{on})}(a) + \mathrm{rad}_t^{\mathrm{S}}(a) + \left[\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a) + T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a) \cdot (\mu^{(\mathrm{off})}(a) - \mu^{(\mathrm{on})}(a))}{N_t(a) + T_{\mathrm{S}}(a)} \right] \right\}.$$

While $\mathcal{E}_t(a)$ incorporates estimation error due to stochastic variations, $\mathcal{E}_t^{\mathrm{S}}(a)$ additionally involves the potential bias of the offline data.

Lemma 4.1. $\Pr(\mathcal{E}_t) \geq 1 - 2K\delta_t$.

The proof of Lemma 4.1 is provided in Appendix B.1.

4.1. Instance Dependent Regret Bounds

Both our instance dependent regret upper and lower bounds depend on the following discrepancy measure

$$\omega(a) = V(a) + (\mu^{(\mathrm{off})}(a) - \mu^{(\mathrm{on})}(a)) \quad (10)$$

on an arm a. Note that $\omega(a)$ depends on both the valid bias bound $V(a)$ and the model parameters $\mu^{(\mathrm{off})}(a), \mu^{(\mathrm{on})}(a))$. By the validity of V, we know that $\omega(a) \in [0, 2V(a)]$.

Theorem 4.2. *Consider policy MIN-UCB, which inputs a valid bias bound V on the underlying instance I and $\delta_t = \frac{1}{2Kt^2}$ for $t = 1, 2, \ldots$. We have $\mathbb{E}[Reg_T(\text{MIN-UCB}, P)] \leq$*

$$O\left(\sum_{a:\Delta(a)>0} \max\left\{\frac{\log(T)}{\Delta(a)} - Sav_0(a), \Delta(a)\right\}\right), \quad (11)$$

where $\sum_{a:\Delta(a)>0}$ is over $\{a \in \mathcal{A} : \Delta(a) > 0\}$, and

$$Sav_0(a) = T_S(a) \cdot \Delta(a) \cdot \max\left\{1 - \frac{\omega(a)}{\Delta(a)}, 0\right\}^2. \quad (12)$$

Theorem 4.2 is proved in the Appendix B.2. We provide a sketch proof in Section 4.2. A full explicit bound is in (30) in the Appendix. Comparing with the regret bound (1) by the vanilla UCB, MIN-UCB provides an improvement on the regret order bound by the saving term $\text{Sav}_0(a)$.

First, note that $\text{Sav}_0(a) \geq 0$ always. The case of $\omega(a) < \Delta(a)$ means that the reward distributions $P_a^{(\text{off})}, P_a^{(\text{on})}$ are "sufficiently close". The offline data on arm a are incorporated by MIN-UCB to improve upon the vanila UCB. By contrast, the case of $\omega(a) \geq \Delta(a)$ means that $P_a^{(\text{off})}, P_a^{(\text{on})}$ are "far apart", so that the offline data on arm a are ignored in the learning process, while MIN-UCB still matches the performance guarantee of the vanilla UCB. We emphasize that $\omega(a)$ and $\Delta(a)$ are both latent, so MIN-UCB conducts the above adaptive choice under uncertainty.

Next, in the case of $\omega(a) < \Delta(a)$ where the offline data on a are beneficial, observe that $\text{Sav}_0(a)$ is increasing in $T_S(a)$. This monotonicity adheres to the intuition that more offline data leads to a higher reward when $P_a^{(\text{off})}, P_a^{(\text{on})}$ are "sufficiently close". The term $\text{Sav}_0(a)$ dcreases as $\omega(a)$ increses. Indeed, a smaller $V(a)$ means a tighter upper bound on $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$ is known to the DM, which leads to a better peformance.

Interestingly, $\text{Sav}_0(a)$ increases when $\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)$ decreases. Paradoxically, $\text{Sav}_0(a)$ increases with $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$, *under the case of $\mu^{(off)}(a) < \mu^{(on)}(a)$*. Indeed, the saving term $\text{Sav}_0(a)$ depends not only on the magnitude of the distributional shift, but also the direction. To gain intuition, consider the case of $\mu^{(\text{off})}(a) < \mu^{(\text{on})}(a)$. On $\mu^{(\text{on})}(a)$, the offline dataset on a provides a pessimistic estimate, which encourages the DM to eliminate arm a (which is sub-optimal), hence improves the DM's performance.

Lastly, we compare to existing baselines. By setting $V(a) = \infty$ for all a, the regret bound (11) reduces to (1) of the vanilla UCB. In the case when $P^{(\text{on})} = P^{(\text{off})}$ and it is known to the DM, setting $V(a) = 0$ for all a reduces the regret bound (11) to (2). In both cases we achieve the state-of-the-art regret bounds, while additionally we quantify the impact of distributional drift to the regret bound. Finally, the ARROW-CB by (Zhang et al., 2019) can be applied to our model. While ARROW-CB does not require any auxiliary input, their regret bound is at least $\Omega(K^{1/3}T^{2/3})$. ARROW-CB provides improvment to purely online policies in the contextual bandit setting, as discussed in Appednix A.1.

We next provides a regret lower bound that nearly matches (11). To proceed, we consider a subset of instances with bounded distributional drift, as defined below.

Definition 4.3. Let $V = (V(a))_{a\in\mathcal{A}} \in \mathbb{R}^K_{\geq 0}$. We define $\mathcal{I}_V = \{I : |\mu^{(\text{on})}(a) - \mu^{(\text{off})}(a)| \leq V(a) \text{ for all } a \in \mathcal{A}\}$.

Knowing a valid bias bound V is equivalent to knowing that the underlying instance I lies in the subset $\mathcal{I}_V$, leading to the following which is equivalent to Theorem 4.2:

Corollary 4.4. *For any $V \in \mathbb{R}^K_{\geq 0}$, the MIN-UCB policy with input V and $\delta_t = 1/(2Kt^2)$ satisfies $\mathbb{E}[Reg_T(\text{MIN-UCB}, P)] \leq$ (11), for all $I \in \mathcal{I}_V$.*

Our regret lower bounds involve instances with Gaussian rewards.

Definition 4.5. An instance I is a Gaussian instance if $P_a^{(\text{on})}, P_a^{(\text{off})}$ are Gaussian distributions with variance one, for every $a \in \mathcal{A}$.

Similar to existing works, we focus on consistent policies.

Definition 4.6. For $C > 0, p \in (0, 1)$ and a collection $\mathcal{I}$ of instances, a non-anticipatory policy π is said to be (C, p)-consistent on $\mathcal{I}$, if for all $I \in \mathcal{I}$, it holds that $\mathbb{E}[\text{Reg}_T(P, \pi)] \leq CT^p$.

For example, the vanilla UCB is $(C, 1/2)$ consistent for some absolute constant $C > 0$.

Theorem 4.7. *Let $V \in \mathbb{R}^K_{\geq 0}$ be arbitrary, and consider a fixed but arbitrary Gaussian instance $I \in \mathcal{I}_V$. For any (C, p)-consistent policy π on $\mathcal{I}_V$, we have the following regret lower bound of π on I: $\mathbb{E}[Reg_T(\pi, P)] \geq$*

$$\sum_{a:\Delta(a)>0} \left\{\frac{2(1-p)}{(1+\epsilon)^2} \cdot \frac{\log T}{\Delta(a)} - Sav_\epsilon(a) + \kappa_\epsilon(a)\right\} \quad (13)$$

holds for any $\epsilon \in (0, 1]$, where

$$Sav_\epsilon(a) = T_S(a) \cdot \Delta(a) \cdot \max\left\{1 - \frac{\omega(a)}{(1+\epsilon)\Delta(a)}, 0\right\}^2, \quad (14)$$

and $\kappa_\epsilon(a) = \frac{1}{2(1+\epsilon)^2\Delta(a)} \log \frac{\epsilon\Delta(a)}{8C}$ is independent of T or $T_S(a)$.

Theorem 4.7 is proved in Appendix C.2. Modulo the additive term $\kappa_\epsilon(a)$, the regret upper bound (11) and lower bound (13) are nearly matching. Both bounds feature the $\Theta((\log T)/\Delta(a))$ regret term that is classical in the literature (Lattimore & Szepesvári, 2020). Importantly, the

lower bound (13) features the regret saving term $\text{Sav}_\epsilon(a)$, which closely matches the term $\text{Sav}_0(a)$ in the upper bound. Indeed, $\epsilon \in (0,1]$ is arbitrary, $\text{Sav}_\epsilon(a)$ tends to the term $\text{Sav}_0(a)$ in the regret upper bound (11) as we tend ϵ to 0. The close match highlights the fundamental nature of the discrepancy measure $\omega(a)$. In addition, the lower bound (13) carries the insight that the offline data on arm a facilitates improvement over the vanilla UCB if and only if $\omega(a) < (1+\epsilon)\Delta(a)$, which is consistent with our insights from the regret upper bound (11) by tending $\epsilon \to 0$.

Lastly, in the case $P^{(\text{off})} = P^{(\text{on})}$, the lower bound (13) reduces to $\Omega(\sum_{a:\Delta(a)>0}[\frac{\log T}{\Delta(a)} - T_{\text{S}}(a)\Delta(a) + \kappa_\epsilon(a)])$, which ascertains the near tightness of (2) by HUCB1 and MonUCB. A nearly tight regret lower bound is novel in the literature.

4.2. Sketch Proof of Theorem 4.2

Theorem 4.2 is a direct conclusion of the following lemma:

Lemma 4.8. *Let a be a sub-optimal arm, that is $\Delta(a) > 0$. Conditioned on the event $\mathcal{E}_t$, if it holds that*

$$N_t(a) > 32 \cdot \frac{\log(4Kt^4)}{\Delta(a)^2} - T_S(a) \cdot \max\left\{1 - \frac{\omega(a)}{\Delta(a)}, 0\right\}^2,$$

then $A_t \neq a$ with certainty.

Note $\frac{\text{Sav}_0(a)}{\Delta(a)} = T_{\text{S}}(a) \cdot \max\{1 - \frac{\omega(a)}{\Delta(a)}, 0\}^2$. The Lemma is fully proved in Appendix B.2, and a sketch proof is below.

Sketch Proof of Lemma 4.8. Consider

Case 1a: $\omega(a) < \Delta(a)$, and $\frac{\text{Sav}_0(a)}{\Delta(a)} \geq \frac{16\log(4Kt^4)}{\Delta(a)^2}$,

Case 1b: $\omega(a) < \Delta(a)$, and $\frac{\text{Sav}_0(a)}{\Delta(a)} < \frac{16\log(4Kt^4)}{\Delta(a)^2}$,

Case 2: $\omega(a) \geq \Delta(a)$.

Both **Cases 1b, 2** imply $N_t(a) > 16 \cdot \frac{\log(4Kt^4)}{\Delta(a)^2}$. Following (Auer et al., 2002), we can then deduce

$$\begin{aligned}\text{UCB}_t(a) &\leq \mu^{(\text{on})}(a) + \frac{\Delta(a)}{2} \\ &< \mu_*^{(\text{on})} \leq \min\left\{\text{UCB}_t(a_*), \text{UCB}_t^{\text{S}}(a_*)\right\},\end{aligned}$$

thus $A_t \neq a$ with certainty. In what follows, we focus on **Case 1a**, the main case when MIN-UCB incorporates the offline dataset in learning. **Case 1a** and $V(a)$ being a valid bias bound imply

$$\frac{\omega(a)}{\Delta(a)} \in [0,1). \tag{15}$$

Then

$$\begin{aligned}&\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a) + T_{\text{S}}(a)}} \leq \sqrt{\frac{2\log(4Kt^4)}{T_{\text{S}}(a)}} \\ &\overset{(\dagger)}{\leq} \left(1 - \frac{\omega(a)}{\Delta(a)}\right)\sqrt{\frac{2\log(4Kt^4)}{\frac{16\log(4Kt^4)}{\Delta(a)^2}}} = \left(1 - \frac{\omega(a)}{\Delta(a)}\right) \cdot \frac{\Delta(a)}{\sqrt{8}}.\end{aligned}$$

($\dagger$) is by the second condition in **Case 1a** and (15). Also,

$$\frac{T_{\text{S}}(a) \cdot \omega(a)}{N_t(a) + T_{\text{S}}(a)} \leq \frac{\omega(a)}{\Delta(a)} \cdot \Delta(a).$$

Consequently, conditioned on event $\mathcal{E}_t$,

$$\begin{aligned}&\text{UCB}_t^{\text{S}}(a) \\ &\leq \mu^{(\text{on})}(a) + 2\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a) + T_{\text{S}}(a)}} + \frac{T_{\text{S}}(a) \cdot \omega(a)}{T_{\text{S}}(a) + N_t(a)} \qquad (16) \\ &\leq \mu^{(\text{on})}(a) + \left(1 - \frac{\omega(a)}{\Delta(a)}\right) \cdot \frac{\Delta(a)}{\sqrt{2}} + \frac{\omega(a)}{\Delta(a)} \cdot \Delta(a) \\ &\overset{(\ddagger)}{<} \mu_*^{(\text{on})} \overset{(\P)}{\leq} \min\left\{\text{UCB}_t(a_*), \text{UCB}_t^{\text{S}}(a_*)\right\},\end{aligned}$$

($\ddagger$) is because $0 \leq 1 - \frac{\omega(a)}{\Delta(a)} < 1$ by (15). (¶, 16) are by the event $\mathcal{E}_t$. Altogether, $A_t \neq a$ with certainty. □

4.3. Instance Indepedent Regret Bounds

Now we analyze the instance independent bound. We set $\delta_t = \delta/(2Kt^2)$, where $\delta \in (0,1)$ is a prescribed confidence paramater.

Theorem 4.9. *Assume $2 \leq K \leq T$. Consider policy MIN-UCB, which inputs a valid bias bound V on the underlying instance and $\delta_t = \delta/(2Kt^2)$ for $t = 1, \ldots$, where $\delta \in (0,1)$. With probability at least $1-\delta$, $Reg_T(\text{MIN-UCB}, P) =$*

$$O\left(\min\left\{\sqrt{KT\log\frac{T}{\delta}}, \left(\sqrt{\frac{\log\frac{T}{\delta}}{\tau_*}} + V_{max}\right) \cdot T\right\}\right), \tag{17}$$

where $V_{max} = \max_{a\in\mathcal{A}} V(a)$, and (τ_*, n_*) is an optimum solution of the following linear program:

$$\begin{aligned}\textit{(LP)}: \max_{\tau,n} \quad & \tau \\ \textit{s.t.} \quad & \tau \leq T_S(a) + n(a) \quad \forall a \in \mathcal{A}, \\ & \sum_{a\in\mathcal{A}} n(a) = T, \\ & \tau \geq 0, n(a) \geq 0 \quad \forall a \in \mathcal{A}.\end{aligned} \tag{18}$$

The proof of Theorem 4.9 is in Appendix B.3. In the minimum in (17), the first term corresponds to the regret

due to $\{\text{rad}_t(A_t)\}_{t=K+1}^T$, in the same vein as the vanilla UCB. The second term corresponds to the regret due to $\{\text{rad}_t^{\text{S}}(A_t)\}_{t=K+1}^T$ caused by the warm-start UCB^{S}. The minimum corresponds to the fact that MIN-UCB adapts to the better of the two UCB's.

We make two observations on τ_*. Firstly, $\tau_* \geq T/K$ for any $\{T_{\text{S}}(a)\}_{a\in\mathcal{A}}$. Indeed, the solution $(\bar{\tau}, \bar{n})$ defined as $\bar{\tau} = \bar{n}(a) = T/K$ for all $a \in \mathcal{A}$ is feasible to (LP). Crucially, we have $\sqrt{KT} \geq T/\sqrt{\tau_*}$, which implies $\text{Reg}_T(\text{MIN-UCB}, P) = o(\sqrt{KT\log(T/\delta)})$ when $V_{\text{max}} = o(\sqrt{K/T})$, and $\text{Reg}_T(\text{MIN-UCB}, P) = O(\sqrt{KT\log(T/\delta)})$ otherwise. Secondly, in the case when $T_{\text{S}}(a) = m$ for all $a \in \mathcal{A}$, it can be verified that $\tau_* = (T/K) + m$ is the optimum to (LP), with $n_*(a) = T/K$ for all a, meaning that $T/\sqrt{\tau_*} = T/\sqrt{(T/K)+m} \leq \sqrt{KT}$ for any $m \in \mathbb{Z}_{\geq 0}$.

The second term in (17) requires a new analysis, as we need to determine the worst-case regret under the heterogeneity in the sample sizes $\{T_{\text{S}}(a)\}_{a\in\mathcal{A}}$. We overcome the heterogeneity by a novel linear program (LP) in (18). The constraints in (LP) are set such that $\{n_*(a)\}_{a\in\mathcal{A}}$ represents a worst-case $\{\mathbb{E}[N_T(a)]\}_{a\in\mathcal{A}}$ that maximizes the instance-independent regret in the case of $P^{(\text{on})} = P^{(\text{off})}$. We incorporate the auxiliary decision variable τ in (LP), where a non-anticipatory incurs at most $O(1/\sqrt{\tau_*})$ regret per time round under the worst case number of arm pulls.

Despite the non-closed-form nature of τ_*, we demonstrate that τ_* is fundamental to a tight regret bound by the following regret lower bound, which matches (17) up to a logarithmic factor.

Theorem 4.10. *Let $V_{max} \in \mathbb{R}_{\geq 0}$ and $\{T_S(a)\}_{a\in\mathcal{A}} \in \mathbb{N}^{\mathcal{A}}$ be fixed but arbitrary, and let there be $K \geq 2$ arms. Set $V(a) = V_{max}$ for all $a \in \mathcal{K}$. For any non-anticipatory policy π, there exists a Gaussian instance $I \in \mathcal{I}_V$ with offline sample size $\{T_S(a)\}_{a\in\mathcal{A}}$, such that $\mathbb{E}[Reg_T(\pi, P)] =$*

$$\Omega\left(\min\left\{\sqrt{KT}, \left(\frac{1}{\sqrt{\tau_*}} + V_{max}\right)\cdot T\right\}\right),$$

where τ_* is the optimum of (18).

Theorem 4.10 is proved in Appendix C.3. We conclude the Section by highlighting that, in the case of $P^{(\text{off})} = P^{(\text{on})}$ no existing work establishes a problem independent regret bound of $o(\sqrt{KT\log(T/\delta)})$. To this end, by setting $V(a) = 0$ for all a, our analysis implies that MIN-UCB achieves a regret of $O(T\sqrt{\log(T/\delta)/\tau_*})$ and there is regret lower bound of $\Omega(T/\sqrt{\tau_*})$, showing the near-optimality of MIN-UCB in the well-aligned setting.

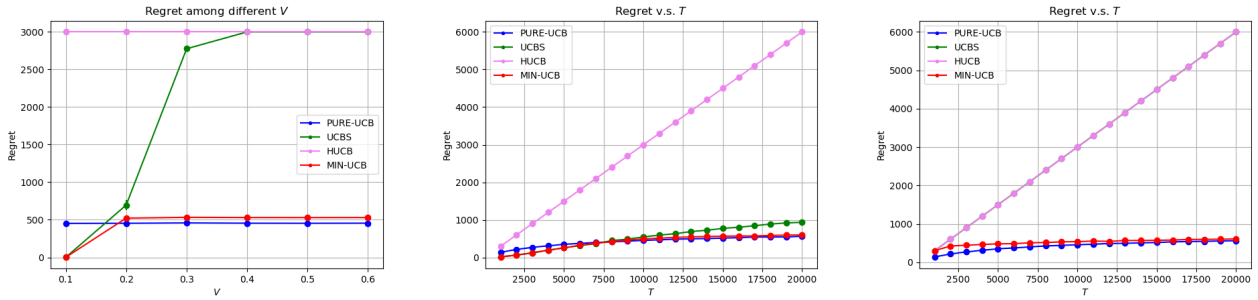
5. Numerical Experiments

We perform numerical experiments to evaluate MIN-UCB. We compare MIN-UCB with three related benchmarks. The first policy is vanilla UCB (Auer et al., 2002) that ignores the offline data, which we call "PURE-UCB". The second policy chooses an arm with the largest $\text{UCB}_t^{\text{S}}(a)$ defined in (8), which we call "UCBS". The third policy is HUCB1 (Shivaswamy & Joachims, 2012), which we call "HUCB".

In all experiments, we simulate our algorithm MIN-UCB and benchmarks on several instances, with fixed $K = 4$. Each arms's offline and online reward follows standard Gaussian distribution with mean $\mu^{\text{off}}(a)$, $\mu^{\text{on}}(a)$, respectively. In each specific instance, we run each algorithms 100 times and report their average and standard error. For simplicity, we assume $V(a) = V$ for any sub-optimal arm a and $T_{\text{S}}(a) = T_{\text{S}}$ for any arm a. In all experiments, we set $\boldsymbol{\mu}^{(\text{on})} = (0.8, 0.5, 0.5, 0.5)$ (hence $a_* = 1$). In Figure 1, we set $\boldsymbol{\mu}^{(\text{off})} = (0.4, 0.6, 0.6, 0.6)$, $V(1) = 0.4$, and vary V from 0.1 to 0.6. In Figure 2, we set $\boldsymbol{\mu}^{(\text{off})} = (0.5, 0.6, 0.6, 0.6)$, $V(1) = 0.3$, and vary T_S from 1000 to 3000, for $V = 0.1, 0.3$, respectively.

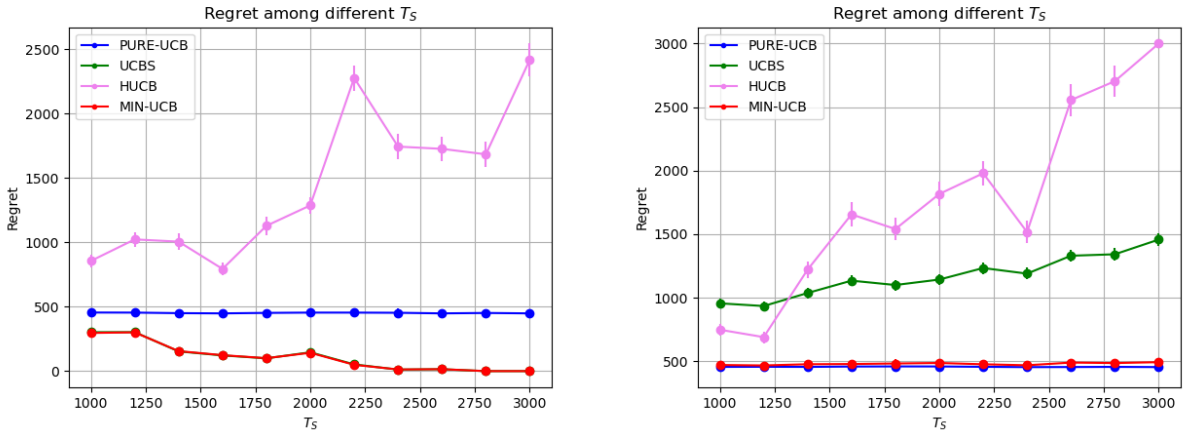
Figure 1a plots the regret on single instance under different valid bias bounds V, with $T = T_{\text{S}} = 10000$ in each trial. Specifically, for $V = 0.2, 0.6$, we illustrate the regret across different T in Figures 1b, 1c. These results demonstrate the benefit of our algorithm. When V is small, indicating that DM has a precise estimation on $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$, our MIN-UCB and UCBS that effectively leverage the information from V and offline dataset perform well. In contrary, when V is large, implying a loose estimation of $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$, the offline dataset becomes challenging to utilize effectively. In such cases, MIN-UCB's performance aligns with PURE-UCB, and remains unaffected by looseness of V that misleads the DM to believe $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$ to be much larger than the actual gap. The stable performance of MIN-UCB is in contrast with the performance of UCBS and HUCB. Thus, the robustness of our algorithm is validated through these numerical results.

Figures 2a and 2b show the regret on different T_{S}, where in the first group $V = 0.1$ (small V) and the second group $V = 0.3$ (large V). Both figures demonstrate our MIN-UCB consistently outperforms alternative methods, irrespective of T_{S} and V. A noteworthy observation is that is that when the offline data is easy to leverage (small V), a larger offline dataset (increased T_{S}) results in larger improvement of our algorithm over vanilla UCB. This observation aligns with intuition, as a larger T_{S} coupled with a smaller V indicates more informative insights into the online distribution. Conversely, when V is large and offline data is hard to utilize, increasing the size of the offline dataset (T_{S}) does not significantly impact the performance of MIN-UCB. This is because our algorithm can adapt to this environment and



(a) Different V (b) $V = 0.2$ (c) $V = 0.6$

Figure 1. Regret on different V



(a) $V = 0.1$ (b) $V = 0.3$

Figure 2. Regret on different T_S

consistently align with vanilla UCB, thereby mitigating the influence of the offline data on performance.

6. Conclusion

We explore a multi-armed bandit problem with additional offline data. Our approach on how to adapt to the possibly biased offline data is novel in the literature, leading to tight and optimal regret bounds. There are some interesting future directions. For example, we can further investigate how to apply it in other online learning models, such as linear bandits and online Markov decision processes. It is also intriguing to study how other forms of metrics that can measure the distribution drift can affect and facilitate the algorithm design.

Acknowledgements

Wang Chi would like to acknowledge the support from the Singapore Ministry of Education AcRF Tier 2 Grant (Grant number: T2EP20121-0035). This research work is partially funded by Institute of Operations Research and Analytics, National University of Singapore. Lixing would like to acknowledge the support from NUS Research Scholarship.

Impact Statement

This paper presents work whose goal is to advance the field of Online Learning. We primarily focus on theoretical research. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Ma-*

chine Learning Research, pp. 12–38. PMLR, 2017.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002. ISSN 0885-6125. URL https://doi.org/10.1023/A:1013689704352.

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

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, may 2010. ISSN 0885-6125.

Besbes, O., Ma, W., and Mouchtaki, O. Beyond iid: data-driven decision-making in heterogeneous environments. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23979–23991. Curran Associates, Inc., 2022.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Bouneffouf, D., Parthasarathy, S., Samulowitz, H., and Wistuba, M. Optimal exploitation of clustering and history information in multi-armed bandit. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2016–2022, 2019.

Bu, J., Simchi-Levi, D., and Xu, Y. Online pricing with offline data: Phase transition and inverse square law. In *International Conference on Machine Learning*, pp. 1202–1210. PMLR, 2020.

Chen, X., Shi, P., and Pu, S. Data-pooling reinforcement learning for personalized healthcare intervention. *arXiv preprint arXiv:2211.08998*, 2022.

Crammer, K., Kearns, M., and Wortman, J. Learning from multiple sources. *Journal of Machine Learning Research*, 9(57):1757–1774, 2008. URL http://jmlr.org/papers/v9/crammer08a.html.

Cutkosky, A., Dann, C., Das, A., Gentile, C., Pacchiano, A., and Purohit, M. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pp. 2276–2285. PMLR, 2021.

Cutkosky, A., Dann, C., Das, A., and Zhang, Q. Leveraging initial hints for free in stochastic linear bandits. In Dasgupta, S. and Haghtalab, N. (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 282–318. PMLR, 29 Mar–01 Apr 2022.

Gur, Y. and Momeni, A. Adaptive sequential experiments with unknown information arrival processes. *Manufacturing & Service Operations Management*, 24(5):2666–2684, 2022.

Hao, B., Jain, R., Lattimore, T., Van Roy, B., and Wen, Z. Leveraging demonstrations to improve online learning: Quality matters. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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

Lee, J., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3340–3348. PMLR, 2021.

Liu, C.-Y. and Li, L. On the prior sensitivity of thompson sampling. In Ortner, R., Simon, H. U., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 321–336, Cham, 2016. Springer International Publishing.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Pacchiano, A., Phan, M., Abbasi Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvari, C. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33:10328–10337, 2020.

Pacchiano, A., Dann, C., and Gentile, C. Data-driven regret balancing for online model selection in bandits. *CoRR*, abs/2306.02869, 2023. URL https://doi.org/10.48550/arXiv.2306.02869.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *Conference on Learning Theory*, pp. 993–1019. PMLR, 2013a.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013b.

Russo, D. and Roy, B. V. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL http://jmlr.org/papers/v17/14-087.html.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

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

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. Distributionally robust batch contextual bandits. *Management Science*, 69 (10):5772–5793, 2023.

Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D. J., Lykouris, T., Dudik, M., and Schapire, R. E. Bayesian decision-making under misspecified priors with applications to meta-learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26382–26394. Curran Associates, Inc., 2021.

Steinhardt, J. and Liang, P. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International conference on machine learning*, pp. 1593–1601. PMLR, 2014.

Tennenholtz, G., Shalit, U., Mannor, S., and Efroni, Y. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pp. 430–439. PMLR, 2021.

Wagenmaker, A. and Pacchiano, A. Leveraging offline data in online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Wang, Z., Zhang, C., Singh, M. K., Riek, L. D., and Chaudhuri, K. Multitask bandit learning through heterogeneous feedback aggregation. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1531–1539. PMLR, 2021.

Wei, C. and Luo, H. More adaptive algorithms for adversarial bandits. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1263–1291. PMLR, 2018.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3583–3634. PMLR, 09–12 Jul 2020a.

Wei, C.-Y., Luo, H., and Agarwal, A. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pp. 3583–3634. PMLR, 2020b.

Ye, L., Lin, Y., Xie, H., and Lui, J. Combining offline causal inference and online bandit learning for data driven decision. *arXiv preprint arXiv:2001.05699*, 2020.

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

A. More Detailed Comparisons with Existing Works

A.1. Comparing with (Zhang et al., 2019)

(Zhang et al., 2019) consider a stochastic contextual K-armed bandit model with possibly biased offline data, which generalizes our setting of stochastic K-armed bandits. On one hand, their proposed algorithm ARROW-CB does not require any knowledge on the discrepancy between the offline and online data, while our proposed algorithm requires knowing an upper bound $V(a)$ to $|\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a)|$ for each a. On the other hand, in the case of stochastic K-armed bandits, the regret bound of ARROW-CB (see Theorem 1 in (Zhang et al., 2019)) is at least the regret bound of the explore-then-commit algorithm (see Chapter 6 in (Lattimore & Szepesvári, 2020)) that ignores all offline data, which in particular does not offer improved regret bound compared to existing baselines that do not utilize offline data.

In more details, the ARROW-CB algorithm inputs an exploration probability parameter $\epsilon \in (0, 1)$, and a set $\Lambda \subset [0, 1]$ of weighted combination parameters that hedges between ignoring all offline data and incorporating all offline data. In their contextual setting where they compare ARROW-CB with a benchmark policy class Π, they establish the following regret upper bound on ARROW-CB on an instance I:

$$\epsilon T + 3\sqrt{T \log(T|\Pi|)} + \frac{32}{\sqrt{\epsilon}} \cdot \sqrt{KT \log(8T|\Lambda|)} + \min_{\lambda \in \Lambda}\{\text{disc}(\lambda, I)\}. \tag{19}$$

The function $\text{disc}(\lambda, I)$ (see Equation (4) in (Zhang et al., 2019)) is a discrepancy measure on the distance between the offline and online models, with the property that $\text{disc}(\lambda, I) \geq 0$ for all λ, I.

On one hand, (19) provides improvements to the conventional regret bound $O(\sqrt{KT \log(|\Pi|T)})$ in terms of the dependency on $|\Pi|$ when $\text{disc}(\lambda, I)$ is sufficiently small. On the other hand, the (19) does not provide any improvement to the conventional regret bounds in the stochastic K-armed bandits setting. Indeed, in K-armed setting we have $|\Pi| = K$, where Π consists of the policies of always pulling arm a, for $a \in \mathcal{A}$. Ignoring the non-negative term $\min_{\lambda \in \Lambda}\{\text{disc}(\lambda, I)\}$ and optimizing ϵ in the remaining terms show that the remaining terms sum to at least

$$K^{1/3}T^{2/3} + 3\sqrt{T \log(TK)} + 32K^{1/3}T^{2/3}\sqrt{\log T},$$

which is no better than the explore-then-commit policy's regret of $O(K^{1/3}T^{2/3}\sqrt{\log T})$.

Finally, one additional distinction is that (Zhang et al., 2019) requires $T_{\text{S}}(1) = \ldots = T_{\text{S}}(K)$, while we allow the offline sample sizes $T_{\text{S}}(1), \ldots, T_{\text{S}}(K)$ to be arbitrary non-negative integers.

A.2. Comparing wtih Online Learning with Advice

The frameworks of (Rakhlin & Sridharan, 2013a; Steinhardt & Liang, 2014; Wei & Luo, 2018; Wei et al., 2020a) utilize a commonly defined quantity $\mathcal{E}$ to quantify the error in the predictions, which are provided in an online manner. We focus our discussion on comparison with (Wei & Luo, 2018). The framework of (Wei & Luo, 2018) provides the following results for adversarial multi-armed bandits with side information. Consider a K-armed bandit model, where an arm $a \in \mathcal{A}$ is associated with loss $\ell_t(a) \in [-1, 1]$ at time t for $t \in \{1, \ldots, T\}$, and the loss vectors $\{\ell_1, \ldots, \ell_T\}$ over the horizon of T time steps are generated by an oblivious adversary. Before choosing an arm $a_t \in \mathcal{A}$ at time t, the agent is endowed with a prediction $m_t \in [-1, 1]^K$, where $m_t(a)$ is a prediction on $\ell_t(a)$. They design a version of optimistic online mirror design algorithm that incoporates the predictions $m_1, \ldots, m_T$, and achieves an expected regret bound of

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a) - \min_{a \in K} \sum_{t=1}^{T} \ell_t(a)\right] = O(\sqrt{K\mathcal{E} \log(T)}), \tag{20}$$

where the expectation is taken over the internalized randomization of the algorithm, where an arm is randomly chosen each round. The quantity $\mathcal{E} = \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2 = O(T)$ represents the prediction errors.

We argue that their framework can only guarantees a $O(\sqrt{KT \log(T)})$ in terms of expected regret bound in our setting, since the prediction error term $\mathcal{E}$ depends on the *realized* losses rather than their means. Indeed, In our stochastic setting (translated to a loss minimization instead of reward maximization model), the loss vectors are $\ell_1, \ldots, \ell_T$ are independent and identically distributed according to a common (but latent) distribution $\mathcal{D}$. Consider the case when the loss of arm a is distributed according to the Bernoulli distribution with mean $\mu(a) \in [1/4, 3/4]$ for each $a \in \mathcal{A}$. Even in the case when we have the

prediction $m_t(a) = \mu(a)$ for each t, namely the predictions reveal the true mean, we still have $(\ell_t(a) - \mu(a))^2 \geq 1/16$ for each arm with certainty, leading to $\mathcal{E} \geq T/16$.

The framework of (Wei & Luo, 2018) can only gauranttee a regret bound of $o(\sqrt{KT})$ when $\mathcal{E} = o(T)$. In our setting, it requires that $\|\ell_t - m_t\|_\infty = o(1)$, meaning that the prediction m_t has to be correlated with, thus contains information of, the realized loss ℓ_t, for $\Omega(T)$ many time rounds t. Different from (Wei & Luo, 2018), we achieve a regret bound of $o(\sqrt{KT})$ in the presense of accurate offline data, instead of receiving hints about the realized online rewards. A similar conclusion (with a worse regret bound than (20)) holds when we compare our results with (Wei et al., 2020a), which presents regret bounds on adversarial contextual bandits with predictions, since their regret bounds are also defined in terms of $\mathcal{E}$ mentioned above.

A.3. Comparing with (Chen et al., 2022)

(Chen et al., 2022) explores a Finite-Horizon Reinforcement Learning Model incorporating possibly biased historical data, another generalization of our stochastic multi-armed bandit model. Similar to our work, their approach "Data-pooling Perturbed LSVI" (Algorithm 2) requires an upper bound on the discrepancy between offline and online data, defined as Δ in their Assumption 1. They propose an cross-validation method and illustrate through a case study that a substantial range of Δ values produce similar performance for their algorithm. However, their approach cannot facilitate the design of an algorithm that achieve our improved regret bounds compared to existing benchmarks that neglects the offline data.

In details, their "Data-pooling Perturbed LSVI" algorithm is based on classical Value Iteration (VI). They adapt the VI on online learning setting by randomly perturbing the estimated value-go-functions, and applying least squares fitting to estimate the Q-function. This approach is fundemantally different from the OFU principle we apply in the multi-armed bandit context. Besides, they follow a different way to combine offline and online data. Specifically, they compute a weighted parameter λ_t that balances between online and offline data, via t, Δ, and $T_{\mathrm{S}}^{\min} = \min_{h,s,a} T_{\mathrm{S}}(h,s,a)$, where $T_{\mathrm{S}}(h,s,a)$ refers to the number of offline samples for pair (h,s,a), refering to epoch h, state s, and action a, respectively. This λ_t is shown in their equation (10). They derive the following regret upper bound:

$$(1+2p_0)^{-1} HSK \sum_{t=1}^{T/(SK)} \left(\varepsilon_R^{\mathrm{DP}}(t, T_{\mathrm{S}}^{\min}) + H(\bar{W}+1)\varepsilon_P^{\mathrm{DP}}(t, T_{\mathrm{S}}^{\min}) + \varepsilon_V^{\mathrm{DP}}(t, T_{\mathrm{S}}^{\min})\right) + 4HT\delta \tag{21}$$

where p_0, $\bar{W}$, δ are their input parameter, H is the length of a horizon, and S is the number of state. $\varepsilon_R^{\mathrm{DP}}(\cdot,\cdot)$, $\varepsilon_P^{\mathrm{DP}}(\cdot,\cdot)$, $\varepsilon_V^{\mathrm{DP}}(\cdot,\cdot)$ are confidence radius defined in their Appendix EC.4.2. They state that when $T_{\mathrm{S}}^{\min} \geq 1$ and Δ is small (See their Theorem EC.1), this regret bound is strictly smaller than the case that without combining historical data.

However, (21) does not offer an explicit closed-form bound, and they do not provide any regret lower bound.More importantly, their implicit bound solely depends on $T_{\mathrm{S}}^{\min} = \min_{h,s,a} T_{\mathrm{S}}(h,s,a)$, which is equivalent to $\min_a T_{\mathrm{S}}(a)$ in our model. In contrast, both our instance dependent bound (Theorem 4.2) and instance independent bound (Theorem 4.9) demonstrate how the difference in $T_S(a)$ among different arms affects the regret. Thus, their approach cannot deliver the tight regret bounds on and insights into our model.

B. Proofs for Regret Upper Bounds

B.1. Proof for Lemma 4.1

The proof uses the Chernoff inequality:

Proposition B.1 (Chernoff Inequality)**.** *Let $G_1, \ldots, G_m$ be independent (though not necessarily identically distributed) 1-subGaussian random variables. For any $\delta \in (0,1)$, it holds that*

$$\Pr\left[\frac{1}{m}\mathbb{E}\left[\sum_{i=1}^m G_i\right] \leq \frac{1}{m}\sum_{i=1}^m G_i + \sqrt{\frac{2\log(2/\delta)}{m}}\right] \geq 1-\delta/2,$$
$$\Pr\left[\frac{1}{m}\sum_{i=1}^m G_i \leq \frac{1}{m}\mathbb{E}\left[\sum_{i=1}^m G_i\right] + \sqrt{\frac{2\log(2/\delta)}{m}}\right] \geq 1-\delta/2.$$

Proof of Lemma 4.1. To prove the Lemma, it suffices to prove that $\Pr[\mathcal{E}_t^{\mathrm{S}}(a)] \geq 1-\delta_t$. Indeed, the inequality implies

$\Pr(\mathcal{E}_t(a)) \geq 1-\delta_t$ by setting $T_{\mathrm{S}}(a)=0$, and then a union bound on the failure probabilities of $\mathcal{E}_t^{\mathrm{S}}(a), \mathcal{E}_t(a)$ for $a \in \mathcal{A}$ establishes the Lemma. Now, we have

$$
\begin{aligned}
&\Pr[\mu^{(\mathrm{on})}(a) \leq \mathrm{UCB}_t^{\mathrm{S}}(a)] \\
=&\Pr\left[\mu^{(\mathrm{on})}(a) \leq \frac{N_t(a)\cdot \hat{R}_t(a)+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{N_t(a)+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2t/\delta_t)}{N_t^+(a)+T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)}{N_t^+(a)+T_{\mathrm{S}}(a)}\cdot V(a).\right] \\
=&\Pr\left[\frac{N_t(a)\mu^{(\mathrm{on})}(a)+T_{\mathrm{S}}(a)\mu^{(\mathrm{off})}(a)}{N_t(a)+T_{\mathrm{S}}(a)} + \frac{T_{\mathrm{S}}(a)(\mu^{(\mathrm{on})}(a)-\mu^{(\mathrm{off})}(a))}{N_t(a)+T_{\mathrm{S}}(a)} \leq \right. \\
&\left. \frac{N_t(a)\cdot \hat{R}_t(a)+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{N_t(a)+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)+T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)}{N_t(a)+T_{\mathrm{S}}(a)}\cdot V(a).\right] \\
\geq&\Pr\left[\frac{N_t(a)\mu^{(\mathrm{on})}(a)+T_{\mathrm{S}}(a)\mu^{(\mathrm{off})}(a)}{N_t(a)+T_{\mathrm{S}}(a)} \leq \frac{N_t(a)\cdot \hat{R}_t(a)+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{N_t(a)+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)+T_{\mathrm{S}}(a)}}\right] && (22) \\
\geq&\Pr_{Y_i\sim P^{(\mathrm{on})}(a)}\left[\frac{n\mu^{(\mathrm{on})}(a)+T_{\mathrm{S}}(a)\mu^{(\mathrm{off})}(a)}{n+T_{\mathrm{S}}(a)} \leq \frac{\sum_{i=1}^n Y_i+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{n+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2t/\delta_t)}{n+T_{\mathrm{S}}(a)}} \text{ for } n=1,2,\ldots,t\right] && (23) \\
\geq&1-\delta_t/2. && (24)
\end{aligned}
$$

Step (22) is by the input assumption that $|\mu^{(\mathrm{on})}(a)-\mu^{(\mathrm{off})}(a)| \leq V(a)$. Step (23) is by a union bound over all possible values of $N_t(a)$. Step (24) is by the Chernoff inequality.

Next,

$$
\begin{aligned}
&\Pr\left(\mathrm{UCB}_t^{\mathrm{S}}(a) \leq \mu^{(\mathrm{on})}(a) + \mathrm{rad}_t^{\mathrm{S}}(a) + \sqrt{\frac{2\log(2N_t^+(a)/\delta_t)}{N_t^+(a)+T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)\cdot(\mu^{(\mathrm{off})}(a)-\mu^{(\mathrm{on})}(a))}{N_t^+(a)+T_{\mathrm{S}}(a)}\right) \\
=&\Pr\left(\frac{N_t(a)\cdot \hat{R}_t(a)+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{N_t(a)+T_{\mathrm{S}}(a)} \leq \mu^{(\mathrm{on})}(a) + \sqrt{\frac{2\log(2N_t^+(a)/\delta_t)}{N_t^+(a)+T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)\cdot(\mu^{(\mathrm{off})}(a)-\mu^{(\mathrm{on})}(a))}{N_t^+(a)+T_{\mathrm{S}}(a)}\right) \\
=&\Pr\left(\frac{N_t(a)\cdot \hat{R}_t(a)+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{N_t(a)+T_{\mathrm{S}}(a)} \leq \frac{N_t(a)\mu^{(\mathrm{on})}(a)+T_{\mathrm{S}}(a)\mu^{(\mathrm{off})}(a)}{N_t(a)+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2N_t^+(a)/\delta_t)}{N_t^+(a)+T_{\mathrm{S}}(a)}}\right) \\
\geq&\Pr_{Y_i\sim P^{(\mathrm{on})}(a)}\left[\frac{\sum_{i=1}^n Y_i+T_{\mathrm{S}}(a)\cdot \hat{X}(a)}{n+T_{\mathrm{S}}(a)} \leq \frac{n\mu^{(\mathrm{on})}(a)+T_{\mathrm{S}}(a)\mu^{(\mathrm{off})}(a)}{n+T_{\mathrm{S}}(a)} + \sqrt{\frac{2\log(2t/\delta_t)}{n+T_{\mathrm{S}}(a)}} \text{ for } n=1,2,\ldots,t\right] && (25) \\
\geq&1-\delta_t/2. && (26)
\end{aligned}
$$

The justifications of (25, 26) are the same as (23, 24) respectively. Altogether, the Lemma is proved.

□

B.2. Proof of instance dependent regret upper bound

B.2.1. Proof of Lemma 4.8

We consider two main cases depending on $\omega(a)$ and $\Delta(a)$:

Case 1: $\omega(a) < \Delta(a)$. The case condition implies that

$$
\max\left\{1-\frac{\omega(a)}{\Delta(a)}, 0\right\}^2 = \left(1-\frac{\omega(a)}{\Delta(a)}\right)^2 > 0, \quad \frac{\omega(a)}{\Delta(a)} \in [0,1).
$$

To proceed. we consider two sub-cases:

Sub-case 1a: $T_{\mathrm{S}}(a) \cdot (1 - \frac{\omega(a)}{\Delta(a)})^2 \geq \frac{16\log(4Kt^4)}{\Delta(a)^2}$**.** Then

$$\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)+T_{\mathrm{S}}(a)}} \leq \sqrt{\frac{2\log(4Kt^4)}{T_{\mathrm{S}}(a)}} \leq \left(1-\frac{\omega(a)}{\Delta(a)}\right)\sqrt{\frac{2\log(4Kt^4)}{\frac{16\log(4Kt^4)}{\Delta(a)^2}}} = \left(1-\frac{\omega(a)}{\Delta(a)}\right)\cdot\frac{\Delta(a)}{\sqrt{8}}. \tag{27}$$

The second inequality is by the lower bound $T_{\mathrm{S}}(a)$ in the case condition in **Sub-case 1a**. In addition,

$$\frac{T_{\mathrm{S}}(a)}{N_t(a)+T_{\mathrm{S}}(a)}\cdot\omega(a) \leq \frac{\omega(a)}{\Delta(a)}\cdot\Delta(a).$$

Consequently, conditioned on event $\mathcal{E}_t$,

$$\begin{aligned}
\mathrm{UCB}_t^{\mathrm{S}}(a) &\leq \mu^{(\mathrm{on})}(a) + 2\cdot\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)+T_{\mathrm{S}}(a)}} + \frac{T_{\mathrm{S}}(a)}{T_{\mathrm{S}}(a)+N_t(a)}\cdot\omega(a) && (28)\\
&\leq \mu^{(\mathrm{on})}(a) + \left(1-\frac{\omega(a)}{\Delta(a)}\right)\cdot\frac{\Delta(a)}{\sqrt{2}} + \frac{\omega(a)}{\Delta(a)}\cdot\Delta(a)\\
&< \mu_*^{(\mathrm{on})} \leq \min\left\{\mathrm{UCB}_t(a_*), \mathrm{UCB}_t^{\mathrm{S}}(a_*)\right\}, && (29)
\end{aligned}$$

The strict inequality in step (29) is by the sub-case condition that $1-\frac{\omega(a)}{\Delta(a)} > 0$. The less-than-equal steps in Lines (28, 29) is by conditioning on the event $\mathcal{E}_t$. Altogether, $A_t \neq a$ with certainty.

Sub-case 1b: $T_{\mathrm{S}}(a) \cdot (1 - \frac{\omega(a)}{\Delta(a)})^2 < \frac{16\log(4Kt^4)}{\Delta(a)^2}$**.** In this case, we clearly have

$$N_t(a) > 32\cdot\frac{\log(4Kt^4)}{\Delta(a)^2} - T_{\mathrm{S}}(a)\cdot\max\left\{1-\frac{\omega(a)}{\Delta(a)}, 0\right\}^2 \geq \frac{16\log(4Kt^4)}{\Delta(a)^2}.$$

Consequently, we have

$$\mathrm{rad}_t(a) = \sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)}} < \sqrt{\frac{2\log(4Kt^4)}{16\cdot\frac{\log(4Kt^4)}{\Delta(a)^2}}} \leq \frac{\Delta(a)}{\sqrt{8}},$$

leading to

$$\mathrm{UCB}_t(a) \leq \mu^{(\mathrm{on})}(a) + 2\cdot\frac{\Delta(a)}{\sqrt{8}} < \mu_*^{(\mathrm{on})} \leq \min\left\{\mathrm{UCB}_t(a_*), \mathrm{UCB}_t^{\mathrm{S}}(a_*)\right\},$$

thus $A_t \neq a$ with certainty.

Case 2: $\omega(a) \geq \Delta(a)$. Then $\max\{1-\frac{\omega(a)}{\Delta(a)}, 0\}^2 = 0$, and $N_t(a) > 32\cdot\frac{\log(4Kt^4)}{\Delta(a)^2}$. Similar to **Case 1b**, we arrive at

$$\mathrm{rad}_t(a) = \sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)}} < \sqrt{\frac{2\log(4Kt^4)}{32\cdot\frac{\log(4Kt^4)}{\Delta(a)^2}}} \leq \frac{\Delta(a)}{4}.$$

Consequently,

$$\mathrm{UCB}_t(a) \leq \mu^{(\mathrm{on})}(a) + 2\cdot\frac{\Delta(a)}{4} < \mu_*^{(\mathrm{on})} \leq \min\left\{\mathrm{UCB}_t(a_*), \mathrm{UCB}_t^{\mathrm{S}}(a_*)\right\},$$

where the first and the last inequalities are by the event $\mathcal{E}_t$. Altogether, $A_t \neq a$ with certainty.

Altogether, all cases are covered and the Lemma is proved.

B.2.2. Proof of Theorem 4.2

Applying Lemma 4.8, we get

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_T(\pi, T)] &= \sum_{t=1}^{T}\mathbb{E}\left[(\mu_*^{(\mathrm{on})} - \mu^{(\mathrm{on})}(A_t))\cdot[\mathbf{1}(\mathcal{E}_t) + \mathbf{1}(\mathcal{E}_t^c)]\right]\\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[(\mu_*^{(\mathrm{on})} - \mu^{(\mathrm{on})}(A_t))\cdot\mathbf{1}(\mathcal{E}_t)\right] + \Delta_{\max}\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}(\mathcal{E}_t^c)\right],
\end{aligned}$$

and

$$\sum_{t=1}^{T}\mathbb{E}\left[(\mu_*^{(\text{on})}-\mu^{(\text{on})}(A_t))\cdot\mathbf{1}(\mathcal{E}_t)\right]=\sum_{a\in\mathcal{A}}\Delta(a)\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}(A_t=a)\mathbf{1}(\mathcal{E}_t)\right]$$
$$\leq\sum_{a\in\mathcal{A}}\max\left\{32\cdot\frac{\log(4KT^4)}{\Delta(a)}-T_{\text{S}}(a)\cdot\Delta(a)\cdot\max\left\{1-\frac{\omega(a)}{\Delta(a)},0\right\}^2,\Delta(a)\right\},$$

and $\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}(\mathcal{E}_t^c)\right]\leq\pi^2/6$. Altogether, we arrive at

$$\frac{\pi^2}{6}\Delta_{\text{max}}+\sum_{a\in\mathcal{A}:\Delta(a)>0}\max\left\{32\cdot\frac{\log(4KT^4)}{\Delta(a)}-T_{\text{S}}(a)\cdot\Delta(a)\cdot\max\left\{1-\frac{\omega(a)}{\Delta(a)},0\right\}^2,\Delta(a)\right\},\tag{30}$$

where $\Delta_{\text{max}}=\max_a\Delta(a)$, which proves the Theorem.

B.3. Proof for instance independent regret upper bounds

B.3.1. Proof of Theorem 4.9

Conditioned on $\cap_{t=1}^{T}\mathcal{E}_t$, we have

$$\begin{aligned}
\text{Reg}&=\sum_{t=1}^{T}\left[\mu_*^{(\text{on})}-\mu^{(\text{on})}(A_t)\right]\\
&\leq K\Delta_{\text{max}}+\sum_{t=K+1}^{T}\left[\min\{\text{UCB}_t(a_*),\text{UCB}_t^{\text{S}}(a_*)\}-\mu^{(\text{on})}(A_t)\right]\\
&\leq K\Delta_{\text{max}}+\sum_{t=K+1}^{T}\left[\min\{\text{UCB}_t(A_t),\text{UCB}_t^{\text{S}}(A_t)\}-\mu^{(\text{on})}(A_t)\right]\\
&\leq K\Delta_{\text{max}}+\sum_{t=K+1}^{T}\left[\min\{\mu^{(\text{on})}(A_t)+2\text{rad}_t(A_t),\mu^{(\text{on})}(A_t)+2\text{rad}_t^{\text{S}}(A(t))\}-\mu^{(\text{on})}(A_t)\right]\qquad(31)\\
&\leq K\Delta_{\text{max}}+2\min\left\{\sum_{t=K+1}^{T}\text{rad}_t(A_t),\sum_{t=K+1}^{T}\text{rad}_t^{\text{S}}(A_t)\right\}.\qquad(32)
\end{aligned}$$

Step (31) is by applying Lemma 4.1, and noting that $\mu^{(\text{off})}(a)-\mu^{(\text{on})}(a)\leq V(a)$ in the upper bound of $\text{UCB}_t^{\text{S}}(a)$ in the Lemma.

The conventional analysis shows that $\sum_{t=K+1}^{T}\text{rad}_t(A_t)=O(\sqrt{KT\log(T/\delta)})$. We next focus on upper bounding

$$\sum_{t=K+1}^{T}\text{rad}_t^{\text{S}}(A_t)=\sum_{t=K+1}^{T}\left[\sqrt{\frac{2\log(2t/\delta_t)}{N_t(a)+T_{\text{S}}(a)}}+\frac{T_{\text{S}}(a)}{N_t(a)+T_{\text{S}}(a)}\cdot V(a)\right]$$
$$\leq\max_{a\in\mathcal{A}}V(a)\cdot T+\sum_{a\in\mathcal{A}}\sum_{n=1}^{N_T(a)}\sqrt{\frac{8\log(T/\delta)}{n+T_{\text{S}}(a)}}.$$

We claim that

$$\sum_{a\in\mathcal{A}}\sum_{n(a)=1}^{N_T(a)}\sqrt{\frac{1}{n(a)+T_{\text{S}}(a)}}\leq\sum_{a\in\mathcal{A}}\sum_{n(a)=1}^{\max\{\lceil\tau_*\rceil-T_{\text{S}}(a),0\}}\sqrt{\frac{1}{n(a)+T_{\text{S}}(a)}}.\tag{33}$$

We show (33) by establishing two major claims:

Claim 1: Suppose that $(N_T^*(a))_{a\in\mathcal{A}} \in \mathbb{N}_{\geq 0}^{\mathcal{A}}$ is an optimal solution to the following maximization problem

$$\text{(IP):} \quad \max_{(N_T(a))_{a\in\mathcal{A}}} \sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{N_T(a)} \sqrt{\frac{1}{n(a)+T_{\mathrm{S}}(a)}} \quad \text{s.t.} \sum_{a\in\mathcal{A}} N_T(a) = T, \quad N_T(a) \in \mathbb{N}_{\geq 0} \text{ for all } a \in \mathcal{A}.$$

Then it must be the case that $N_T^*(a) \leq \max\{\lceil \tau_* \rceil - T_{\mathrm{S}}(a), 0\}$ for all $a \in \mathcal{A}$, where we recall that τ_* is the optimum of (LP).

Claim 2: . For an optimal solution $(\tau_*, \{n_*(a)\}_{a\in\mathcal{A}})$ to (LP), it holds that

$$n_*(a) = \max\{\tau_* - T_{\mathrm{S}}(a), 0\} \text{ for every arm } a. \tag{34}$$

We see that **Claim 1** is immediately useful for proving (33), since then

$$\sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{N_T(a)} \sqrt{\frac{1}{n(a)+T_{\mathrm{S}}(a)}} \leq \sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{N_T^*(a)} \sqrt{\frac{1}{n(a)+T_{\mathrm{S}}(a)}} \leq \sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{\max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\}} \sqrt{\frac{1}{n(a)+T_{\mathrm{S}}(a)}},$$

where the first inequality is by the optimality of $(N_T^*(a))_{a\in\mathcal{A}}$ to (IP), and the fact that any realized $(N_T(a))_{a\in\mathcal{A}}$ must be feasible to (IP). The second inequality is a direct consequence of **Claim 1**. **Claim 2** is an auxiliary claim for proving **Claim 1**. We postpone the proofs of these two claims to the end of the proof of the Theorem.

Given (33), for each $a \in \mathcal{A}$ we then have

$$\sum_{n=1}^{\max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\}} \sqrt{\frac{1}{n+T_{\mathrm{S}}(a)}} \leq \frac{\max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\}}{\lceil\tau_*\rceil} \sum_{t=1}^{\lceil\tau_*\rceil} \sqrt{\frac{1}{t}} \leq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\} \cdot \frac{4}{\sqrt{\tau_*}}. \tag{35}$$

Applying the feasibility of (τ_*, n_*) to (LP), we have

$$\max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\} \leq \max\{\tau_* - T_{\mathrm{S}}(a), 0\} + 1 \leq n_*(a) + 1.$$

Combining the above with (35) gives

$$\sum_{a\in\mathcal{A}} \sum_{n=1}^{\max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\}} \sqrt{\frac{1}{n+T_{\mathrm{S}}(a)}} \leq \sum_{a\in\mathcal{A}} (n_*(a)+1) \cdot \frac{4}{\sqrt{\tau_*}} \leq T \cdot \frac{8}{\sqrt{\tau_*}},$$

hence $\sum_{a\in\mathcal{A}} \sum_{n=1}^{N_T(a)} \sqrt{\frac{8\log(T/\delta)}{n+T_{\mathrm{S}}(a)}} = O(T\sqrt{\log(T/\delta)/\tau_*})$. Altogether, the Theorem is proved after we provide the proofs to **Claims 1, 2**:

Proving Claim 1 We establish the claim by a contradiction argument. Suppose there exists an arm a such that $N_T^*(a) \geq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\} + 1$. Firstly, we assert that there must exist another arm $a' \in \mathcal{A} \setminus \{a\}$ such that $N_T^*(a') \leq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a'), 0\} - 1$. If not, we then have $N_T^*(a) \geq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a), 0\} + 1$ and $N_T^*(a'') \geq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a''), 0\}$ for all $a'' \in \mathcal{A}$, but then

$$\sum_{k\in\mathcal{A}} N_T^*(k) > \sum_{k\in\mathcal{A}} \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(k), 0\} \geq \sum_{k\in\mathcal{A}} \max\{\tau_* - T_{\mathrm{S}}(k), 0\} \overset{(\dagger)}{=} \sum_{k\in\mathcal{A}} n_*(k) = T,$$

violating the constraint $\sum_{k\in\mathcal{A}} N_T^*(k) = T$. Note that the equality $(\dagger)$ is by **Claim 2**. Thus, the claimed arm a' exists. To this end, the condition $N_T^*(a') \leq \max\{\lceil\tau_*\rceil - T_{\mathrm{S}}(a'), 0\} - 1$ implies that $\lceil\tau_*\rceil > T_{\mathrm{S}}(a')$, since otherwise $N_T^*(a') \leq -1$, which violates the non-negativity constraint on $N_T^*(a)$. Altogether we have established the existence of two distinct arms $a, a' \in \mathcal{A}$ such that

$$N_T^*(a) + T_{\mathrm{S}}(a) \geq \max\{\lceil\tau_*\rceil, T_{\mathrm{S}}(a)\} + 1, \text{ in particular } N_T^*(a) \geq 1,$$
$$N_T^*(a') + T_{\mathrm{S}}(a') \leq \max\{\lceil\tau_*\rceil, T_{\mathrm{S}}(a')\} - 1 = \lceil\tau_*\rceil - 1.$$

To establish the contradiction argument, consider another solution $(\tilde{N}_T(k))_{k\in\mathcal{A}}$ to the displayed optimization problem in the claim, where

$$\tilde{N}_T(k) = \begin{cases} N_T^*(k) - 1 & \text{if } k = a \\ N_T^*(k) + 1 & \text{if } k = a' \\ N_T^*(k) & \text{if } k \in \mathcal{A} \setminus \{a, a'\}. \end{cases}$$

By the property that $N_T^*(a) \geq 1$, $\tilde{N}_T(a) \geq 0$, and $(\tilde{N}_T(k))_{k\in\mathcal{A}}$ is a feasible solution. But then we have

$$\begin{aligned} &\sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{\tilde{N}_T(a)} \sqrt{\frac{1}{n(a) + T_{\mathrm{S}}(a)}} - \sum_{a\in\mathcal{A}} \sum_{n(a)=1}^{N_T^*(a)} \sqrt{\frac{1}{n(a) + T_{\mathrm{S}}(a)}} \\ =&\sqrt{\frac{1}{N_T^*(a') + T_{\mathrm{S}}(a') + 1}} - \sqrt{\frac{1}{N_T^*(a) + T_{\mathrm{S}}(a)}} \\ \geq&\sqrt{\frac{1}{\lceil \tau_* \rceil}} - \sqrt{\frac{1}{\max\{\lceil \tau_* \rceil, T_{\mathrm{S}}(a)\} + 1}} > 0, \end{aligned} \tag{36}$$

which contradicts the assumed optimality of $(N_T^*(a))_{a\in\mathcal{A}}$. Thus **Claim 1** is shown.

Proving Claim 2 . Now, since $(\tau_*, \{n_*(a)\}_{a\in\mathcal{A}})$ is feasible to the (LP), the constraints in the (LP) give us that $n_*(a) \geq \max\{\tau_* - T_{\mathrm{S}}(a), 0\}$. We establish our asserted equality by a contradition argument. Suppose there is an arm a' such that $n_*(a') - \max\{\tau_* - T_{\mathrm{S}}(a'), 0\} = \epsilon > 0$. Then it can be verified that the solution $(\tau', \{n'(a)\}_{a\in\mathcal{A}})$ defined as $\tau' = \tau_* + \epsilon/K$, $n'(a) = n_*(a) + \epsilon/K$ for all $a \in \mathcal{A} \setminus \{a'\}$ and $n'(a') = n_*(a) - \frac{K-1}{K}\epsilon$ is also a feasible solution to (LP). But then $\tau' > \tau_*$, which contradicts with the optimality of $(\tau_*, \{n_*(a)\}_{a\in\mathcal{A}})$. Thus (34) is shown.

C. Proofs for Regret Lower Bounds

C.1. Notational Set Up and Auxiliary Results

Notational Set Up. We recall the notational set-up in Section 3. Consider a non-anticipatory policy π and an instance I with reward distribution P. We denote $\rho_{P,\pi}$ as the joint probability distribution function on $S, A_1, R_1, \ldots, A_T, R_T$, the concatenation of the offline dataset S and the online trajectory under policy π on instance I. For a $\sigma(S, A_1, R_1, \ldots, A_T, R_T)$-measurable event E, we denote $\Pr_{P,\pi}(E)$ as the probability of E holds under $\rho_{P,\pi}$. For a $\sigma(S, A_1, R_1, \ldots, A_T, R_T)$-measurable random variable Y, we denote $\mathbb{E}_{P,\pi}[Y]$ as the expectation of Y under the joint probability distribution function $\rho_{P,\pi}$. In our analysis, we make use of Y being $\mathrm{Reg}_T(\pi, P)$ or $N_T(a)$ for an arm a. To lighten our notational burden, we abbreviate $\mathbb{E}_{P,\pi}[\mathrm{Reg}_T(\pi, P)]$ as $\mathbb{E}[\mathrm{Reg}_T(\pi, P)]$.

Auxiliary Results To establish the regret lower bounds and the impossibility result, we need three auxiliary results, namely Claim 1 and Theorems C.1, C.2. Claim 1 is on the KL-divergence between Gaussian random variables:

Claim 1. For $P_i = \mathcal{N}(\mu_i, \sigma^2)$, where $i \in \{1, 2\}$, we have $\mathrm{KL}(P_1, P_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

The following, dubbed Bretagnolle–Huber inequality, is extracted from Theorem 14.2 from (Lattimore & Szepesvári, 2020):

Theorem C.1. *Let $\mathbb{P}, \mathbb{Q}$ be probability disributions on $(\Omega, \mathcal{F})$. For an event $E \in \sigma(\mathcal{F})$, it holds that*

$$\Pr_{\mathbb{P}}(E) + \Pr_{\mathbb{Q}}(E^c) \geq \frac{1}{2} \exp\left(-KL(\mathbb{P}, \mathbb{Q})\right).$$

Lastly, the derivation of the chain rule in Theorem C.2 largely follows from the derivation of Lemma 15.1 in (Lattimore & Szepesvári, 2020):

Theorem C.2. *Consider two instances I_P, I_Q that share the same arm set $\mathcal{A}$, online phase horizon T, offline sample size $\{T_S(a)\}_{a\in\mathcal{A}}$, but have two different reward distributions $P = (P^{(on)}, P^{(off)})$, $Q = (Q^{(on)}, Q^{(off)})$. For any non-anticipatory policy π, it holds that*

$$KL(\rho_{P,\pi}, \rho_{Q,\pi}) = \sum_{a\in\mathcal{A}} \mathbb{E}_{P,\pi}[N_T(a)] \cdot KL(P_a^{(on)}, Q_a^{(on)}) + \sum_{a\in\mathcal{A}} T_S(a) \cdot KL(P_a^{(off)}, Q_a^{(off)}).$$

We provide a proof to theorem C.2 for completeness sake:

Proof of Theorem C.2. The proof largely follow the well-known chain rule in the multi-armed bandit literature, for example see Lemma 15.1 in (Lattimore & Szepesvári, 2020). We start by explicitly expressing the joint probability function $\rho_{P,\pi}$ on

$$S = \{\{X_s(a)\}_{s=1}^{T_{\mathrm{S}}(a)}\}_{a\in\mathcal{A}}, A_1, R_1, \ldots, A_T, R_T,$$

under reward distribution P and non-anticipatory policy π. In coherence with our focus on Gaussian instances, we only consider the case when all the random rewards (offline or online) are continuous random variables with support on $\mathbb{R}$. Generalizing the argument to general reward distributions only require notational changes. By an abuse in notation, we denote $P_a^{((\mathrm{off})}(x_s(a))$ as the probability density function (with variable $x_s(a)$) of the offline reward distribution with arm a, and likewise for the online reward distribution.

To ease the notation, we denote $x = \{x_s(a)\}_{s=1}^{T_{\mathrm{S}}(a)}$. Then $\rho_{P,\pi}$ is expressed as

$$\begin{aligned}
&\rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T) \\
&= \underbrace{\left[\prod_{a\in\mathcal{A}} \prod_{s=1}^{T_{\mathrm{S}}(a)} P_a^{(\mathrm{off})}(x_s(a))\right]}_{\text{on offline rewards}} \cdot \underbrace{\prod_{t=1}^{T} \left[\pi_t(a_t|x, a_1, r_1, \ldots, a_{t-1}, r_{t-1}) \cdot P_{a_t}^{(\mathrm{on})}(r_t)\right]}_{\text{on online arms and rewards}}.
\end{aligned} \tag{37}$$

Likewise, by replacing reward distribution P with Q but keeping the fixed policy π unchanged, we have

$$\begin{aligned}
&\rho_{Q,\pi}(x, a_1, r_1, \ldots, a_T, r_T) \\
&= \left[\prod_{a\in\mathcal{A}} \prod_{s=1}^{T_{\mathrm{S}}(a)} Q_a^{(\mathrm{off})}(x_s(a))\right] \cdot \prod_{t=1}^{T} \left[\pi_t(a_t|x, a_1, r_1, \ldots, a_{t-1}, r_{t-1}) \cdot Q_{a_t}^{(\mathrm{on})}(r_t)\right].
\end{aligned}$$

The KL divergence between $\rho_{P,\pi}$ and $\rho_{Q,\pi}$ is

$$\begin{aligned}
&\mathrm{KL}(\rho_{P,\pi}, \rho_{Q,\pi}) \\
&= \int_{x, r_1, \ldots, r_T} \sum_{a_1, \ldots, a_T \in \mathcal{A}} \rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T) \log\left[\frac{\rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T)}{\rho_{Q,\pi}(x, a_1, r_1, \ldots, a_T, r_T)}\right] \mathrm{d}x\, \mathrm{d}r_1 \ldots \mathrm{d}r_T,
\end{aligned}$$

where $\int_x = \int_{\{x_s(a)\}_{s=1}^{T_{\mathrm{S}}(a)} \in \mathbb{R}^{\sum_{a\in\mathcal{A}} T_{\mathrm{S}}(a)}}$, and $\mathrm{d}x = \prod_{a\in\mathcal{A}} \prod_{s=1}^{T_{\mathrm{S}}(a)} \mathrm{d}x_s(a)$.

We use the explicit expressions of $\rho_{P,\pi}, \rho_{Q,\pi}$ to decompose the log term:

$$\begin{aligned}
&\log\left[\frac{\rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T)}{\rho_{Q,\pi}(x, a_1, r_1, \ldots, a_T, r_T)}\right] \\
&= \sum_{a\in\mathcal{A}} \sum_{s=1}^{T_{\mathrm{S}}(a)} \log\left(\frac{P_a^{(\mathrm{off})}(x_s(a))}{Q_a^{(\mathrm{off})}(x_s(a))}\right) \\
&\quad + \sum_{t=1}^{T} \underbrace{\pi_t(a_t|x, a_1, r_1, \ldots, a_{t-1}, r_{t-1}) \log\left(\frac{\pi_t(a_t|x, a_1, r_1, \ldots, a_{t-1}, r_{t-1})}{\pi_t(a_t|x, a_1, r_1, \ldots, a_{t-1}, r_{t-1})}\right)}_{=0} + \sum_{t=1}^{T} \log\left(\frac{P_{a_t}^{(\mathrm{on})}(r_t)}{Q_{a_t}^{(\mathrm{on})}(r_t)}\right) \\
&= \sum_{a\in\mathcal{A}} \sum_{s=1}^{T_{\mathrm{S}}(a)} \log\left(\frac{P_a^{(\mathrm{off})}(x_s(a))}{Q_a^{(\mathrm{off})}(x_s(a))}\right) + \sum_{t=1}^{T} \log\left(\frac{P_{a_t}^{(\mathrm{on})}(r_t)}{Q_{a_t}^{(\mathrm{on})}(r_t)}\right).
\end{aligned}$$

In the second line, the middle sum is equal to zero, which can be interpreted as the fact that we are evaluating the same

policy π on reward distributions P, Q. Finally, marginalizing and making use of (37) gives, for each $a \in \mathcal{A}$,

$$\int_{x,r_1,\ldots,r_T} \sum_{a_1,\ldots,a_T \in \mathcal{A}} \rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T) \log\left(\frac{P_a^{(\text{off})}(x_s(a))}{Q_a^{(\text{off})}(x_s(a))}\right) \mathrm{d}x\, \mathrm{d}r_1 \ldots \mathrm{d}r_T$$
$$= \int_{x_s(a)} P_a^{(\text{off})}(x_s(a)) \log\left(\frac{P_a^{(\text{off})}(x_s(a))}{Q_a^{(\text{off})}(x_s(a))}\right) \mathrm{d}x_x(a) = \mathrm{KL}(P_a^{(\text{off})}, Q_a^{(\text{off})}), \tag{38}$$

where the first equality in (38) follows by integrating with respect to r_t over $\mathbb{R}$ and then summing over $a_t \in \mathcal{A}$ in the order of $t = T, \ldots, 1$, and then integrating over all variables in x except $x_s(a)$. For each $t \in \{1, \ldots, T\}$ during the online phase we have

$$\int_{x,r_1,\ldots,r_T} \sum_{a_1,\ldots,a_T \in \mathcal{A}} \rho_{P,\pi}(x, a_1, r_1, \ldots, a_T, r_T) \log\left(\frac{P_{a_t}^{(\text{on})}(r_t)}{Q_{a_t}^{(\text{on})}(r_t)}\right) \mathrm{d}x\, \mathrm{d}r_1 \ldots \mathrm{d}r_T$$
$$= \int_{x,r_1,\ldots,r_t} \sum_{a_1,\ldots,a_t \in \mathcal{A}} \left[\prod_{a \in \mathcal{A}} \prod_{s=1}^{T_{\text{S}}(a)} P_a^{(\text{off})}(x_s(a))\right] \cdot \prod_{\tau=1}^{t} \left[\pi_\tau(a_\tau | x, a_1, r_1, \ldots, a_{\tau-1}, r_{\tau-1}) \cdot P_{a_\tau}^{(\text{on})}(r_\tau)\right]$$
$$\log\left(\frac{P_{a_t}^{(\text{on})}(r_t)}{Q_{a_t}^{(\text{on})}(r_t)}\right) \mathrm{d}x\, \mathrm{d}r_1 \ldots \mathrm{d}r_t$$
$$= \sum_{a_t \in \mathcal{A}} \left\{\int_{x,r_1,\ldots,r_{t-1}} \sum_{a_1,\ldots,a_{t-1} \in \mathcal{A}} \left[\prod_{a \in \mathcal{A}} \prod_{s=1}^{T_{\text{S}}(a)} P_a^{(\text{off})}(x_s(a))\right] \cdot \prod_{\tau=1}^{t-1} \left[\pi_\tau(a_\tau | x, a_1, r_1, \ldots, a_{\tau-1}, r_{\tau-1}) \cdot P_{a_\tau}^{(\text{on})}(r_\tau)\right]\right.$$
$$\left. \times\, \pi_t(a_t | x, a_1, r_1, \ldots, a_{t-1}, r_{t-1})\, \mathrm{d}x\, \mathrm{d}r_1 \ldots \mathrm{d}r_{t-1}\right\} \cdot \int_{r_t} P_{a_t}^{(\text{on})}(r_t) \log\left(\frac{P_{a_t}^{(\text{on})}(r_t)}{Q_{a_t}^{(\text{on})}(r_t)}\right) \mathrm{d}r_t$$
$$= \sum_{a_t \in \mathcal{A}} \Pr_{P,\pi}(A_t = a_t) \int_{r_t} P_{a_t}^{(\text{on})}(r_t) \log\left(\frac{P_{a_t}^{(\text{on})}(r_t)}{Q_{a_t}^{(\text{on})}(r_t)}\right) \mathrm{d}r_t$$
$$= \sum_{a \in \mathcal{A}} \mathbb{E}_{P,\pi}[\mathbf{1}(A_t = a)] \cdot \mathrm{KL}(P_a^{(\text{on})}, Q_a^{(\text{on})}). \tag{39}$$

Summing (38) over $s \in \{1, \ldots, T_{\text{S}}(a)\}$ and $a \in \mathcal{A}$, as well as summing (39) over $t \in \{1, \ldots, T\}$ establish the Theorem. □

C.2. Proof for instance dependent regret lower bound

C.2.1. Proof for Theorem 4.7

We denote $\mu^{(\text{off})}(a), \mu^{(\text{on})}(a)$ as the means of the Gaussian distributions $P_a^{(\text{off})}, P_a^{(\text{on})}$ respestively, for each arm $a \in \mathcal{A}$. In addition, recall the notation that $\mu_*^{(\text{on})} = \max_{a \in \mathcal{A}} \mu^{(\text{on})}(a)$, $\Delta(a) = \mu_*^{(\text{on})} - \mu^{(\text{on})}(a)$ and $\omega(a) = V(a) + (\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a))$.

Consider an arbitrary but fixed arm a with $\Delta(a) > 0$. We claim that

$$\mathbb{E}_{P,\pi}[N_T(a)] \geq \frac{2(1-p)}{(1+\epsilon)^2} \cdot \frac{\log T}{\Delta(a)^2} + \frac{2}{(1+\epsilon)^2 \Delta(a)^2} \cdot \log\left(\frac{\epsilon \Delta(a)}{8C}\right)$$
$$- T_{\text{S}}(a) \cdot \max\left\{\left(1 - \frac{\omega(a)}{(1+\epsilon)\Delta(a)}\right), 0\right\}^2 \tag{40}$$

The Lemma then follows by applying (40) on each sub-optimal arm, thus it remains to prove (40). To proceed, we consider another Gaussain instance $Q = (Q^{(\text{off})}, Q^{(\text{on})})$, where

$$Q_k^{(\text{off})} = P_k^{(\text{off})} \text{ for all } k \in \mathcal{A} \setminus \{a\},$$

but

$$Q_a^{(\text{on})} = \mathcal{N}(\mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a), 1),$$

$$Q_a^{(\text{off})} = \begin{cases} \mathcal{N}(\mu^{(\text{off})}(a), 1) & \text{if } \mu^{(\text{off})}(a) \geq \mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a) - V(a) \\ \mathcal{N}(\mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a) - V(a), 1) & \text{if } \mu^{(\text{off})}(a) < \mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a) - V(a) \end{cases}. \tag{41}$$

It is clear that $P \in \mathcal{I}_V$ implies $Q \in \mathcal{I}_V$. Indeed, if the second case in (41) holds, then the mean of $Q_a^{(\text{on})}-$ the mean of $Q_a^{(\text{off})}$ is equal to $V(a)$, while if the first case hold, then we evidently have

$$\underbrace{\mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a)}_{\text{mean of } Q_a^{(\text{on})}} + V(a) \geq \mu^{(\text{off})}(a) \geq \underbrace{\mu^{(\text{on})}(a) + (1+\epsilon)\Delta(a)}_{\text{mean of } Q_a^{(\text{on})}} - V(a),$$

thus $Q \in \mathcal{I}_V$. Consider the event $E = \{N_T(a) \geq T/2\}$. By the assumed consistency of π, we have

$$\begin{aligned} 2CT^p &\geq \mathbb{E}[\text{Reg}_T(\pi, P)] + \mathbb{E}[\text{Reg}_T(\pi, Q)] && (42) \\ &\geq \frac{T}{2} \cdot \epsilon\Delta(a) \cdot \left[\Pr_{P,\pi}(E) + \Pr_{Q,\pi}(E^c)\right] && (43) \\ &\geq \frac{T}{4} \cdot \epsilon\Delta(a) \cdot \exp\left[-\mathbb{E}_{\pi,P}[N_T(a)] \cdot \text{KL}(P_a^{(\text{on})}, Q_a^{(\text{on})}) - T_S(a)\text{KL}(P_a^{(\text{off})}, Q_a^{(\text{off})})\right]. && (44) \end{aligned}$$

Step (42) is by the definition of consistency. Step (43) is implied by our construction that in instance P, arm a is sub-optimal with optiamlity gap $\Delta(a) \geq \epsilon\Delta(a)$, and in instance Q, arm a is the unique optimal arm and other arms have optimality gap at least $\epsilon\Delta(a)$. Step (44) is by the Chain rule Theorem C.2, as well as the BH inequality. By our set up, we have

$$\begin{aligned} \text{KL}(P_a^{(\text{on})}, Q_a^{(\text{on})}) &= \frac{(1+\epsilon)^2\Delta(a)^2}{2}, \\ \text{KL}(P_a^{(\text{off})}, Q_a^{(\text{off})}) &= \frac{\max\{(1+\epsilon)\Delta(a) - [V(a) + (\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a))], 0\}^2}{2}. \end{aligned}$$

Plugging in we get

$$\begin{aligned} 2CT^p \geq \frac{T}{4} \cdot \epsilon\Delta(a) \cdot \exp\Bigg[&-\mathbb{E}_{\pi,P}[N_T(a)] \cdot \frac{(1+\epsilon)^2\Delta(a)^2}{2} \\ &-T_S(a)\frac{\max\{(1+\epsilon)\Delta(a) - [V(a) + (\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a))], 0\}^2}{2}\Bigg], \end{aligned} \tag{45}$$

which is equivalent to

$$\begin{aligned} &\mathbb{E}_{\pi,P}[N_T(a)] \cdot \frac{(1+\epsilon)^2\Delta(a)^2}{2} + T_S(a)\frac{\max\{(1+\epsilon)\Delta(a) - [V(a) + (\mu^{(\text{off})}(a) - \mu^{(\text{on})}(a))], 0\}^2}{2} \\ \geq &(1-p)\log T + \log\left(\frac{\epsilon\Delta(a)}{8C}\right). \end{aligned}$$

Rearranging leads to the claimed inequality (40), and the Lemma is proved.

C.3. Proof for instance independent regret lower bound

C.3.1. Proof for Theorem 4.10

We proceed by a case analysis:

Case 1a: $2\sqrt{KT} > T \cdot (V_{\max} + 1/\sqrt{\tau_*})$, and $V_{\max} \leq 1/\sqrt{\tau_*}$. We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\max}\right)\right\}\right) = \Omega\left(\frac{T}{\sqrt{\tau_*}}\right).$$

Case 1b: $2\sqrt{KT} > T \cdot (V_{\max} + 1/\sqrt{\tau_*})$, and $V_{\max} > 1/\sqrt{\tau_*}$. We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\max}\right)\right\}\right) = \Omega(T \cdot V_{\max}).$$

Case 2: $2\sqrt{KT} \leq (V_{\textbf{max}} + 1/\sqrt{\tau_*}) \cdot T$**.** We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\text{max}}\right)\right\}\right) = \Omega(\sqrt{KT}).$$

We establish the three cases in what follows.

Case 1a: $2\sqrt{KT} > T \cdot (V_{\textbf{max}} + 1/\sqrt{\tau_*})$**, and** $V_{\textbf{max}} \leq 1/\sqrt{\tau_*}$**.** We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\text{max}}\right)\right\}\right) = \Omega\left(\frac{T}{\sqrt{\tau_*}}\right).$$

Now, without loss of generality, we assume that $1 \in \text{argmax}_{a\in\mathcal{A}} T_{\text{S}}(a)$. Consider the following Gaussian reward distributions $P = (P^{(\text{on})}, P^{(\text{off})})$ with $P_a^{(\text{on})} = P_a^{(\text{off})}$ (thus $P \in \mathcal{I}_V$ for any $V \in \mathbb{R}^K_{\geq 0}$), defined as

$$P_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1 \\ \mathcal{N}(0, 1) & \text{if } a \in \mathcal{A} \setminus \{1\} \end{cases}, \text{ where } \Delta = \frac{1}{\sqrt{\tau_*}}.$$

Consider the values $(\tilde{\tau}, \tilde{n})$ defined as $\tilde{n}(a) = \mathbb{E}_{P,\pi}[N_T(a)]$ for each $a \in \mathcal{A}$, and $\tilde{\tau} = \min_{a\in\mathcal{A}}\{T_{\text{S}}(a) + \tilde{n}(a)\}$. Evidently, $(\tilde{\tau}, \tilde{n})$ is feasible to (LP), and thus we have $\tilde{\tau} \leq \tau_*$, the optimum of (LP). In particular, there exists an arm $a \in \mathcal{A}$ such that $T_{\text{S}}(a) + \mathbb{E}_{P,\pi}[N_T(a)] \leq \tau_*$. To this end, let's consider two situations:

Situation (i): $T_{\text{S}}(a)+\mathbb{E}_{P,\pi}[N_T(a)] > \tau_*$ **for all** $a \in \mathcal{A}\setminus\{1\}$**.** The condition immeidately implies $T_{\text{S}}(1)+\mathbb{E}_{P,\pi}[N_T(1)] \leq \tau_*$. Then we can deduce that $T_{\text{S}}(1) + \mathbb{E}_{P,\pi}[N_T(1)] < T_{\text{S}}(a) + \mathbb{E}_{P,\pi}[N_T(a)]$ for all $a \in \mathcal{A} \setminus \{1\}$, which further implies that $\mathbb{E}_{P,\pi}[N_T(1)] < \mathbb{E}_{P,\pi}[N_T(a)]$ for all $a \in \mathcal{A} \setminus \{1\}$ since $T_{\text{S}}(1) = \max_{a\in\mathcal{A}} T_{\text{S}}(a)$. The above implies that $\mathbb{E}_{P,\pi}[N_T(1)] < T/K$, which implies that

$$\mathbb{E}[\text{Reg}_T(\pi, P)] > \frac{(K-1)T}{K}.$$

Situation (ii): $T_{\text{S}}(k) + \mathbb{E}_{P,\pi}[N_T(k)] \leq \tau_*$ **for a** $k \in \mathcal{A} \setminus \{1\}$**.** In this case, consider the following Gaussian reward distribution $Q = (Q^{(\text{on})}, Q^{(\text{off})})$ with $Q^{(\text{on})} = Q^{(\text{off})}$, defined as

$$Q_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1 \\ \mathcal{N}(2\Delta, 1) & \text{if } a = k \\ \mathcal{N}(0, 1) & \text{if } a \in \mathcal{A} \setminus \{1, k\} \end{cases}, \text{ where } \Delta = \frac{1}{\sqrt{\tau_*}}.$$

To this end, note that $P_a^{(\text{off})} = Q_a^{(\text{off})} = P_a^{(\text{on})} = Q_a^{(\text{on})}$ for all $a \in \mathcal{A} \setminus \{k\}$, but $P_k^{(\text{on})} = P_k^{(\text{off})} \neq Q_k^{(\text{on})} = Q_k^{(\text{off})}$. In addition, both P, Q belong to $\mathcal{I}_V$. Consider the event $E = \{N_T(1) < T/2\}$. We have

$$\begin{aligned}
\mathbb{E}[\text{Reg}_T(\pi, P)] + \mathbb{E}[\text{Reg}_T(\pi, Q)] &\geq \frac{T}{2} \cdot \Delta \cdot \left[\Pr_{P,\pi}(E) + \Pr_{Q,\pi}(E^c)\right] && (46)\\
&\geq \frac{T}{4} \cdot \Delta \cdot \exp\left[-\mathbb{E}_{P,\pi}[N_T(k)] \cdot \text{KL}(P_k^{(\text{on})}, Q_k^{(\text{on})}) - T_{\text{S}}(k)\text{KL}(P_k^{(\text{off})}, Q_k^{(\text{off})})\right] && (47)\\
&= \frac{T}{4} \cdot \Delta \cdot \exp\left[-\left(T_{\text{S}}(k) + \mathbb{E}_{P,\pi}[N_T(k)]\right) \cdot \text{KL}(P_k^{(\text{on})}, Q_k^{(\text{on})})\right] \\
&= \frac{T}{4\sqrt{\tau_*}} \cdot \exp\left[-\left(T_{\text{S}}(k) + \mathbb{E}_{P,\pi}[N_T(k)]\right) \cdot \frac{1}{2\tau_*}\right] \\
&\geq \frac{T}{4\sqrt{\tau_*}} \cdot \exp\left[-\tau_* \cdot \frac{1}{2\tau_*}\right] = \frac{1}{4\sqrt{e}} \cdot \frac{T}{\sqrt{\tau_*}}. && (48)
\end{aligned}$$

Altogether, in **Case 1a**, we either have

$$\mathbb{E}[\text{Reg}_T(\pi, P)] > \frac{(K-1)T}{K} = \Omega\left(\frac{T}{\sqrt{\tau_*}}\right),$$

or

$$\max\{\mathbb{E}[\text{Reg}_T(\pi, P)], \mathbb{E}[\text{Reg}_T(\pi, Q)]\} \geq \frac{1}{8\sqrt{e}} \cdot \frac{T}{\sqrt{\tau_*}}.$$

Case 1b: $2\sqrt{KT} > T \cdot (V_{\max} + 1/\sqrt{\tau_*})$, and $V_{\max} > 1/\sqrt{\tau_*}$. We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\max}\right)\right\}\right) = \Omega(T \cdot V_{\max}).$$

Now, set $\Delta = V_{\max}/2$, and consider the Guassian instance with reward disribution P:

$$P_a^{(\text{off})} = N(0,1) \text{ for all } a \in \mathcal{A}, \text{ and } P_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1, \\ \mathcal{N}(0,1) & \text{if } a \in \mathcal{A} \setminus \{1\}. \end{cases}$$

Now, there exists an arm $a' \in \mathcal{A} \setminus \{1\}$ such that $\mathbb{E}_{P,\pi}[N_T(a')] \leq T/(K-1)$. Consider the Gaussian instance with

$$Q_a^{(\text{off})} = \mathcal{N}(0,1) \text{ for all } a \in \mathcal{A}, \text{ and } Q_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1, \\ \mathcal{N}(2\Delta, 1) & \text{if } a = a', \\ \mathcal{N}(0,1) & \text{if } a \in \mathcal{A} \setminus \{1, a'\}. \end{cases}$$

To this end, note that $P_a^{(\text{off})} = Q_a^{(\text{off})}$ for all $a \in \mathcal{A}$, and $P_a^{(\text{on})} = Q_a^{(\text{on})}$ for all $a \in \mathcal{A} \setminus \{a'\}$, but $P_{a'}^{(\text{on})} \neq Q_{a'}^{(\text{on})}$. In addition, both P, Q belong to $\mathcal{I}_V$. Consider the event $E = \{N_T(1) < T/2\}$. We have

$$\begin{aligned}
\mathbb{E}[\text{Reg}_T(\pi, P)] + \mathbb{E}[\text{Reg}_T(\pi, Q)] \geq & \frac{T}{2} \cdot \Delta \cdot \left[\Pr_{P,\pi}(E) + \Pr_{Q,\pi}(E^c)\right] && (49) \\
\geq & \frac{T}{4} \cdot \Delta \cdot \exp\left[-\mathbb{E}_{P,\pi}[N_T(a')] \cdot \text{KL}(P_{a'}^{(\text{on})}, Q_{a'}^{(\text{on})})\right] && (50) \\
\geq & \frac{T}{4} \cdot \Delta \cdot \exp\left[-\frac{T}{K-1} \cdot 2 \cdot \Delta^2\right] \\
= & \frac{T}{8} \cdot V_{\max} \cdot \exp\left[-\frac{T}{K-1} \cdot \frac{V_{\max}^2}{2}\right] \\
> & \frac{T}{8} \cdot V_{\max} \cdot \exp\left[-\frac{K}{K-1}\right] \geq \frac{1}{8e^2} \cdot T \cdot V_{\max}. && (51)
\end{aligned}$$

Step (49) is again by Theorem C.1. Step (50) is by the Chain rule (Theorem C.2), and our observations on the constructed P, Q. Step (54) is by the choice of arm a' and the KL divergence between $P_{a'}^{(\text{on})}, Q_{a'}^{(\text{on})}$. The strict inequality in (51) is by the case assumption that $2\sqrt{KT} > T \cdot (V_{\max} + 1/\tau_*) > T \cdot V_{\max}$, which implies $\frac{T}{K-1} \cdot \frac{V_{\max}^2}{2} < \frac{K}{K-1}$. The larger than equal in (51) is by the assumption that $K \geq 2$. Altogether, the desire regret lower bound of

$$\max\{\mathbb{E}[\text{Reg}_T(\pi, P)], \mathbb{E}[\text{Reg}_T(\pi, Q)]\} \geq \frac{1}{16e^2} \cdot T \cdot V_{\max}$$

is achieved.

Case 2: $2\sqrt{KT} \leq (V_{\max} + 1/\sqrt{\tau_*}) \cdot T$. We derive a regret lower bound of

$$\Omega\left(\min\left\{\sqrt{KT}, T \cdot \left(\frac{1}{\sqrt{\tau_*}} + V_{\max}\right)\right\}\right) = \Omega(\sqrt{KT}),$$

largely by following **Case 1b** with a different Δ, as well as the proof of Lemma 15.2 in (Lattimore & Szepesvári, 2020). Recall $\sqrt{KT} \geq T/\sqrt{\tau_*}$ since $\tau^* \geq T/K$, so the case condition implies $\sqrt{KT} \leq TV_{\max}$, meaning $\sqrt{K/T} \leq V_{\max}$. Now, set $\Delta = (1/2)\sqrt{(K-1)/T}$, and consider the Guassian instance with reward disribution P:

$$P_a^{(\text{off})} = \mathcal{N}(0,1) \text{ for all } a \in \mathcal{A}, \text{ and } P_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1, \\ \mathcal{N}(0,1) & \text{if } a \in \mathcal{A} \setminus \{1\}. \end{cases}$$

Now, there exists an arm $a' \in \mathcal{A} \setminus \{1\}$ such that $\mathbb{E}_{P,\pi}[N_T(a')] \leq T/(K-1)$. Consider the Gaussian instance with

$$Q_a^{(\text{off})} = \mathcal{N}(0,1) \text{ for all } a \in \mathcal{A}, \text{ and } Q_a^{(\text{on})} = \begin{cases} \mathcal{N}(\Delta, 1) & \text{if } a = 1, \\ \mathcal{N}(2\Delta, 1) & \text{if } a = a', \\ \mathcal{N}(0,1) & \text{if } a \in \mathcal{A} \setminus \{1, a'\}. \end{cases}$$

To this end, note that $P_a^{(\text{off})} = Q_a^{(\text{off})}$ for all $a \in \mathcal{A}$, and $P_a^{(\text{on})} = Q_a^{(\text{on})}$ for all $a \in \mathcal{A} \setminus \{a'\}$, but $P_{a'}^{(\text{on})} \neq Q_{a'}^{(\text{on})}$. In addition, both P, Q belong to $\mathcal{I}_V$. Consider the event $E = \{N_T(1) < T/2\}$. We have

$$\begin{aligned} \mathbb{E}[\text{Reg}_T(\pi, P)] + \mathbb{E}[\text{Reg}_T(\pi, Q)] \geq & \frac{T}{2} \cdot \Delta \cdot \left[\Pr_{P,\pi}(E) + \Pr_{Q,\pi}(E^c) \right] & (52) \\ \geq & \frac{T}{4} \cdot \Delta \cdot \exp\left[-\mathbb{E}_{P,\pi}[N_T(a')] \cdot \text{KL}(P_{a'}^{(\text{on})}, Q_{a'}^{(\text{on})}) \right] & (53) \\ \geq & \frac{T}{4} \cdot \Delta \cdot \exp\left[-\frac{T}{K-1} \cdot 2 \cdot \Delta^2 \right]. & (54) \end{aligned}$$

Step (52) is again by Theorem C.1. Step (53) is by the Chain rule (Theorem C.2), and our observations on the constructed P, Q. Step (54) is by the choice of arm a' and the KL divergence between $P_{a'}^{(\text{on})}, Q_{a'}^{(\text{on})}$. Finally, putting in $\Delta = (1/2)\sqrt{(K-1)/T}$ gives us

$$\max\{\text{Reg}_T(\pi, P), \text{Reg}_T(\pi, Q)\} \geq \frac{1}{8\sqrt{e}} \cdot \sqrt{(K-1)T}.$$

Altogether, the three cases cover all possibilites and the Theorem is proved.