On the price of exact truthfulness in incentive-compatible online learning with bandit feedback: a regret lower bound for WSU-UX

Ali Mortazavi
University of Victoria
alithemorty@gmail.com

Junhao Lin[*]
University of Waterloo
linjunhao9385@gmail.com

Nishant A. Mehta
University of Victoria
nmehta@uvic.ca

Abstract

In one view of the classical game of prediction with expert advice with binary outcomes, in each round, each expert maintains an adversarially chosen belief and honestly reports this belief. We consider a recently introduced, strategic variant of this problem with selfish (reputation-seeking) experts, where each expert strategically reports in order to maximize their expected future reputation based on their belief. In this work, our goal is to design an algorithm for the selfish experts problem that is incentive-compatible (IC, or *truthful*), meaning each expert's best strategy is to report truthfully, while also ensuring the algorithm enjoys sublinear regret with respect to the expert with the best belief. Freeman et al. (2020) recently studied this problem in the full information and bandit settings and obtained truthful, no-regret algorithms by leveraging prior work on wagering mechanisms. While their results under full information match the minimax rate for the classical ("honest experts") problem, the best-known regret for their bandit algorithm WSU-UX is $O(T^{2/3})$, which does not match the minimax rate for the classical ("honest bandits") setting. It was unclear whether the higher regret was an artifact of their analysis or a limitation of WSU-UX. We show, via explicit construction of loss sequences, that the algorithm suffers a worst-case $\Omega(T^{2/3})$ lower bound. Left open is the possibility that a different IC algorithm obtains $O(\sqrt{T})$ regret. Yet, WSU-UX was a natural choice for such an algorithm owing to the limited design room for IC algorithms in this setting.

1 INTRODUCTION

In the problem of prediction with expert advice (Vovk, 1995), we have K experts and T days and a fixed loss function $\ell(x_t, y_t)$. Each day $t \in [T]$, the learner has access to the advice $b_{i,t}$ of each expert $i \in [K]$ about the outcome y_t. Using the expert advice, the learner predicts x_t and then outcome y_t is revealed. The error of the learner is measured by $\ell(x_t, y_t)$ at round t. The goal of the learner is to achieve low regret with respect to the cumulative loss of the best expert in hindsight. One common approach to this problem in the literature is to, for any round t, maintain a set of weights over experts $w_{t,i}$ for all $i \in [K]$, select the advice of expert i with probability $\frac{w_{t,i}}{\|\boldsymbol{w}_t\|_1}$, and update the weights appropriately once the outcome y_t is revealed. See for example the multiplicative weight update (MWU) method (Arora et al., 2012) and Hedge (Freund and Schapire, 1997).

Roughgarden and Schrijvers (2017) considered this problem for binary outcomes (i.e. $y_t \in \{0, 1\}$) where the experts are strategic: in each round t, each expert forms a belief $b_{i,t} \in [0, 1]$ about the binary outcome y_t, (i.e. $b_{i,t} = \Pr(y_t = 1) \in [0, 1]$) and reports $r_{i,t} \in [0, 1]$ in such a way as to maximize its own future *reputation* among the pool of experts. See the protocol in Algorithm 1 which shows this framework.

Moreover, considering the class of learning algorithms that maintain weights over experts $w_{t,i}$ for $i \in [K]$, Roughgarden and Schrijvers (2017) assumed that each expert i at round t associates its own current reputation with the weight $w_{t,i}$ and its future reputation simply as its weight in the next round $w_{t+1,i}$. This type of expert is called a *myopic* expert as it does not consider the impact of the decision on its long-term reputation.

[*]Work completed while at University of Victoria

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

Algorithm 1: Protocol for Prediction With Selfish (Reputation Seeking) Experts for binary outcomes

```
Input: $T$, $K$, $\ell(X,Y) : [0,1] \times \{0,1\} \to \mathbb{R}$
for $t = 1, \ldots, T$ do
    The learner chooses a distribution $\boldsymbol{\pi_t} \in \Delta_K$
     over experts and draws an expert $I_t$.
    Each expert $i \in [K]$ forms a belief $b_{i,t} \in [0,1]$
     about the distribution of outcome $y_t$.
    Each expert $i$ reports a prediction $r_{i,t} \in [0,1]$
     with the goal of
     maximizing their own future reputation.
    Nature reveals the outcome $y_t \in \{0,1\}$.
    Learner incurs loss of
     $\mathbb{E}[\ell(r_{I_t,t}, y_t)] = \sum_{j \in [K]} \pi_{t,i} \ell(r_{j,t}, y_t)$
```

Given this notion of future reputation, the design of the algorithm would impact how each expert would report. Consider round t and expert i and a fixed weight-based learning algorithm; depending on the learning algorithm, there is a function f that determines the weight $w_{t+1,i}$ as

$$w_{t+1,i} = f(r_{i,t}, y_t, r_{-i,t}, h_{t-1}), \tag{1}$$

where $r_{-i,t}$ denotes the reports of the experts other than expert i and h_{t-1} is all the information revealed by the end of round $t-1$.

Moreover, assume that expert i has perfect information about $r_{-i,t}$ and h_{t-1}. Assuming expert i has a belief $b_{i,t}$ about the distribution of y_t, then it reports $r_{i,t}$ to maximize its expected reputation:

$$\begin{aligned} r_{i,t} &= \arg\max_{r \in [0,1]} \mathbb{E}_{b_{i,t}} \left[w_{t+1,i} \right] \\ &= \arg\max_{r \in [0,1]} \mathbb{E}_{b_{i,t}} \left[f(r, y_t, r_{-i,t}, h_{t-1}) \right], \end{aligned}$$

where the expectation is over the randomness of the outcome y_t as if it were drawn based on the expert's belief $b_{i,t}$. Observe that depending on function f, the value the expert i reports $r_{i,t}$ can be different than its belief $b_{i,t}$. If truth-telling is always a dominating strategy, meaning that no matter the other experts' reports $r_{-i,t}$, the best response is $r_{i,t} = b_{i,t}$, then the algorithm is called *incentive-compatible.* The design of incentive-compatible algorithms is desirable for two reasons:

- Quality of prediction: The regret guarantee for an incentive-compatible online learning algorithm holds not only for the expert with the best reports but also holds for the expert with the best beliefs, which the algorithm does not have direct access to. This guarantee is called *belief regret.*[1]
- Natural strategy: An expert does not need to take into consideration the reports of other experts. Moreover, when a simple strategy (truth-telling) is strictly dominating, it is reasonable to expect that agents will choose that strategy.

As observed by Roughgarden and Schrijvers (2017), the design of incentive-compatible online learning algorithms is intimately connected to the problem of designing proper scoring rules (see Definition 3 in Section 2.3). This implies that when the loss function is proper, the problem is easy. For instance, for any proper loss function, such as squared loss, MWU which uses the update rule

$$w_{t+1,i} = w_{t,i}(1 - \eta \ell(r_{i,t}, y_t)),$$

is incentive-compatible.

However, for absolute loss, which is not a proper loss function, Roughgarden and Schrijvers (2017, Corollary 31) showed that, under some mild restrictions[2], no weight-based randomized algorithm can achieve no-regret.

Yet, as observed by Freeman et al. (2020), even for proper loss functions such as squared loss, for another natural variation of incentive for experts who, in any round t, want to maximize their expected *normalized* weight (i.e. the probability of being selected) in the next round based on their private belief about the outcome y_t, the classical multiplicative weight algorithm (MWU) fails to be truthful[3]. Freeman et al. (2020) designed the Weighted-Score Update rule which is truthful and also achieves $O(\sqrt{T})$ regret in the full-information setting. However, in their extension to the multi-armed bandit setting (which they refer to as the partial information setting), their algorithm WSU-UX achieves $O(T^{2/3})$ regret, which does not match the minimax optimal rate $O(\sqrt{T})$ in the classical "honest experts" problem.

Although the experimental results by Freeman et al. (2020) suggested that WSU-UX performs similarly to EXP3 (Auer et al., 2002) which has minimax optimal regret of $O(\sqrt{T})$ in the classical "honest experts" problem, it remained an open problem whether the $O(T^{2/3})$ regret of WSU-UX is due to an artifact in the analysis or if instead the algorithm cannot achieve lower regret.

[1] Frongillo et al. (2021) used the term "regret with respect to the true beliefs".

[2] The class of algorithms that are considered by Roughgarden and Schrijvers (2017) are those that are a natural extension of deterministic weighted majority algorithms, where the weight update has some mild restrictions.

[3] It is easy to show that Hedge and MWU are not truthful. However, note that Frongillo et al. (2021) showed that Hedge is approximately truthful. Under some assumptions, approximately truthfulness is enough to get good belief regret, but in this paper, we only focus on exactly-truthful/incentive-compatible algorithms: the algorithms where truth-telling is the only dominant strategy.

Main Question The main question that we are interested in understanding is: in the setting of Freeman et al. (2020), whether learning with reputation-seeking experts under bandit feedback is strictly harder than the classical bandit problem.

Contribution We take one step toward answering this question by showing that WSU-UX, which is a very natural choice for this problem, can not achieve regret better than $\Omega(T^{2/3})$ in the worst case. In particular, we show that for any choice of hyperparameters for WSU-UX, for large enough T, there exists a loss sequence where the belief regret is $\Omega(T^{2/3})$. The construction of the loss sequence is fairly simple but, for the set of non-trivial hyperparameters (see Section 3.2 for a description of this set) requires a highly intricate analysis. In particular, for the non-trivial case, we design a loss sequence such that (i) the best expert[4] has an estimated loss with large variance for a constant fraction of T rounds, and at the same time (ii) the best expert outperforms the other experts at the end. The core technical difficulty is showing that both (i) and (ii) happen simultaneously.

2 MODEL AND PRELIMINARIES

2.1 Problem setting

We first describe the learning protocol. In the setting of Freeman et al. (2020), the protocol is the same as Protocol 1, where the loss function is squared loss (which is a proper loss). In the full-information version of the problem, all experts offer reports, while under bandit feedback, only the selected expert offers a report. The future reputation of each expert is defined as the probability of being selected in the next round. More concretely, each expert i at round t with belief $b_{i,t} \in [0,1]$ about binary outcome y_t strategically reports $r_{i,t} \in [0,1]$ to maximize their probability of being selected by the algorithm in the next round $t+1$. The goal of the learner is to minimize its belief regret, the regret with respect to the expert with the best belief.

Definition 1. Let $\boldsymbol{\pi}_t \in [K]$ be the learner's probability distribution in round t. Then the learner's *belief regret* $\mathbb{E}[\mathcal{R}_T]$ after T rounds is defined as

$$\mathbb{E}\left[\sum_{t\in[T]}\sum_{j\in[K]}\pi_{t,j}\ell(r_{j,t},y_t) - \min_{i\in[K]}\sum_{t\in[T]}\ell(b_{i,t},y_t)\right].$$

Note that the learner incurs loss according to reports $r_{i,t}$ whereas the performance of the best expert is measured with respect to $b_{i,t}$. In general, $r_{i,t}$ need not equal $b_{i,t}$. However, if a learning algorithm is incentive-compatible, meaning that truth-telling is the only strictly dominant strategy, it is reasonable to assume that $r_{i,t} = b_{i,t}$. In this case, low classical regret implies low belief regret. Next, we restate the definition of incentive-compatibility from Freeman et al. (2020).

Definition 2 (Freeman et al. (2020))**.** An online learning algorithm is incentive-compatible if for every timestep $t \in [T]$, every expert i with belief $b_{i,t}$, every report $r_{i,t}$, every vector of reports of the other experts $r_{-i,t}$, and every history of reports $(r_{t'})_{t'<t}$ and outcomes $(y_{t'})_{t'<t}$,

$$\begin{aligned}&\mathbb{E}_{y_t\sim\mathrm{Bern}(b_{i,t})}\left[\pi_{t+1,i} \mid (b_{i,t}, r_{-i,t}), y_t, (y_{t'})_{t'<t}, (r_{t'})_{t'<t}\right]\\ &\geq \mathbb{E}_{y_t\sim\mathrm{Bern}(b_{i,t})}\left[\pi_{t+1,i} \mid (r_{i,t}, r_{-i,t}), y_t, (y_{t'})_{t'<t}, (r_{t'})_{t'<t}\right],\end{aligned}$$

where $y \sim \mathrm{Bern(b)}$ denotes a random variable taking value 1 with probability b and 0 otherwise.

2.2 Motivation for Bandit Setting

To motivate the bandit version of the problem, consider the following example. A forecasting agency wishes to forecast an event and has a choice of which forecaster to employ. The selected forecaster will be given a fixed payment (say $1000) from the agency to research the event, will then develop their belief about the likelihood of the event occurring, and will finally decide what probability forecast (report) to give the forecasting agency. Any forecaster that is not selected will not receive a payment and will never provide a report to the forecasting agency. Naturally, the agency desires accurate reports, and so its goal is to select the forecaster whose belief (which can be developed only after the forecaster is funded and hence was selected) is the most accurate. To this end, the agency should ensure that the *future* expected payment given to a selected forecaster (for the next event) incentivizes the forecaster to report its belief honestly for the current event. The forecaster's incentive is exactly equal to the future probability of being selected (a quantification of the forecaster's reputation) since, if selected in the future, the selected forecaster will again receive a fixed payment. The agency should thus ensure that a forecaster's future probability of being selected is directly proportional to the accuracy of the forecaster's report, an accuracy which is known once the outcome has been realized.

2.3 Preliminaries

In this subsection, we overview fundamental concepts relevant to incentive-compatibility. We first recall the notion of proper scoring rules, which can be used

[4] The best expert here is the expert whose belief has the lowest cumulative loss over T rounds.

to elicit information from an expert (Gneiting and Raftery, 2007), (Buja et al., 2005).

Definition 3. Let $\mathcal{Y}$ denote the outcome space and $\mathcal{R} \subseteq \Delta(\mathcal{Y})$ denote the distributional report space. A scoring rule $s : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ is *proper* if for any $b, r \in \mathcal{R}$, we have

$$\mathbb{E}_{Y\sim b}\left[s(b,Y)\right] \geq \mathbb{E}_{Y\sim b}\left[s(r,Y)\right],$$

and *strictly proper* if the inequality becomes tight only when $r = b$.

This implies that when a scoring rule is (strictly) proper, an expert with belief b about the distribution of outcome $Y \in \mathcal{Y}$ would (uniquely) maximize their expected score by reporting $r = b$.

A (strictly) proper loss function ℓ is defined similarly where truthful reporting of the belief (strictly) *minimizes* the expected loss. We assume $\mathcal{R} = [0, 1]$ and $\mathcal{Y} = \{0, 1\}$ in this paper.

Next, let us view any online algorithm for the problem of prediction with expert advice as follows.

Definition 4 (Probability-based update class of an online learning algorithm)**.** An online learner M maintains a distribution $\boldsymbol{\pi}_t$ over K experts. In round t, the learner draws an expert $I_t = i$ with probability $\pi_{t,i}$. Then, the outcome y_t is revealed, and the learner incurs loss $\ell_{I_t,t} = \ell(r_{I_t,t}, y_t)$ and updates $\boldsymbol{\pi}_{t+1}$ only as a function of $\boldsymbol{r}_t = (r_{t,1}, \ldots, r_{t,K})$, $\boldsymbol{\pi}_t = (\pi_{t,1}, \ldots, \pi_{t,K})$, and outcome y_t.

Note that many algorithms can be described as probability-based update algorithms, such as MWU and Hedge. In order to make a distinction, we describe the precise definition of Hedge and MWU using the probability-based update description.

Definition 5. Hedge initializes the weights $w_{1,i} = \frac{1}{K}$ for all $i \in [K]$, updates the weights in each round based on the update

$$w_{t+1,i} = w_{t,i} \cdot \exp\left(-\eta\ell(r_{i,t}, y_t)\right), \tag{2}$$

and chooses $\pi_{t,i} = \frac{w_{t,i}}{\sum_j w_{t,j}}$.

MWU does the same except it uses the update

$$w_{t+1,i} = w_{t,i} \cdot (1 - \eta\ell(r_{i,t}, y_t)).$$

Note that $1 - \eta\ell(r_{i,t}, y_t)$ is a linear approximation of $\exp\left(-\eta\ell(r_{i,t}, y_t)\right)$ around 0.

Unlike the setting of Roughgarden and Schrijvers (2017) where using MWU with a proper loss function ℓ implies incentive-compatibility, in this setting we do not achieve incentive-compatibility since the normalization would impact the incentive.[5] Indeed, in the MWU algorithm, in round t, depending on the outcome y_t, the sum of the weights of all experts (the normalization factor) can be different. This will skew the incentive of an expert who wants to report to maximize the expected normalized weight.

2.3.1 Connection to Wagering Mechanism

Toward getting an update rule that is incentive compatible, Freeman et al. (2020) observed a connection between online learning algorithms and *wagering mechanisms*. In a wagering mechanism, each player reports their prediction about a random outcome and at the same time wagers (bets) a non-negative amount of money on their prediction. Once the outcome is realized, the mechanism will pay each player a payment based on the quality of their prediction and the amount they wagered.

More concretely, consider the specific setting of wagering mechanisms where there are K fixed players called experts, and there is an unknown Bernoulli outcome $y \in \{0, 1\}$. Each expert $i \in [K]$ with belief $b_i \in [0, 1]$ about the probability that $y = 1$ wagers $m_i > 0$ and reports $r_i \in [0, 1]$ with the goal of maximizing the expected payment from the mechanism. Note that r_i may or may not be equal to b_i. Once the random outcome y is revealed, the mechanism takes the vector of reports $\boldsymbol{r} = (r_1, \ldots, r_K)$, wagers $\boldsymbol{m} = (m_1, \ldots, m_K)$, and the realization y and outputs a K dimensional vector of $\Gamma(\boldsymbol{r}, \boldsymbol{m}, y) \in \mathbb{R}^K$ where $\Gamma_i(\boldsymbol{r}, \boldsymbol{m}, y) \in \mathbb{R}$ is the payment that the mechanism will pay to expert i.

A mechanism is called incentive-compatible if any expert i with belief b_i strictly maximizes their expected payment by reporting $r_i = b_i$, i.e.,

$$b_i = \arg\max_{r\in\mathcal{R}} \mathbb{E}_{y\sim\text{Bern}(b_i)}\left[\Gamma_i\left((r_i, \boldsymbol{r}_{-i}), \boldsymbol{m}, y\right)\right]$$

for any fixed vector $\boldsymbol{r}_{-i}$ of reports of the other experts and vector of wagers $\boldsymbol{m}$.

The Weighted-Score Wagering Mechanism (WSWM) is an incentive-compatible wagering mechanism defined as follows.

Definition 6 (Lambert et al. (2008))**.** The *Weighted-Score Wagering Mechanism* is a wagering mechanism that maps any vectors $\boldsymbol{r}$ and $\boldsymbol{m}$ and outcome y to payment $\Gamma = (\Gamma_1, \ldots, \Gamma_K)$, where

$$\Gamma_i^{\text{WSWM}}(\boldsymbol{r}, \boldsymbol{m}, y) = m_i\left(1 - \ell(r_i, y) + \sum_{j\in[K]} m_j\ell(r_j, y)\right)$$

[5] Note that Hedge is not incentive-compatible even in the setting of Roughgarden and Schrijvers (2017).

is the payment for expert i and ℓ is a strictly proper loss function.

This mechanism has several essential properties (Lambert et al., 2008). First, the mechanism is budget-balanced meaning $\sum_j \Gamma_j(\boldsymbol{r}, \boldsymbol{m}, y) = \sum_j m_j$, and moreover, the payment is non-negative, i.e., $\Gamma_i(\boldsymbol{r}, \boldsymbol{m}, y) \geq 0$. Observe that designing an incentive-compatible probability-based online learning algorithm can be seen as designing an incentive-compatible wagering mechanism that is budget balanced with non-negative payment as follows.

Consider the probability-based description of any online learning algorithm. This algorithm wants to reallocate $\boldsymbol{\pi}_t$ to $\boldsymbol{\pi}_{t+1}$. To do that, the algorithm at round t asks for the reports of the experts. Once the outcome y_t is revealed, the algorithm uses the reports of the experts $\boldsymbol{r}_t = (r_{t,1}, \ldots, r_{t,K})$, the wagers vector $\boldsymbol{\pi}_t = (\pi_{t,1}, \ldots, \pi_{t,K})$, and outcome y_t to set the probability in the next round as

$$\pi_{t+1,i} = \Gamma_i(\boldsymbol{r}_t, \boldsymbol{\pi}_t, y_t).$$

Since the payment $\Gamma_i(\boldsymbol{r}_t, \boldsymbol{\pi}_t, y_t)$ is incentive-compatible, experts will report truthfully.

Therefore, among the algorithms in the class of online probability-based update online learning algorithms defined in Definition 4, the only incentive-compatible ones are the ones where their update can be described as a wagering mechanism update that is non-negative and budget-balanced.

Using a wagering mechanism itself does not imply a no-regret guarantee; however, Freeman et al. (2020) designed a wagering mechanism called Weighted-Score Update (WSU)[6] in which the mechanism uses the update

$$\begin{aligned}
\pi_{t+1,i} &= \Gamma_i^{\mathrm{WSU}}(\boldsymbol{r}_t, \eta\boldsymbol{\pi}_t, y_t) \\
&= \Gamma_i^{\mathrm{WSWM}}(\boldsymbol{r}_t, \eta\boldsymbol{\pi}_t, y_t) + \Gamma_i^{\mathrm{Const}}(\boldsymbol{r}_t, (1-\eta)\boldsymbol{\pi}_t, y_t) \\
&= \Gamma_i^{\mathrm{WSWM}}(\boldsymbol{r}_t, \eta\boldsymbol{\pi}_t, y_t) + (1-\eta)\pi_{t,i}
\end{aligned}$$

for any $i \in [K]$, where $\eta \in (0, 0.5)$ and $\Gamma_i^{\mathrm{Const}}$ is simply a mechanism that returns the input wagers. Using the definition of WSWM, this update may be written as

$$\pi_{t+1,i} = \pi_{t,i}\left(1 - \eta\left(\ell_{t,i} - \sum_j \pi_{t,j}\ell_{t,j}\right)\right). \tag{3}$$

Freeman et al. (2020) showed that WSU with the update of form (3) can achieve $O(\sqrt{T \ln K})$ regret in the full-information setting.

```
Algorithm 2: WSU-UX (Freeman et al., 2020)
Input: $\eta, \gamma \in (0, 1/2)$ such that $\frac{\eta K}{\gamma} \leq 1/2$, and
       loss sequence $\ell(x, y)$.
Set $\pi_{1,i} = \frac{1}{K}, \forall i \in [K]$
for $t \in [T]$ do
    The learner chooses expert $I_t$ according to
      distr. $\tilde{\pi}_{t,i} = (1-\gamma)\pi_{t,i} + \frac{\gamma}{K}, \forall i \in [K]$.
    Arm $i = I_t$ forms a belief $b_{i,t} \in [0, 1]$.
    Arm $i = I_t$ reports a report $r_{i,t} \in [0, 1]$ with
      the goal of maximizing $\mathbb{E}_{y_t \sim \mathrm{Bern(b_{i,t})}}[\pi_{t+1,i}]$.
    Nature reveals the outcome $y_t \in \{0, 1\}$
    The learner computes $\hat{\ell}_{i,t} = \frac{\ell(r_{i,t}, y_t)}{\tilde{\pi}_{i,t}}$ for $i = I_t$
      and $\hat{\ell}_{j,t} = 0, \forall j \neq I_t$.
    The learner updates
      $\pi_{t+1,i} = \pi_{t,i}\left(1 - \eta\left(\hat{\ell}_{t,i} - \sum_{j=1}^{K} \pi_{t,j}\hat{\ell}_{t,j}\right)\right)$.
```

Interestingly, an apparently unnoticed connection is that the update of form (3) recovers the same update as the ML-Prod update of Gaillard et al. (2014) if in all rounds all experts use the same learning rate η.

2.4 Existing Bandit Results

Freeman et al. (2020) extended their result to the bandit case by designing the Weighted-Score Update with Uniform Exploration (WSU-UX) algorithm described in Algorithm 2, and they showed a $O(T^{2/3}(K \ln K)^{1/3})$ upper bound on the regret of this algorithm. In particular, in their algorithm, they used the common technique of constructing unbiased importance-weighted loss estimates. The algorithm then applies the WSU update on the estimated losses to update the probability distribution. For some technical reasons, they additionally needed to mix the probability distribution over arms ($\boldsymbol{\pi}_t$) with a uniform distribution with weight $\gamma \in [0, 1]$ to get the probability distribution ($\tilde{\boldsymbol{\pi}}_t$) from which an arm is selected, i.e., $\tilde{\pi}_{t,i} = (1-\gamma)\pi_{t,i} + \gamma\frac{1}{k}$. The two technical reasons for using γ are as follows:

1. To make sure that after each update, $\pi_{t,i}$ is still a valid probability distribution.

2. Their regret upper bound can be extremely large in case they do not mix (i.e. $\gamma = 0$).

Note that mixing with uniform distribution is not for the purpose of getting high probability bounds, as they bound pseudo-regret.[7]

Indeed, they showed for WSU-UX with learning rate

[6] They do not call their update rule as a wagering mechanism, but it can be viewed as a wagering mechanism.

[7] "Regret" in this work is actually pseudo-regret, which equals expected regret under oblivious beliefs and outcomes.

η and mixing weight γ,

$$\mathbb{E}[\mathcal{R}_T] \le \gamma T + \frac{\eta K T}{\gamma} + \frac{\ln K}{\eta} + 2\eta K T.$$

The best choice of η and γ attains the regret

$$\mathbb{E}[\mathcal{R}_T] \le 2(4T)^{2/3}(K \ln K)^{1/3},$$

which is $O(T^{2/3})$ in terms of T.

It was unclear whether the higher regret was an artifact of their analysis or a limitation of WSU-UX. If there is a tighter analysis of WSU-UX's regret, then for some valid γ, η, we would have $\mathbb{E}[\mathcal{R}_T] = o(T^{2/3})$. However, we show that for any valid γ, η, (see Section 3.2 for a description of valid hyperparamters) for large enough T, there exists a loss sequence for which $\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3})$, implying that WSU-UX cannot achieve regret better than $O(T^{2/3})$.

3 REGRET LOWER BOUND FOR WSU-UX

3.1 Potential Analysis View Comparision between EXP3 and WSU-UX

A potential(-based) analysis is a common way to analyze the regret of online learning algorithms (Cesa-Bianchi and Lugosi, 2003). We compare the potential analysis of WSU-UX and EXP3, beginning with the full-information variation of each algorithm and then turning to the implications in the bandit setting.

In the potential analysis of Hedge, for any $i \in [K]$ and $t \in [T]$, we define $\Phi^{\text{HEDGE}}_{t,i} := w_{t,i}$ with $w_{t,i}$ as in Definition 5. We define $\Phi^{\text{HEDGE}}_{t} := \sum_{j\in[K]} w_{t,j}$. By non-negativity of $w_{t,i}$, we have

$$\frac{1}{\eta} \ln\left(\Phi^{\text{HEDGE}}_{T+1,i}\right) \le \frac{1}{\eta} \ln\left(\Phi^{\text{HEDGE}}_{T+1}\right), \tag{4}$$

where η is the learning rate of the algorithm. From the LHS and RHS of (4), we can extract the cumulative loss of expert i and the cumulative loss of the learning algorithm respectively. However, we might not be able to exactly extract these two quantities from the potentials as there might be some error terms involved in the extraction process. Indeed, for Hedge, the LHS is exactly the cumulative loss of expert i; however, the RHS can only be upper bounded by the cumulative loss of the learner plus some extra terms:

$$\frac{1}{\eta} \ln\left(\Phi^{\text{HEDGE}}_{T+1}\right) \le \sum_{t\in[T]} \sum_j \pi_{t,j}\ell_{t,j} + \underbrace{\frac{\ln K}{\eta}}_{\text{exploration term}} + \eta \underbrace{\sum_{t\in[T]} \left[\sum_j \pi_{t,j} \left(\ell_{t,j}\right)^2\right]}_{\text{Second order error}}. \tag{5}$$

These two terms will appear in the regret analysis.

On the other hand, note that the WSU update can be written as a linear approximation of the Hedge update at the point $\bar{\ell}_t := \sum_j \pi_{t,j}\ell_{t,j}$ (see Appendix C for details). This means that WSU just uses a linear approximation of Hedge when updating the potential. This change in potential function will impact the process of extracting the regret from the potential.

For WSU, the potential is defined as $\Phi^{\text{WSU}}_{t,i} := \pi_{t,i}$ and $\Phi^{\text{WSU}}_{t} := \sum_{j\in[K]} \pi_{t,j} = 1$. By non-negativity of $\pi_{t,i}$ we have

$$\frac{1}{\eta} \ln\left(\Phi^{\text{WSU}}_{T+1,i}\right) \le \frac{1}{\eta} \ln\left(\Phi^{\text{WSU}}_{T+1}\right) = 0. \tag{6}$$

Now, the RHS of (6) (which is 0) does not involve any second-order error term. In fact, since WSU is normalized, the RHS does not give us information about the regret. However, we can extract the difference between the cumulative loss of the algorithm and expert i from the LHS of (6), and this extraction process would lead to a second-order error term. Indeed we have

$$\frac{1}{\eta} \ln\left(\Phi^{\text{WSU}}_{T+1,i}\right) \ge \sum_{t\in[T]} \left[\sum_j \pi_{t,j}\ell_{t,j} - \ell_{t,i}\right] - \underbrace{\frac{\ln K}{\eta}}_{\substack{\text{exploration}\\ \text{term}}} - \eta \underbrace{\sum_{t\in[T]} \left[\sum_j \pi_{t,j}\ell_{t,j} - \ell_{t,i}\right]^2}_{\text{Second-order error}}. \tag{7}$$

These two terms in (7) will appear in the regret.

Comparing (7) and (5): Note that the error term in (7) is a second-order version of $\left[\sum_j \pi_{t,j}\ell_{t,j} - \ell_{t,i}\right]$ for a *fixed* i whereas the second-order term in (5), which is $\left[\sum_j \pi_{t,j}\left(\ell_{t,j}\right)^2\right]$, is a *weighted average* of $\left(\ell_{t,j}\right)^2$ weighted by π_t.

Implication for Bandit Case Now, in the bandit case where we use $\hat{\ell}_{t,i}$ to be an unbiased estimated loss for quantity $\ell_{t,i}$, the expectation of the second-order term in (5) is $\mathbb{E}\left[\sum_j \pi_{t,j}\left(\hat{\ell}_{t,j}\right)^2\right] = O(K)$, whereas the expectation of the second-order term in (7) is $\mathbb{E}\left[\left(\sum_j \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,i}\right)^2\right] \le \mathbb{E}\left[\left(\sum_j \pi_{t,j}\hat{\ell}_{t,j}\right)^2\right] + \mathbb{E}\left[\left(\hat{\ell}_{t,i}\right)^2\right] = 2K + \mathbb{E}\left[\frac{1}{\tilde{\pi}_{t,i}}\right] = O(\frac{K}{\gamma})$. This difference makes the regret bound for WSU-UX larger.

However, it is not clear whether the potential-based analysis is tight or if there might be a way to get a better regret upper bound. Next, we show our main result, a lower bound demonstrating that it is not possible to get a better upper bound.

3.2 Lower Bound Proof

We show that WSU-UX cannot achieve regret better than $\Omega(T^{2/3})$. The following theorem restricts the focus to valid settings of the hyperparameters, which we define after the theorem.

Theorem 7. *For any valid set of hyperparameters (η, γ) there exists T_0 such that for any $T \geq T_0$,*

$$\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3}).$$

The notion of valid hyperparameters is taken from the restrictions imposed by Freeman et al. (2020). In particular, the restrictions are $\frac{\eta K}{\gamma} \leq 1/2$ and $\eta, \gamma \in (0, 1/2)$. See the beginning of the Appendix A for more information about the restrictions.

Now, we further partition the set of valid hyperparameters (η, γ) into two cases:

- the trivial case: $\eta < T^{-2/3}$ or $\gamma > T^{-1/3}$;
- the non-trivial case: $\eta \geq T^{-2/3}$ and $\gamma \leq T^{-1/3}$.

In the trivial case, either the learning rate η is too small, causing the algorithm to take a long time to concentrate on the optimal expert and incurring a large regret of order $\Omega(T^{2/3})$, or γ is so large that the uniform exploration would cause the algorithm to incur a large regret of $\Omega(T^{2/3})$. The proof for the trivial case, along with all other results in this paper, can be found in the appendix. For the non-trivial case, we show the following.

Theorem 8. *For $K = 2$ and for any valid set of (η, γ) in the non-trivial case, there exists T_0 such that for any $T \geq T_0$, we have a loss sequence $\{\ell_t\}_{t=1}^T$ such that*

$$\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3}). \tag{8}$$

3.3 High-level Proof for Non-Trivial Case

In this subsection, we give a high-level proof of Theorem 8. We first introduce the following loss sequence.

Definition 9. For any T, we define

$$\{\ell_t\}_{t=1}^T = \begin{cases} \ell_{t,1} = 1, \ell_{t,2} = 0 & \text{for } 1 \leq t \leq \frac{T}{100} \\ \ell_{t,1} = 0, \ell_{t,2} = 1 & \text{for } \frac{T}{100} < t \leq T \end{cases}.$$

Moreover, we call the set of rounds $\{t : 1 \leq t \leq \frac{T}{100}\}$ Phase 1, where only arm 1 incurs loss, and the set of rounds $\{t : \frac{T}{100} < t \leq T\}$ Phase 2, where only arm 2 incurs loss.

From now on, by $\{\ell_t\}_{t=1}^T$ we mean the loss sequence defined in Definition 9. Note that in this loss sequence, the best arm is arm 1. Our goal is to show that this particular loss sequence forces the algorithm to incur large regret. We do this by decomposing the regret into three terms, as described in Theorem 10.

Theorem 10. *When running WSU-UX for any valid choice of (η, γ) in the non-trivial case, there exists T_0 such that for any $T \geq T_0$, for loss sequence $\{\ell_t\}_{t=1}^T$, we have for some constants $c_1, c_2, c_3 > 0$,*

$$\mathbb{E}\left[\mathcal{R}_T\right] \geq c_1 \frac{1}{\eta} + c_2 \frac{\eta T K}{\gamma} + c_3 \gamma T. \tag{9}$$

Note that Theorem 10 implies $\mathbb{E}\left[\mathcal{R}_T\right] = \Omega(T^{2/3})$ as the RHS of (9) can be lower bounded by $\Omega(T^{2/3})$.

3.3.1 Proof of Theorem 10

It remains to show (9). To do that, we first introduce the following key lemma that follows similar steps as Lemma 4.3 of Freeman et al. (2020) but with all the inequalities in the reverse direction, which implies a second-order lower bound.

Lemma 11 (Second-Order Lower Bound)**.** *For any valid choice of (η, γ) when running WSU-UX on loss sequence $\{\ell_t\}_{t=1}^T$, we get*

$$\begin{aligned} &\sum_{t=1}^T \sum_{j \in [K]} \pi_{t,j} \hat{\ell}_{t,j} - \sum_{t=1}^T \hat{\ell}_{t,1} \\ &\geq \frac{\ln \pi_{T+1,1} + \ln K}{\eta} + \frac{\eta}{4} \sum_{t=1}^T (\hat{\ell}_{t,1} - \sum_{j \in [K]} \pi_{t,j} \hat{\ell}_{t,j})^2. \end{aligned} \tag{10}$$

Next, to convert (10) to the lower bound in (9), we list three claims that hold when running WSU-UX on the loss sequence in Definition 9 given hyperparameters falling in the non-trivial case. Note that each claim corresponds to a term on the right-hand side of (9).

Claim 1 (Concentration on best arm at the end)**.** *For large enough T, there exists $c_1 > 0$ such that*

$$\mathbb{E}\left[\ln \pi_{T+1,1} + \ln K\right] \geq c_1. \tag{11}$$

Claim 2 (Second moment lower bound)**.** *For large enough T, there exists c_2 such that*

$$\mathbb{E}\left[\sum_{t=1}^T \left(\hat{\ell}_{t,1} - \sum_{j \in [1,2]} \pi_{t,j} \hat{\ell}_{t,j}\right)^2\right] \geq 4c_2 \frac{TK}{\gamma}.$$

Claim 3 (Bias induced by uniform exploration)**.** *For large enough T, there exists c_3 such that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j\in[1,2]}\tilde{\pi}_{t,j}\hat{\ell}_{t,j}-\sum_{j\in[1,2]}\pi_{t,j}\hat{\ell}_{t,j}\right)\right]\geq c_3\gamma T.$$

Given all three claims and Lemma 11, the proof of Theorem 10 is straightforward. In loss sequence $\{\ell_t\}_{t=1}^{T}$, the best arm is arm 1; therefore, Claim 3 gives

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T]&=\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j\in[K]}\tilde{\pi}_{t,j}\hat{\ell}_{t,j}-\sum_{t=1}^{T}\hat{\ell}_{t,1}\right]\\&\geq\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j}-\sum_{t=1}^{T}\hat{\ell}_{t,1}\right]+c_3\gamma T.\end{aligned}\tag{12}$$

To further lower bound (12), we take the expectation of both sides of (10) and use Claims 1 and 2 to get

$$\mathbb{E}\left[\mathcal{R}_T\right]\geq c_1\frac{1}{\eta}+c_2\frac{\eta TK}{\gamma}+c_3\gamma T.$$

3.3.2 Subtlety of Showing the Claims

The proof of the claims is not straightforward and requires subtle work. At a high level, we designed loss sequence $\{\ell_t\}_{t=1}^{T}$ such that for a constant fraction of rounds $\frac{T}{200}\leq t\leq\frac{T}{100}$, arm 1 (the best arm) has a cumulative loss linearly worse than arm 2. This can be used to show Claim 2. However, during these rounds $\frac{T}{200}\leq t\leq\frac{T}{100}$, arm 1 keeps getting small probability and hence very large estimated loss. Yet, because the algorithm has some uniform exploration, it picks arm 1 frequently. Therefore, at the beginning of round $t\geq\frac{T}{100}$, the probability update is very slow. Therefore, it is not obvious whether the algorithm can allow $\pi_{T,1}$ to recover at the end or not.

The next section gives a technical overview of how we prove the claims. We perform a careful analysis of the probability updates, leveraging a recently shown multiplicative form of Azuma's inequality (Kuszmaul and Qi, 2021) in some key steps to show that indeed with probability at least $1-O(\frac{1}{T^2})$, $\pi_{T_1+T_2+1,1}\geq 1/4$. We then show that with high probability $\pi_{T+1,1}\geq 3/4$. This implies Claim 1 and also Claim 3.

4 SHOWING THE CLAIMS

The proof of Claim 2 is fairly straightforward. The proof requires the following lemma.

Lemma 12. *In WSU-UX with two arms $(i,\bar{i})$,*

$$\mathbb{E}[\pi_{t+1,i}\mid\mathcal{F}_{t-1}]=(1-C_{t,i})\pi_{t,i}+C_{t,i}\,\pi_{t,i}^2,\tag{13}$$

where $C_{t,i}:=\eta\left(\ell_{t,i}-\ell_{t,\bar{i}}\right)$ and $\mathcal{F}_{t-1}$ is the history up until the end of round $t-1$.

This lemma can be used to show that for t in Phase 1, i.e., $t\leq\frac{T}{100}$, $\pi_{t,1}$ decreases in a multiplicative way and we have

$$\mathbb{E}[\pi_{t+1,1}]\leq(1-\frac{\eta}{2})\,\mathbb{E}[\pi_{t,1}].$$

The above inequality can be used to show that for a constant fraction of rounds $\frac{T}{200}\leq t\leq\frac{T}{100}$ we have

$$\mathbb{E}[\pi_{t,1}]\leq\frac{1}{KT}.\tag{14}$$

The above fact can be further utilized to show that the summation of the second moments of the estimated loss differences in Claim 2 is large, which proves Claim 2.

Claims 1 and 3 require more sophisticated techniques. To demonstrate them, we need to analyze the behavior of $\pi_{t,1}$ for $t\in[T]$ when running the algorithm on the loss sequence $\{\ell_t\}_{t=1}^{T}$. Note that since we are in the bandit case, $\pi_{t,1}$ is a random variable.

Recall Phases 1 and 2 from Definition 9, the definition of the loss sequence. We need to further decompose these phases into multiple sub-phases as defined below.

Definition 13. Define $T_1=\frac{1}{100}T,T_2=\frac{2}{10}T,T_3=\frac{1}{10}T,T_4=\frac{69}{100}T$, and then define (sub-)phases as follows:

- Phase 1: $\mathcal{T}_1=\{t:1\leq t\leq T_1\}$,
- Phase 2.1: $\mathcal{T}_2=\{t:T_1+1\leq t\leq T_1+T_2\}$,
- Phase 2.2: $\mathcal{T}_3=\{t:T_1+T_2+1\leq t\leq T_1+T_2+T_3\}$,
- Phase 2.3: $\mathcal{T}_4=\{t:T_1+T_2+T_3+1\leq t\leq T_1+T_2+T_3+T_4\}$.

Moreover, we define T' and M, which are intermediate numbers that are going to be used in our analysis regarding high probability statements in this section.

Definition 14. We define M and T' as follows

$$M:=\frac{1}{\ln 2}\left[\underbrace{\ln\left(\frac{2K}{\gamma}\right)}_{\propto(\ln T)}+\underbrace{2(1+\varepsilon_1)(1+\frac{\eta K}{\gamma})\eta T_1}_{\propto(\eta T_1)}\right]\tag{15}$$

$$T':=\frac{1}{1-\varepsilon_2}\,(\frac{4}{3-\gamma})\,\frac{2}{\eta}\,M\tag{16}$$

where $\varepsilon_1=\sqrt{\frac{6\ln T}{\frac{2\gamma}{K}T_1}}$ and $\varepsilon_2=\sqrt{\frac{4\ln T}{\frac{3-\gamma}{4}T_2}}$.

Note that since we are in the non-trivial case ($\eta \geq T^{-2/3}$ and $\gamma \leq T^{-1/3}$), as T goes to ∞, we have $\varepsilon_1, \varepsilon_2 \to 0$ and moreover $M \approx c\,\eta T_1$ and $T' = c' \frac{M}{\eta} \approx c\,c'\,T_1$. Note that for large enough T, we have

$$T' \leq T_2. \tag{17}$$

We now give a high-level picture of how we prove Claims 1 and 3, which also will give more intuition about M and T'.

Proof Sketch of Claims 1 and 3. Let $\mathcal{E}_1 = \{\pi_{T_1+1,1} \geq 2^{-M}\}$ be the event that arm 1's probability at the end of Phase 1 is not too small, where M is defined in (15). Let $\mathcal{E}_2 = \{\pi_{T_1+T_2+1,1} \geq \frac{1}{4}\}$ be the event that arm 1's probability at the end of Phase 2.2 has recovered to $\frac{1}{4}$.

First, we show that with high probability, the algorithm does not pull arm 1 too many times in Phase 1. In particular, we show that at the end of Phase 1, the probability $\pi_{T_1+1,1}$ of selecting arm 1 is lower bounded by 2^{-M} (hence, $\mathcal{E}_1$ happens). To prove this result, we leverage a recent multiplicative form of Azuma's inequality for martingales (Kuszmaul and Qi, 2021).

The next key step is to show that if $\mathcal{E}_1$ happens, then with high probability $\mathcal{E}_2$ happens. Since we already showed that $\mathcal{E}_1$ happens with high probability, it follows that with high probability we have that $\pi_{T_1+T_2+1,1} \geq 1/4$. Now, to show this key step, we proceed as follows. First, we observe that in Phase 2.1, it is only via pulls of arm 2 that the probability of arm 1 can increase. Moreover, initially, the rate of update of arm 1 is $\frac{\pi_{t+1,1}}{\pi_{t,1}}$, which is very close to 1. Therefore, we first analyze how many pulls of arm 2 suffice for arm 1's probability to double, and we then analyze how many doublings are needed to satisfy event $\mathcal{E}_2$. Finally, we again use a martingale analysis (multiplicative Azuma) to show that within the rounds of Phase 2.1, the sufficient number of pulls of arm 2 occur with high probability and hence event $\mathcal{E}_2$ happens.

We then show that conditional on event $\mathcal{E}_2$, from Phase 2.2 onwards, the probability $\pi_{t,2}$ goes to zero exponentially quickly as t increments beyond $T_1 + T_2$. Therefore, the probability $\pi_{t,1}$ converges to 1 exponentially quickly. This, combined with Lemma 12, is essentially what is needed to prove Claim 3. For Claim 1, we use a careful analysis based on Chebyshev's inequality to show that with probability exponentially close to 1, $\pi_{T,1}$ is at least $3/4$. This is essentially what allows us to control the expected log probability term in Claim 1 and hence what allows Claim 1 to go through. □

Acknowledgements

AM and NM were supported by NSERC Discovery Grant RGPIN- 2018-03942 and a JP Morgan Faculty Research Award.

References

Arora, S., Hazan, E., and Kale, S. (2012). The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 3:13.

Cesa-Bianchi, N. and Lugosi, G. (2003). Potential-based algorithms in on-line prediction and game theory. *Machine Learning*, 51:239–261.

Freeman, R., Pennock, D., Podimata, C., and Vaughan, J. W. (2020). No-regret and incentive-compatible online learning. In *International Conference on Machine Learning*, pages 3270–3279. PMLR.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Frongillo, R., Gomez, R., Thilagar, A., and Waggoner, B. (2021). Efficient competitions and online learning with strategic forecasters. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 479–496.

Gaillard, P., Stoltz, G., and Van Erven, T. (2014). A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Kivinen, J. and Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63.

Kuszmaul, W. and Qi, Q. (2021). The multiplicative version of Azuma's inequality, with an application to contention analysis. *arXiv preprint arXiv:2102.05077*.

Lambert, N. S., Langford, J., Wortman, J., Chen, Y., Reeves, D., Shoham, Y., and Penno k, D. M. (2008). Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 170–179.

Roughgarden, T. and Schrijvers, O. (2017). Online prediction with selfish experts. *Advances in Neural Information Processing Systems*, 30.

Vovk, V. G. (1995). A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 51–60.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. YES
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [YES]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [YES in appendix]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [ot Applicable]

A Main Result And High-level proof

We give a high-level proof of the main theorem. For the convenience of the reader, we re-state the lemmas along with their proofs.

Theorem 7. *For any valid set of hyperparameters (η, γ) there exists T_0 such that for any $T \geq T_0$,*

$$\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3}).$$

We remind the reader that we restrict our attention to the set of valid hyperparameters: $\frac{\eta K}{\gamma} \leq 1/2$ and $\eta, \gamma \in (0, 1/2)$. The notion of valid hyperparameters is taken from the restrictions imposed by Freeman et al. (2020). We briefly explain some of the restrictions.

We note that we need $\frac{\eta K}{\gamma} \leq 1$ to make sure that we get a valid probability distribution π_{t+1} after updating in each round t. Moreover, to get sublinear regret, we need $\gamma \leq 1/2$; otherwise, uniform exploration will pick the suboptimal arm frequently, which can cause linear regret. Since $K \geq 1$, the inequality $\frac{\eta K}{\gamma} \leq 1$ implies $\eta \leq \frac{\gamma}{K} \leq \gamma \leq 1/2$ and therefore $\eta \leq 1/2$.

Our analysis focuses on the restriction $\frac{\eta K}{\gamma} \leq c$ for $c = 1/2$. Thus far, for technical reasons related to certain inequalities, we are not sure whether our particular analysis can be made to go through for a larger c (that is still less than 1). The main concern arises as c gets closer to 1. However, using advanced Taylor approximation-based inequalities, this might be possible.

Now, we further partition the set of valid hyperparameters (η, γ) into two cases:

- The trivial case: $\eta < T^{-2/3}$ or $\gamma > T^{-1/3}$
- The non-trivial case: $\eta \geq T^{-2/3}$ and $\gamma \leq T^{-1/3}$.

The proof strategy for this theorem is that for any valid hyperparameters (η, γ), we show there exists a loss sequence such that the algorithm will incur an expected regret of $\Omega(T^{2/3})$.

A.1 Trivial Case

We show that in the trivial case $\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3})$.

Lemma 15. *For $K = 2$ and for any setting (η, γ) in the trivial case where $\eta < T^{-2/3}$ or $\gamma > T^{1/3}$,*

$$\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3}).$$

We prove this by proving the following Lemma 16 for the case where $\eta < T^{-2/3}$ and Lemma 17 for the case where $\gamma > T^{-1/3}$.

Lemma 16. *For WSU-UX run with any $\eta < T^{-2/3}$ and any $\gamma \geq 0$ there exists a loss sequence $\{\ell_t\}_{t=1}^T$ and $c = \frac{1}{200} > 0$ such that*

$$\mathbb{E}\left[\mathcal{R}_T\right] \geq c T^{2/3}.$$

Before proving Lemma 16, let us restate and prove Lemma 12. We will use this lemma numerous times in the course of proving the main result.

Lemma 12. *In WSU-UX with two arms $(i, \bar{i})$,*

$$\mathbb{E}[\pi_{t+1,i} \mid \mathcal{F}_{t-1}] = (1 - C_{t,i})\pi_{t,i} + C_{t,i}\, \pi_{t,i}^2, \tag{13}$$

where $C_{t,i} := \eta \left(\ell_{t,i} - \ell_{t,\bar{i}}\right)$ and $\mathcal{F}_{t-1}$ is the history up until the end of round $t - 1$.

Proof of Lemma 12. When we have two arms, the update rule for arm i can be expressed as

$$\begin{aligned}\pi_{t+1,i} &= \pi_{t,i}\left(1-\eta\left(\hat{\ell}_{t,i}-\sum_{j\in\{i,\bar{i}\}}\pi_{t,j}\hat{\ell}_{t,j}\right)\right)\\ &= \pi_{t,i}-\eta\pi_{t,i}\hat{\ell}_{t,i}+\eta\pi_{t,i}\left(\sum_{j\in\{i,\hat{i}\}}\pi_{t,j}\hat{\ell}_{t,j}\right).\end{aligned}$$

Now, taking the expectation of both sides conditional on the past, we get

$$\begin{aligned}\mathbb{E}\left[\pi_{t+1,i}\mid\mathcal{F}_{t-1}\right] &= \mathbb{E}_{t-1}\left[\pi_{t+1,i}\right]\\ &= \mathbb{E}_{t-1}\left[\pi_{t,i}-\eta\pi_{t,i}\hat{\ell}_{t,i}+\eta\pi_{t,i}\sum_{j\in\{i,\bar{i}\}}\pi_{t,j}\hat{\ell}_{t,j}\right]\\ &= \pi_{t,i}-\eta\pi_{t,i}\mathbb{E}_{t-1}[\hat{\ell}_{t,i}]+\eta\pi_{t,i}^2\mathbb{E}_{t-1}\left[\hat{\ell}_{t,i}\right]+\eta(\pi_{t,i})(\pi_{t,\bar{i}})\mathbb{E}_{t-1}\left[\hat{\ell}_{t,\bar{i}}\right]\\ &= \pi_{t,i}-\eta\pi_{t,i}\ell_{t,i}+\eta\pi_{t,i}^2\ell_{t,i}+\eta(\pi_{t,i})(1-\pi_{t,i})\ell_{t,\bar{i}}.\end{aligned}$$

Now, rearranging the terms, we get

$$\begin{aligned}\mathbb{E}_t\left[\pi_{t+1,i}\right] &= \left(1-\eta\ell_{t,i}+\eta\ell_{t,\bar{i}}\right)\pi_{t,i}+\eta\left(\ell_{t,i}-\ell_{t,\bar{i}}\right)\pi_{t,i}^2\\ &= \left(1-C_{t,i}\right)\pi_{t,i}+\left(C_{t,i}\right)\pi_{t,i}^2,\end{aligned}$$

where we recall that $C_{t,i}=\eta\left(\ell_{t,i}-\ell_{t,\bar{i}}\right)$. □

Now, we are ready to prove Lemma 16.

Proof of Lemma 16. Consider loss sequence with two arms, $\{\ell_t\}_{t=1}^T$ where $\ell_{t,1}=0, \ell_{t,2}=1$ for $1\le t\le T$. In this case, we show that regret simplifies to the number of times we pull arm 2. In particular,

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{\mathrm{T}}] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{2}\tilde{\pi}_{t,j}\ell_{t,j}-\sum_{t=1}^{T}\ell_{t,1}\right]\\ &= \sum_{t=1}^{T}\mathbb{E}\left[\tilde{\pi}_{t,2}\right]\ell_{t,2} && (\ell_{t,1}=0,\forall t\in[T])\\ &= \sum_{t=1}^{T}\mathbb{E}[\tilde{\pi}_{t,2}]. && (\ell_{t,2}=1,\forall t\in[T])\end{aligned}$$

Remember that

$$\tilde{\pi}_{t,i}=(1-\gamma)\pi_{t,i}+\frac{\gamma}{2}. \tag{18}$$

Now, taking the expectation of both sides of (18) for $i=2$, we get

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\tilde{\pi}_{t,2}] &= \sum_{t=1}^{T}\mathbb{E}[(1-\gamma)\pi_{t,2}+\frac{\gamma}{2}]\\ &= (1-\gamma)\sum_{t=1}^{T}\mathbb{E}[\pi_{t,2}]+\frac{\gamma}{2}T. \qquad (19)\end{aligned}$$

Next, we lower bound the first term. Observe that by applying Lemma 12 on any round $t>1$, we get

$$\begin{aligned}\mathbb{E}[\pi_{t,2}|\mathcal{F}_{t-2}] &= (1-\eta)\,\pi_{t-1,2}+\eta\,\pi_{t-1,2}^2\\ &\ge (1-\frac{\eta}{2})\,\pi_{t-1,2}, \qquad (20)\end{aligned}$$

where the last inequality comes from the fact that $\pi_{t-1,2} \leq 1/2$. This fact is true because we know the initial value for $\pi_{1,2} = \frac{1}{2}$ in the first round, as this quantity can only decrease, $\mathbb{E}[\pi_{t,2}^2] \leq \frac{1}{2}\mathbb{E}[\pi_{t,2}]$. Now, taking expectation of both sides of (20), we get

$$\mathbb{E}[\pi_{t,2}] \geq (1 - \frac{\eta}{2})\mathbb{E}[\pi_{t-1,2}]. \tag{21}$$

Now, using (21) recursively, we get

$$\begin{aligned}\mathbb{E}[\pi_{t,2}] &\geq \pi_{1,2}\,(1 - \frac{\eta}{2})^{t-1} \\ &\geq \frac{1}{2}e^{-\eta(t-1)}. && (1 - \frac{\eta}{2} \geq e^{-\eta}, \forall \eta : 0 < \eta \leq 0.5)\end{aligned}$$

Setting $T' = \min\{\lfloor \frac{\ln 100}{\eta} \rfloor + 1, T\} \leq T$, we have

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\pi_{t,2}] \geq \sum_{t=1}^{T} \frac{1}{2}e^{-\eta(t-1)} &\geq \sum_{t=1}^{T'} \frac{1}{2}e^{-\eta(t-1)} \\ &\geq \sum_{t=1}^{T'} \frac{1}{200} = \frac{T'}{200} \geq \frac{1}{200}\min\{\frac{\ln 100}{\eta}, T\}.\end{aligned} \tag{22}$$

Now, by using (22), we can further lower bound (19) to get

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}[\tilde{\pi}_{t,2}] &= (1-\gamma)\sum_{t=1}^{T} \mathbb{E}[\pi_{t,2}] + \frac{\gamma}{2}T \\ &\geq (1-\gamma)\left(\frac{1}{200}\min\left\{\frac{\ln 100}{\eta}, T\right\}\right) + \gamma\frac{T}{2} && \text{(from (22))} \\ &\geq \min\left\{\frac{1}{200}\min\left\{\frac{\ln 100}{\eta}, T\right\}, \frac{T}{2}\right\} && (\gamma\alpha + (1-\gamma)\beta \geq \min\{\alpha, \beta\}) \\ &= \frac{1}{200}\min\left\{\frac{\ln 100}{\eta}, T\right\} \\ &\geq \frac{1}{200}\min\left\{\frac{1}{\eta}, T\right\} \\ &\geq \frac{1}{200}T^{2/3} && (\eta > T^{-2/3}) \\ &= cT^{2/3}.\end{aligned}$$

□

We now present the next lemma.

Lemma 17. *For WSU-UX run with any $\gamma > T^{-1/3}$ and any $\eta \geq 0$ there exists a loss sequence $\{\ell_t\}_{t=1}^T$ and $c = \frac{1}{2} > 0$ such that*

$$\mathbb{E}\left[\mathcal{R}_T\right] \geq cT^{2/3}.$$

Proof. Consider the same loss sequence as Lemma 16 with two arms: $\{\ell_t\}_{t=1}^T$ where $\ell_{t,1} = 0, \ell_{t,2} = 1$ for $1 \leq t \leq T$. The best arm is arm 1. Since we have a uniform exploration of γ, in any round $t \in [T]$, we pick arm 2 with probability $\tilde{\pi}_{t,2} = (1-\gamma)\pi_{t,2} + \gamma\frac{1}{2} \geq \gamma\frac{1}{2}$. Therefore, the algorithm incurs at least $\gamma\frac{1}{2}$ loss for each round. Hence

$$\mathbb{E}\left[\mathcal{R}_T\right] = \sum_{t=1}^{T} \mathbb{E}\left[\tilde{\pi}_{t,2}\right] \geq \sum_{t=1}^{T} \frac{\gamma}{2} = \frac{\gamma T}{2} > cT^{2/3},$$

where the last inequality holds because $\gamma > T^{-1/3}$. □

A.2 Non-trivial Case

For the non-trivial case, we have the following theorem as stated in the main text.

Theorem 8. *For $K = 2$ and for any valid set of (η, γ) in the non-trivial case, there exists T_0 such that for any $T \geq T_0$, we have a loss sequence $\{\ell_t\}_{t=1}^T$ such that*

$$\mathbb{E}[\mathcal{R}_T] = \Omega(T^{2/3}). \tag{8}$$

Note that in the non-trivial case, we always consider the loss sequence defined in Definition 9. For the convenience of the reader, we recall that for any T, we define $\{\ell_t\}_{t=1}^T$ as

$$\{\ell_t\}_{t=1}^T = \begin{cases} \ell_{t,1} = 1, \ell_{t,2} = 0 & \text{for } 1 \leq t \leq \frac{T}{100} \\ \ell_{t,1} = 0, \ell_{t,2} = 1 & \text{for } \frac{T}{100} < t \leq T \end{cases}.$$

To prove Theorem 8, by Lemma 18 we first observe that any bound of the form $c_1 \frac{1}{\eta} + c_2 \frac{\eta K T}{\gamma} + c_3 \gamma T$ can be lower bounded by $c_4 K^{\frac{1}{3}} T^{\frac{2}{3}}$.

Lemma 18. *For any choice of $c_1, c_2, c_3 > 0$, and any valid choice of $\gamma, \eta > 0$, there exists $c_4 > 0$ such that*

$$c_1 \frac{1}{\eta} + c_2 \frac{\eta K T}{\gamma} + c_3 \gamma T \geq c_4 K^{\frac{1}{3}} T^{\frac{2}{3}} = \Omega(T^{2/3}).$$

Proof. Observe that $c_1 \frac{1}{\eta} + c_2 \frac{\eta K T}{\gamma} \geq 2\sqrt{\frac{c_1}{\eta} \frac{c_2 \eta K T}{\gamma}}$. Therefore, $c_1 \frac{1}{\eta} + c_2 \frac{\eta K T}{\gamma} + c_3 \gamma T \geq 2\sqrt{\frac{c_1 c_2 K T}{\gamma}} + c_3 \gamma T$. Define $f(\gamma) := 2\sqrt{\frac{c_1 c_2 K T}{\gamma}} + c_3 \gamma T$ for $\gamma > 0$. Note that since f is convex, the minimum is attained at γ^* where $f'(\gamma^*) = 0$. It is easy to see that $\gamma^* = c_3^{-\frac{2}{3}} \cdot (\frac{c_1 c_2 K}{T})^{\frac{1}{3}}$ and therefore, we get

$$\begin{aligned} f(\gamma) &\geq f(\gamma^*) \\ &= f(c_3^{-\frac{2}{3}} \cdot (\frac{c_1 c_2 K}{T})^{\frac{1}{3}}) \\ &= 2\sqrt{\frac{c_1 c_2 K T}{c_3^{-\frac{2}{3}} \cdot (\frac{c_1 c_2 K}{T})^{\frac{1}{3}}}} + c_3^{\frac{1}{3}} (c_1 c_2 K)^{\frac{1}{3}} T^{\frac{2}{3}} \\ &= 2\sqrt{(c_1 c_2 c_3 K)^{\frac{2}{3}} T^{\frac{4}{3}}} + (c_1 c_2 c_3 K)^{\frac{1}{3}} T^{\frac{2}{3}} \\ &= 2(c_1 c_2 c_3 K)^{\frac{1}{3}} T^{\frac{2}{3}} + (c_1 c_2 c_3 K)^{\frac{1}{3}} T^{\frac{2}{3}} \\ &= 3(c_1 c_2 c_3 K)^{\frac{1}{3}} T^{\frac{2}{3}}. \end{aligned}$$

Therefore, there exists $c_4 = 3(c_1 c_2 c_3)^{\frac{1}{3}} > 0$ such that $c_1 \frac{1}{\eta} + c_2 \frac{\eta K T}{\gamma} + c_3 \gamma T \geq c_4 K^{\frac{1}{3}} T^{\frac{2}{3}}$. □

We will then show that for large enough T, the regret on this particular loss sequence can be lower bounded as follows.

Theorem 10. *When running WSU-UX for any valid choice of (η, γ) in the non-trivial case, there exists T_0 such that for any $T \geq T_0$, for loss sequence $\{\ell_t\}_{t=1}^T$, we have for some constants $c_1, c_2, c_3 > 0$,*

$$\mathbb{E}[\mathcal{R}_T] \geq c_1 \frac{1}{\eta} + c_2 \frac{\eta T K}{\gamma} + c_3 \gamma T. \tag{9}$$

Theorem 10 along with Lemma 18 proves Theorem 8.

A.2.1 Proving Theorem 10 using Claims

In this section, we show Theorem 10 given Claim 1, 2, and 3. We first restate and prove the second-order lower bound lemma.

Lemma 11 (Second-Order Lower Bound)**.** *For any valid choice of (η, γ) when running WSU-UX on loss sequence $\{\ell_t\}_{t=1}^T$, we get*

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j}-\sum_{t=1}^{T}\hat{\ell}_{t,1}\\ &\geq \frac{\ln\pi_{T+1,1}+\ln K}{\eta}+\frac{\eta}{4}\sum_{t=1}^{T}(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})^2. \qquad (10)\end{aligned}$$

Proof of Lemma 11. We can express $\pi_{T+1,1}$ as follows.

$$\pi_{T+1,1}=\frac{1}{K}\prod_{t=1}^{T}(1-\eta(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})).$$

Taking the logarithm of both sides, we get

$$\ln(\pi_{T+1,1})=-\ln K+\sum_{t=1}^{T}\ln\left(1-\eta(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})\right). \qquad (23)$$

Next, we observe that for any fixed $t\in[T]$, we have $-1\leq\eta(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})\leq 1/2$ (by Lemma 20 below). Note that for $-1\leq x\leq 1/2$ we have $\ln(1-x)\leq -x-x^2/4$ (by Lemma 19 below). Therefore, we can show

$$\ln\left(1-\eta(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})\right)\leq -\eta(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})-\frac{\eta^2}{4}(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})^2$$

Taking the summation over T rounds and combining with (23), we get

$$\ln\pi_{T+1,1}\leq -\ln K-\eta\sum_{t=1}^{T}(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})-\frac{\eta^2}{4}\sum_{t=1}^{T}(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})^2.$$

Rearranging and dividing by η, we get

$$\sum_{t=1}^{T}\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j}-\hat{\ell}_{t,1}\geq\frac{\ln\pi_{T+1,1}+\ln K}{\eta}+\frac{\eta}{4}\sum_{t=1}^{T}(\hat{\ell}_{t,1}-\sum_{j\in[K]}\pi_{t,j}\hat{\ell}_{t,j})^2.$$

□

The previous proof used the following two simple lemmas.

Lemma 19. $\ln(1-x)\leq -x-x^2/4$ *when* $-1\leq x\leq 1/2$.

Proof. Let $f(x)=\ln(1-x)+x+x^2/4$. Then $f'(x)=-1/(1-x)+1+x/2=\frac{-x(1+x)}{2-2x}$. Therefore $f'(x)\geq 0$ for $-1\leq x\leq 0$, and $f'(x)<0$ for $0<x\leq 1/2$. The maximum is attained at $x=0$. As a result, $f(x)\leq f(0)=0$ when $-1<x\leq 1/2$. □

Lemma 20. *For* $0\leq\eta,\gamma\leq 1/2$ *where* $\frac{\eta K}{\gamma}\leq 1/2$, *and loss sequence* $\{\ell_t\}_{t=1}^T$, *for any round* $t\in[T]$, *we have*

$$-1\leq\eta\left(\hat{\ell}_{t,1}-\sum_{j\in\{1,2\}}\pi_{t,j}\hat{\ell}_{t,j}\right)\leq 1/2.$$

Proof. For $1 \leq t \leq \frac{T}{100}$, we have $\ell_{t,2} = 0$, and hence $\hat{\ell}_{t,2} = 0$. Therefore, we have

$$\begin{aligned}\eta\left(\hat{\ell}_{t,1} - \sum_{j\in\{1,2\}} \pi_{t,j}\hat{\ell}_{t,j}\right) &= \eta\left(\hat{\ell}_{t,1} - \pi_{t,1}\hat{\ell}_{t,1}\right) \\ &= \eta\ell_{t,1}\frac{1-\pi_{t,1}}{\tilde{\pi}_{t,1}}\mathbb{1}[I_t = 1].\end{aligned}$$

Observe that

$$\begin{aligned}-1 < 0 &\leq \eta\,\ell_{t,1}\frac{1-\pi_{t,1}}{\tilde{\pi}_{t,1}}\mathbb{1}[I_t = 1] \\ &\leq \eta\frac{1}{\tilde{\pi}_{t,1}} \\ &\leq \eta\frac{K}{\gamma} && (\tilde{\pi}_{t,1} \geq \frac{\gamma}{K}) \\ &\leq 1/2.\end{aligned}$$

For $\frac{T}{100} < t \leq T$, we have $\ell_{t,1} = 0$, and hence $\hat{\ell}_{t,1} = 0$. Therefore

$$\begin{aligned}\eta\,(\hat{\ell}_{t,1} - \sum_{j\in\{1,2\}} \pi_{t,j}\hat{\ell}_{t,j}) &= \eta\,(-\pi_{t,2}\hat{\ell}_{t,2}) \\ &= -\eta\,\ell_{t,2}\frac{\pi_{t,2}}{\tilde{\pi}_{t,2}}\mathbb{1}[I_t = 2] \\ &= -\eta\frac{\pi_{t,2}}{\pi_{t,2}(1-\gamma) + \frac{\gamma}{K}}\ell_{t,2}\mathbb{1}[I_t = 2].\end{aligned}$$

Now, we can simply show

$$\begin{aligned}1/2 \geq 0 &\geq -\eta\frac{\pi_{t,2}}{\pi_{t,2}(1-\gamma) + \frac{\gamma}{K}}\ell_{t,2}\,\mathbb{1}[I_t = 2] \\ &\geq -\eta\frac{\pi_{t,2}}{\pi_{t,2}(1-\gamma)}\ell_{t,2}\,\mathbb{1}[I_t = 2] \\ &\geq -\frac{\eta}{1-\gamma} \\ &\geq -\frac{\eta}{1/2} && (\gamma \leq 1/2) \\ &\geq -1. && (\eta \leq 1/2)\end{aligned}$$

□

Then, as mentioned in the main part of the paper, we can use Claims 1, 2, and 3 to convert the RHS of (10) to the lower bound in (9).

B Proof of claims

In this part, we prove the claims.

B.1 Proof of Claim 2

We first restate Claim 2.

Claim 2 (Second moment lower bound)**.** *For large enough T, there exists c_2 such that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{\ell}_{t,1} - \sum_{j\in[1,2]} \pi_{t,j}\hat{\ell}_{t,j}\right)^2\right] \geq 4c_2\frac{TK}{\gamma}.$$

In order to prove Claim 2, we need to introduce the following lemma.

Lemma 21. *When running WSU-UX on loss sequence $\{\ell_t\}_{t=1}^T$ and hyperparameter defined in the non-trivial case and for large enough T, for $\frac{T_1}{2} \le t \le T_1$, we have*

$$\mathbb{E}[\pi_{t,1}] \le \frac{1}{KT}.$$

Proof. According to Lemma 12, which can be found in Appendix A.1, we have $\mathbb{E}\left[\pi_{t+1,1} \mid \mathcal{F}_{t-1}\right] = (1 - C_t)\,\pi_{t,1} + C_t\,\pi_{t,1}^2$ where $C_t = \eta\,(\ell_{t,1} - \ell_{t,2})$. For t in $\frac{T_1}{2} \le t \le T_1$, we have $\pi_{t,1} \le 1/2$ and $C_t = \eta$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\pi_{t+1,1} \mid \mathcal{F}_{t-1}\right] &= (1-\eta)\,\pi_{t,1} + \eta\,\pi_{t,1}^2 \\
&\le (1-\eta)\,\pi_{t,1} + \frac{\eta}{2}\,\pi_{t,1} \\
&= (1 - \frac{\eta}{2})\,\pi_{t,1}.
\end{aligned}$$

Taking the expectation over all possible $\mathcal{F}_{t-1}$, we get

$$\mathbb{E}[\pi_{t+1,1}] \le (1 - \frac{\eta}{2})\,\mathbb{E}[\pi_{t,1}].$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\pi_{1,\frac{T_1}{2}}\right] = \mathbb{E}\left[\pi_{1,1}\right] \prod_{s=1}^{\lceil \frac{T_1}{2} \rceil - 1} \frac{\mathbb{E}\left[\pi_{s+1,1}\right]}{\mathbb{E}\left[\pi_{s,1}\right]} &\le \frac{1}{2}(1 - \frac{\eta}{2})^{\frac{T_1}{2}} \\
&\le \frac{1}{2}(e^{-\frac{\eta}{2}})^{\frac{T}{200}} && (1 + x \le e^x, T_1 = \frac{T}{100}) \\
&\le \frac{1}{2}\exp\left(-\frac{1}{400}T^{\frac{1}{3}}\right). && (\eta \ge T^{-2/3})
\end{aligned}$$

Since we are in the non-trivial case and $\eta \ge T^{-2/3}$, we have $\eta T \ge T^{1/3}$. Since $\frac{1}{2}e^{-\frac{T^{1/3}}{400}}$ converges to zero exponentially, whereas $\frac{1}{KT}$ convergence to zero at a slower rate, we can say for large enough T that $\frac{1}{2}e^{-\frac{\eta T}{400}} \le \frac{1}{2}e^{-\frac{T^{1/3}}{400}} \le \frac{1}{KT}$. □

Now, we are ready to prove Claim 2.

Proof of Claim 2. For $\frac{T_1}{2} \le t \le T_1$, we have $\ell_{t,2} = 0$, therefore $\hat{\ell}_{t,2} = 0$. Now we can lower bound

$\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{\ell}_{t,1}-\sum_{j\in[1,2]}\pi_{t,j}\hat{\ell}_{t,j}\right)^2\right]$ as follows:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{\ell}_{t,1}-\sum_{j\in[1,2]}\pi_{t,j}\hat{\ell}_{t,j}\right)^2\right] &\geq \mathbb{E}\left[\sum_{t=T/100}^{T/200}\left(\hat{\ell}_{t,1}-\sum_{j\in[1,2]}\pi_{t,j}\hat{\ell}_{t,j}\right)^2\right]\\
&= \mathbb{E}\left[\sum_{t=T/200}^{T/100}\left(\hat{\ell}_{t,1}-\pi_{t,1}\hat{\ell}_{t,1}\right)^2\right]\\
&= \mathbb{E}\left[\sum_{t=T/200}^{T/100}\left(1-\pi_{t,1}\right)^2\hat{\ell}_{t,1}^2\right]\\
&\geq \mathbb{E}\left[\sum_{t=T/200}^{T/100}(1-1/2)^2\hat{\ell}_{t,1}^2\right] && \left(\forall t, \frac{T}{200}\leq t\leq\frac{T}{100}:\pi_{t,1}\leq 1/2\right)\\
&= \frac{1}{4}\sum_{t=T/200}^{T/100}\mathbb{E}\left[\left(\frac{\ell_{t,1}}{\tilde{\pi}_{t,1}}\mathbb{1}\left[I_t=1\right]\right)^2\right]\\
&= \frac{1}{4}\sum_{t=T/200}^{T/100}\mathbb{E}\left[\mathbb{E}_{t-1}\left[\left(\frac{\ell_{t,1}}{\tilde{\pi}_{t,1}}\right)^2\left(\mathbb{1}\left[I_t=1\right]\right)^2\right]\right]\\
&= \frac{1}{4}\sum_{t=T/200}^{T/100}\mathbb{E}\left[\left(\frac{1}{\tilde{\pi}_{t,1}}\right)^2\mathbb{E}_{t-1}\left[\left(\mathbb{1}\left[I_t=1\right]\right)^2\right]\right] && \left(\forall t, \frac{T}{200}\leq t\leq\frac{T}{100}:\ell_{t,1}=1\right)\\
&= \frac{1}{4}\sum_{t=T/200}^{T/100}\mathbb{E}\left[\frac{1}{\tilde{\pi}_{t,1}}\right]\\
&\geq \frac{1}{4}\sum_{t=T/200}^{T/100}\frac{1}{\mathbb{E}\left[\tilde{\pi}_{t,1}\right]},
\end{aligned}
$$

where the last inequality comes from applying Jensen's inequality $\mathbb{E}\left[\frac{1}{X}\right]\geq\frac{1}{\mathbb{E}[X]}$. Note that for large enough T, for $\frac{T}{200}\leq T\leq\frac{T}{100}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\pi}_{1,t}\right] &= \mathbb{E}\left[\pi_{1,t}(1-\gamma)+\frac{\gamma}{K}\right]\\
&\leq \mathbb{E}\left[\pi_{1,t}\right]+\frac{\gamma}{K}\\
&\leq \frac{1}{KT}+\frac{\gamma}{K} && \text{(Lemma 21)}\\
&\leq \frac{2\gamma}{K}, && \left(\text{since } \frac{1}{T}\leq\gamma\right)
\end{aligned}
$$

where the last inequality comes from the fact that we have $\frac{\eta K}{\gamma}\leq 1/2$ for WSU-UX. This implies $\gamma\geq 2\eta K\geq 2KT^{-2/3}\geq\frac{1}{T}$, where we use the fact that $\eta\geq T^{-2/3}$. As a result, we get

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{\ell}_{t,1}-\sum_{j\in[1,2]}\pi_{t,j}\hat{\ell}_{t,j}\right)^2\right] &\geq \frac{1}{4}\sum_{t=T/200}^{T/100}\frac{1}{\mathbb{E}[\tilde{\pi}_{t,1}]}\\
&\geq \frac{1}{4}\sum_{t=T/200}^{T/100}\frac{K}{2\gamma}\\
&\geq \frac{1}{4}\left(\frac{T}{100}-\frac{T}{200}\right)\frac{K}{2\gamma}=\frac{1}{1600}\frac{TK}{\gamma}.
\end{aligned}
$$

Therefore, for $c_2 = \frac{1}{4 \cdot 1600}$, Claim 2 holds. □

B.2 Proof Sketch of Claims 1 and 3

In this subsection, we prove Claims 1 and 3 using several technical lemmas without stating their proof. We will prove all the technical lemmas in the next subsection.

We first recall the notion of phases here.

Definition 13. Define $T_1 = \frac{1}{100}T, T_2 = \frac{2}{10}T, T_3 = \frac{1}{10}T, T_4 = \frac{69}{100}T$, and then define (sub-)phases as follows:

- Phase 1: $\mathcal{T}_1 = \{t : 1 \le t \le T_1\}$,
- Phase 2.1: $\mathcal{T}_2 = \{t : T_1 + 1 \le t \le T_1 + T_2\}$,
- Phase 2.2: $\mathcal{T}_3 = \{t : T_1 + T_2 + 1 \le t \le T_1 + T_2 + T_3\}$,
- Phase 2.3: $\mathcal{T}_4 = \{t : T_1 + T_2 + T_3 + 1 \le t \le T_1 + T_2 + T_3 + T_4\}$.

We recall the definitions of M and T' as well.

Definition 14. We define M and T' as follows

$$M := \frac{1}{\ln 2}\left[\underbrace{\ln\left(\frac{2K}{\gamma}\right)}_{\propto(\ln T)} + \underbrace{2(1+\varepsilon_1)(1+\frac{\eta K}{\gamma})\eta T_1}_{\propto(\eta T_1)}\right] \tag{15}$$

$$T' := \frac{1}{1-\varepsilon_2}\,(\frac{4}{3-\gamma})\,\frac{2}{\eta}\,M \tag{16}$$

where $\varepsilon_1 = \sqrt{\frac{6\ln T}{\frac{2\gamma}{K}T_1}}$ and $\varepsilon_2 = \sqrt{\frac{4\ln T}{\frac{3-\gamma}{4}T_2}}$.

Next, we restate two events $\mathcal{E}_1$ and $\mathcal{E}_2$.

Definition 22. Let

$$\mathcal{E}_1 = \{\pi_{T_1+1,1} \ge 2^{-M}\}$$

be the event that arm 1's probability at the end of Phase 1 is not too small, where M is defined in (15).

Definition 23. Let $\mathcal{E}_2 = \{\pi_{T_1+T_2+1,1} \ge \frac{1}{4}\}$ be the event that arm 1's probability at the end of Phase 2.2 has recovered to $\frac{1}{4}$.

Next, we have the following lemma stating that, with high probability, $\pi_{T_1+1,1}$ is not too small.

Lemma 24. *When we run WSU-UX with any valid parameters η, γ on specific loss sequence $\{\ell_t\}_{t=1}^T$, for any (ε, δ), where $\varepsilon = \sqrt{\frac{3\ln\frac{1}{\delta}}{\frac{2\gamma}{K}T_1}} \in (0,1]$, at the end of phase 1, we have that with probability at least $1-\delta$,*

$$\pi_{T_1+1,1} \ge 2^{-\mathcal{M}},$$

where $\mathcal{M} = \frac{1}{\ln 2}\left(2(1+\varepsilon)(1+\frac{\eta K}{\gamma})\eta T_1 + \ln\frac{2K}{\gamma}\right)$. In particular, for large enough T, by choosing $\delta = \frac{1}{T^2}$, we get $\pi_{T_1+1,1} \ge 2^{-M}$ where M is defined in (15).

We used a recently developed multiplicative form of Azuma's inequality for martingales (Kuszmaul and Qi, 2021) to show Lemma 24. This lemma shows that when T is large enough, with high probability, $\pi_{T_1+1,1}$ does not become too small, i.e., Event $\mathcal{E}_1$ happens. Next, we will show that $\pi_{T_1+T_2+1,1}$ recovers to $1/4$ with high probability. To show this, observe that Phase 2.1 is the phase where $\pi_{t,1}$ can start to recover. By each pull of arm 2 in Phase 2.1, the probability $\pi_{t,1}$ increases, whereas by pulls of arm 1, $\pi_{t,1}$ does not change. Hence, we first find an upper bound on the number of pulls of arm 2 needed so that $\pi_{t,1}$ recovers to $1/4$. At the beginning of Phase 2.1, after each pull of arm 2, the rate of update $\frac{\pi_{t+1,1}}{\pi_{t,1}} \approx 1+\varepsilon$ is very close to 1. Therefore, we first focus on finding an upper bound on *the number of pulls of arm 2 needed for $\pi_{t,1}$ to double.*

Lemma 25. *Consider a round $t_0 > T_1$, where $0 < \pi_{t_0,1} \leq \frac{1}{4}$. If arm 2 is pulled m times in rounds $t_1, \dots, t_m$ where $t_0 \leq t_1 < t_2 < \dots < t_m \leq T$ and $m \geq \frac{2}{\eta} \cdot \frac{1}{1-2\pi_{t_0,1}}$, then we have $\pi_{t_m+1,1} \geq 2\pi_{t_0,1}$.*

Next, Lemma 26 is a cumulative version of Lemma 25 where we show an upper bound on the total number of pulls of arm 2 needed so that $\pi_{t,1}$ doubles for $M-2$ times.

Lemma 26. *Consider a round $t > T_1$ where we have $2^{-m} \leq \pi_{t,1} \leq 2^{-(m-1)}$, where $m \in \mathbb{Z}_+$. Then if arm 2 is pulled $k = \frac{2}{\eta} m$ times in rounds $t, t+1, \dots, T_1$, and round t' denotes the round just after the k^{th} pull, then we have $\pi_{t',1} \geq \frac{1}{4}$.*

Lemma 26 indicates that given Event $\mathcal{E}_1$ happened, $k = \frac{2}{\eta} M$ pulls of arm 2 suffice to ensure that $\pi_{t,1} \geq 1/4$. The next lemma shows that given Event $\mathcal{E}_1$ happened, then with high probability, arm 2 is going to be picked in T_2 rounds at least k times. This along with Lemma 26 implies that $\pi_{T_1+T_2+1,1}$ recovers.

Lemma 27. *When T is large enough, with probability at least $1 - 2\frac{1}{T^2}$, Event $\mathcal{E}_2$ happens. i.e $\pi_{T_1+T_2+1,1} \geq 1/4$.*

We next show that given $\mathcal{E}_2$, the expectation of $\pi_{t,2}$ goes to 0 exponentially quickly as t increments beyond $T_1 + T_2$.

Lemma 28. *Assume Event $\mathcal{E}_2$ happens. Define $t_0 = T_1 + T_2 + 1$, and $1 \leq \tau \leq T_3 + T_4$, and time step $t = t_0 + \tau$. Then we have*

$$\mathbb{E}[\pi_{t,2} | \mathcal{E}_2] \leq \frac{3}{4} \exp{(-\frac{\eta}{4}\tau)}.$$

Now, conditional on Event $\mathcal{E}_2$, by using Chebyshev's inequality and Lemma 28, we show that it is very unlikely that $\pi_{T,1}$ is constantly smaller than 1.

Lemma 29. *Condition on Event $\mathcal{E}_2$ defined in Definition 23, we have*

$$\Pr\left(\pi_{T+1,1} \leq \frac{3}{4} \,\middle|\, \mathcal{E}_2\right) \leq \frac{75}{4} e^{-c\eta T}.$$

Using conditional expectation and Lemma 29, we prove Claim 1 in the next subsection.

Next, we introduce the following lemma.

Lemma 30. *For large enough T, we have $\mathbb{E}\left[\pi_{t,2}\right] \leq \frac{1}{4}$ for $T_1 + T_2 + T_3 < t \leq T$.*

We prove Claim 3 using Lemma 30 in Appendix B.3.

B.3 Complete Proof of Claims 1 and 3

B.3.1 High Probability Lemma for Event $\mathcal{E}_1$

Proof of Lemma 24. We define τ to be the largest t in $1 \leq t \leq T_1$ such that $\pi_{t,1} > \frac{\gamma}{K}$. Since $\pi_{t,1}$ are random variables that depend on internal random bits of the algorithm, τ is also a random variable. Observe that for the loss sequence $\{\ell_t\}_{t=1}^{T}$ we are considering, for all $1 \leq t \leq T_1 + 1$ we have $\pi_{t,1} \geq \pi_{t+1,1}$. This implies that

$$\pi_{t,1} > \frac{\gamma}{K} \qquad \forall t : 1 \leq t \leq \tau \tag{24}$$

$$\pi_{t,1} \leq \frac{\gamma}{K} \qquad \forall t : \tau + 1 \leq t \leq T_1. \tag{25}$$

Using τ, we can express $\pi_{T_1+1,1}$ as follows[8]:

$$\pi_{T_1+1,1} = \pi_{1,1} \prod_{s=1}^{T_1} \frac{\pi_{s+1,1}}{\pi_{s,1}}$$
$$= \underbrace{\left(\pi_{1,1} \prod_{s=1}^{\tau-1} \frac{\pi_{s+1,1}}{\pi_{s,1}}\right)}_{\text{first term}} \underbrace{\left(\frac{\pi_{\tau+1,1}}{\pi_{\tau,1}}\right)}_{\text{second term}} \underbrace{\left(\prod_{s=\tau+1}^{T_1} \frac{\pi_{s+1,1}}{\pi_{s,1}}\right)}_{\text{third term}}. \tag{26}$$

Clearly, the first term in (26) can be lower bounded as

$$\pi_{1,1} \prod_{s=1}^{\tau-1} \frac{\pi_{s+1,1}}{\pi_{s,1}} = \pi_{\tau-1,1} > \frac{\gamma}{K}, \tag{27}$$

where we used (24). Next, observe that for rounds $1 \le t \le T_1$, in phase one, we can simply lower bound $\frac{\pi_{s+1,1}}{\pi_{s,1}}$ as follows:

$$\begin{aligned}\frac{\pi_{s+1,1}}{\pi_{s,1}} &= 1 - \eta\Big(\hat{\ell}_{s,1} - \sum_{j\in[2]} \pi_{s,j}\hat{\ell}_{s,j}\Big) \\ &= 1 - \eta\frac{1-\pi_{s,1}}{(1-\gamma)\pi_{s,1} + \frac{\gamma}{K}} \mathbb{1}[I_s = 1] \\ &\ge 1 - \eta\frac{1-0}{0+\frac{\gamma}{K}} \mathbb{1}[I_s = 1] && (\pi_{s,1} \ge 0) \\ &= 1 - \frac{\eta K}{\gamma} \mathbb{1}[I_s = 1] \\ &= (1 - \frac{\eta K}{\gamma})^{\mathbb{1}[I_s=1]}. \end{aligned} \tag{28}$$

Therefore, the second term can be lower bounded by

$$\underbrace{\frac{\pi_{\tau+1,1}}{\pi_{\tau,1}}}_{\text{second term}} \ge (1 - \frac{\eta K}{\gamma})^{\mathbb{1}[I_\tau=1]} \ge 1/2, \tag{29}$$

since $\frac{\eta K}{\gamma} \le 1/2$ in WSU-UX. Moreover, we have

$$\begin{aligned}\underbrace{\prod_{s=\tau+1}^{T_1} \frac{\pi_{s+1,1}}{\pi_{s,1}}}_{\text{third term}} &\ge \prod_{s=\tau+1}^{T_1} (1 - \frac{\eta K}{\gamma})^{\mathbb{1}[I_s=1]} && \text{(by 28)} \\ &= (1 - \frac{\eta K}{\gamma})^{\sum_{s=\tau+1}^{T_1} \mathbb{1}[I_s=1]}. \end{aligned} \tag{30}$$

Therefore, by plugging (27), (29), and (30) into the right hand side of (26) we obtain

$$\pi_{T_1+1,1} \ge \frac{\gamma}{2K}(1 - \frac{\eta K}{\gamma})^{\sum_{s=\tau+1}^{T_1} \mathbb{1}[I_s=1]}, \tag{31}$$

where the right-hand side is a random variable that depends on τ. Note that $\mathbb{1}[\tau = t' | \mathcal{F}_{t'}]$ is measurable, meaning that in round t', given access to the past history, we can deterministically tell whether $\tau = t'$ or not. Next, we

[8] For convention, the product over an empty set is assumed to be 1. e.g. if $\tau = 1$, then the first term in the right-hand side of (26) which is a product over the empty set is assumed to be 1. Similarly, if $\tau = T_1$, the third term in the right-hand side of (26) is assumed to be 1.

will show that for any possible value t' that τ can take, and particular history $\mathcal{F}_{t'}$ up until the end of round t' for which $\tau = t'$, [9] there is a suitable upper bound on the following term

$$\sum_{s=t'+1}^{T_1} \mathbb{1}[I_s = 1 \mid \mathcal{F}_{t'}] \tag{32}$$

that holds with probability at least $1-\delta$. This implies a uniform high probability lower bound on (31), which completes the proof.

It remains to show this uniform upper bound for (32). In order to show this, we set up a martingale. In particular, we define $I_{t,i} := \mathbb{1}[I_t = i]$. Consider any fixed history up until round t' denoted by $\mathcal{F}_{t'}$ such that $\tau = t'$. Then for any t where $t'+1 \le t \le T_1$, we have

$$\begin{aligned}
\mathbb{E}\left[I_{t,1} \,|\, \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[(1-\gamma)\pi_{t,1} + \frac{\gamma}{K} \,\middle|\, \mathcal{F}_{t-1}\right] && \text{(Uniform Exploration by WSU-UX)}\\
&\le \mathbb{E}\left[\pi_{t,1} + \frac{\gamma}{K} \,\middle|\, \mathcal{F}_{t-1}\right]\\
&\le \frac{2\gamma}{K} =: q, && \text{(from 25)} \qquad (33)
\end{aligned}$$

where the last inequality comes from (24). Define $Z_{t'} := 0$ and for any t where $t'+1 \le t \le T_1$, let $W_t := I_{t,1} - \mathbb{E}[I_{t,1} \mid \mathcal{F}_{t-1}]$ and $Z_t := \sum_{s=t'+1}^{t} W_s$. Observe that for t in $t'+1 \le t \le T_1$, we have

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = Z_{t-1} + \mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = Z_{t-1};$$

therefore $(Z_t)_{t\in\{t',\dots,T_1\}}$ is a martingale. Moreover, for t in $t'+1 \le t \le T_1$, we have $Z_t - Z_{t-1} = W_t \in [-A_t, B_t]$ for $A_t = \mathbb{E}\left[I_{t,1} \mid \mathcal{F}_{t-1}\right]$ and $B_t = 1 - A_t$, simply because $I_{t,1} \in \{0,1\}$. Observe that we have

$$\sum_{t=t'+1}^{T_1} A_t = \sum_{t=t'+1}^{T_1} \mathbb{E}\left[I_{t,1} \,|\, \mathcal{F}_{t-1}\right] \le \sum_{t=t'+1}^{T_1} q \le \sum_{t=1}^{T_1} q = qT_1 = \frac{2\gamma}{K} T_1 =: \mu,$$

where the first inequality comes from (33). We define $c := A_t + B_t = 1$ and apply Theorem 10 of Kuszmaul and Qi (2021) to get that for all $\varepsilon > 0$,

$$\Pr\left(Z_{T_1} - Z_{t'} \ge \varepsilon\mu \,|\, \mathcal{F}_{t'}\right) \le \exp\left(-\frac{\varepsilon^2 \mu}{(2+\varepsilon)c}\right).$$

Using our definition of the martingale sequence, and noting that $c = 1$, we get

$$\Pr\left(\sum_{t=t'+1}^{T_1} I_{t,1} \ge \sum_{t=t'+1}^{T_1} \mathbb{E}\left[I_{t,1} \,|\, \mathcal{F}_{t-1}\right] + \varepsilon\mu \,\middle|\, \mathcal{F}_{t'}\right) \le \exp\left(-\frac{\mu\varepsilon^2}{2+\varepsilon}\right).$$

Using $\sum_{t=t'+1}^{T_1} \mathbb{E}\left[I_{t,1} \mid \mathcal{F}_{t-1}\right] \le \mu$ and by imposing the restriction $\varepsilon \le 1$, we have for $\varepsilon \in (0,1]$,

$$\Pr\left(\sum_{t=t'+1}^{T_1} I_{t,1} \ge (1+\varepsilon)\mu \,\middle|\, \mathcal{F}_{t'}\right) \le \exp\left(-\frac{\mu\varepsilon^2}{3}\right).$$

Equivalently, conditional on $\mathcal{F}_{t'}$, with probability at least $1-\delta$, where $\varepsilon = \sqrt{\frac{3\ln\frac{1}{\delta}}{\frac{2\gamma}{K}T_1}} \in (0,1]$, we have

$$\begin{aligned}
\sum_{t=t'+1}^{T_1} I_{t,1} &\le (1+\varepsilon)\mu\\
&= (1+\varepsilon)\frac{2\gamma}{K} T_1. \qquad (34)
\end{aligned}$$

[9] Observe that this t' is well defined since $\mathcal{F}_{t'-1}$, the history up until round $t'-1$, is enough to determine whether $\tau = t'$ or $\tau \ne t'$.

This is the suitable upper bound we wanted for the quantity in (32). In particular, we have

$$\begin{aligned}
(1-\frac{\eta K}{\gamma})^{\left[\sum_{t=t'+1}^{T_1} I_t\right]} &\geq \left(e^{(-\frac{\eta K}{\gamma})(1+\frac{\eta K}{\gamma})}\right)^{\left[\sum_{t=t'+1}^{T_1} I_t\right]} && (1-x \geq e^{(-x-x^2)} \text{ for } 0 < x \leq \frac{1}{2}) \\
&\geq \left(e^{(-\frac{\eta K}{\gamma})(1+\frac{\eta K}{\gamma})}\right)^{(1+\varepsilon)\frac{2\gamma}{K}T_1} && \text{from (34)} \\
&= \exp\left(-2(1+\varepsilon)(1+\frac{\eta K}{\gamma})\eta T_1\right). && (35)
\end{aligned}$$

Combining this lower bound with (31), we get the following statement.

For any possible value for τ and any fixed $\mathcal{F}_{t'}$ satisying $\tau = t'$, for any $\varepsilon = \sqrt{\frac{3\ln\frac{1}{\delta}}{\frac{2\gamma}{K}T_1}} \in (0,1]$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\pi_{T_1+1,1} = \frac{\gamma}{2K}(1-\frac{\eta K}{\gamma})^{\left[\sum_{s=\tau+1}^{T_1}\mathbb{1}[I_s=1]\right]} &\geq \frac{\gamma}{2K}\exp\left(-2(1+\varepsilon)(1+\frac{\eta K}{\gamma})\eta T_1\right) \\
&= \exp\left(-\left(2(1+\varepsilon)(1+\frac{\eta K}{\gamma})\eta T_1 + \ln\frac{2K}{\gamma}\right)\right) \\
&= 2^{-\mathcal{M}}. && (36)
\end{aligned}$$

Since (36) holds true for any possible value of τ and any fixed $\mathcal{F}_{t'}$ where $\tau = t'$, it holds true in general. Moreover, note that when T is large enough, we can choose $\delta = \frac{1}{T^2}$ since $\varepsilon = \sqrt{\frac{6\ln T}{\frac{2\gamma}{K}T_1}} \in (0,1]$ for large enough T. As a result, we get

$$\pi_{T_1+1,1} \geq 2^{-M},$$

where M is defined in (15). □

B.3.2 Upper bound on the Number of Pulls of Arm 2

Proof of Lemma 25. We can lower bound $\pi_{t_m+1,1}$ as

$$\begin{aligned}
\pi_{t_m+1,1} = 1-\pi_{t_m+1,2} &= 1-\pi_{t_0,2}\prod_{s=t_0}^{t_m}\frac{\pi_{s+1,1}}{\pi_{s,1}} \\
&= 1-\pi_{t_0,2}\prod_{s=t_0}^{t_m}\left(1-\eta\frac{1-\pi_{s,2}}{(1-\gamma)\pi_{s,2}+\frac{\gamma}{K}}\mathbb{1}[I_s=2]\right) \\
&\geq 1-\pi_{t_0,2}\prod_{s=t_0}^{t_m}\left(1-\eta\frac{\pi_{s,1}}{1}\mathbb{1}[I_s=2]\right) && \left((1-\gamma)\pi_{s,2}+\frac{\gamma}{K}\leq 1\right) \\
&\geq 1-\pi_{t_0,2}\prod_{s=t_0}^{t_m}\left(1-\eta\pi_{t_0,1}\mathbb{1}[I_s=2]\right) && \left(\pi_{s,1}\geq\pi_{t_0,1}\right) \\
&= 1-\pi_{t_0,2}\prod_{s=t_0}^{t_m}\left(1-\eta\pi_{t_0,1}\right)^{\mathbb{1}[I_s=2]} \\
&= 1-\pi_{t_0,2}\left(1-\eta\pi_{t_0,1}\right)^{m} \\
&\geq 1-\left(1-\eta\pi_{t_0,1}\right)^{m}. && (\pi_{t_0,2}\leq 1)
\end{aligned}$$

Moreover,

$$\begin{aligned}
(1-\eta\pi_{t_0,1})^m &\leq \exp\left(-\eta\pi_{t_0,1}m\right) && (1-x\leq \exp(x), \forall x\in\mathbb{R})\\
&\leq \exp\left(\frac{-2\pi_{t_0,1}}{1-2\pi_{t_0,1}}\right) && (m\geq \frac{2}{\eta\left(1-2\pi_{t_0,1}\right)})\\
&\leq 1-2\pi_{t_0,1}, && (e^{\frac{-2x}{1-2x}}\leq 1-2x, \forall x\in(0,\frac{1}{2}])
\end{aligned}$$

which means

$$\pi_{t_m+1,1}\geq 1-\left(1-\eta\pi_{t_0,1}\right)^m\geq 2\pi_{t_0,1}.$$

□

Proof of Lemma 26. Note that if $m\leq 2$, then we already have $\pi_{t,1}\geq\frac{1}{4}$, and since $t\geq T_1$, any pulls of arm 2 only increase $\pi_{t,1}$. Hence, after k pulls we have $\pi_{t',1}\geq\pi_{t,1}\geq\frac{1}{4}$.

Consider the case where $m\geq 3$. We have $2^{-m}\leq\pi_{t,1}\leq 2^{m-1}$. Now we want to upper bound the number of pulls it takes so that $1/4\leq\pi_{t',1}\leq 1/2$. Suppose we require k_1 pulls for the first doubling of π, k_2 for the second doubling, and so forth. This means we need $k=\sum_{i=1}^{m-2}k_i$ pulls before we get $1/4\leq\pi_{t',1}\leq 1/2$. Next, we upper bound each k_i. To do this, we denote all the rounds after t, in which we pull arm 2 as follows

$$\underbrace{t_1^{(1)},t_2^{(1)},\ldots,t_{k_1}^{(1)}}_{\text{rounds before 1st doubling}}\quad\underbrace{t_1^{(2)},\ldots,t_{k_2}^{(2)}}_{\text{rounds before 2nd doubling}}\quad\cdots\quad\underbrace{t_1^{(i)},t_2^{(i)},\ldots t_{k_i}^{(i)}}_{\text{rounds before i}^{th}\text{ doubling}}\quad\cdots\quad\underbrace{t_1^{(m_2-2)},\ldots,t_{k_{m_2-2}}^{(m_2-2)}}_{\text{rounds before }(m-2)\text{ doubling}}\quad,\quad\underbrace{t_k}_{1/4\leq\pi_{t_k,1}}$$

Now we can upper bound each k_i as follows:

$$\begin{aligned}
k_i &\leq \frac{2}{\eta}\frac{1}{1-2\pi_{t_1}^{(i)}} && (\text{Lemma 25})\\
&\leq \frac{2}{\eta}\frac{1}{1-2^{-m+i+1}}\ . && (\pi_{t_1}^{(i)}\leq 2^{-m+i})
\end{aligned}$$

To see why the first inequality holds, observe that we start from round ${t_1}^{(i)}$, and Lemma 25 has an upper bound on the number of pulls needed to get doubled.

Therefore, we get

$$\sum_{i=1}^{m-2}k_i\leq\sum_{i=1}^{m-2}\frac{2}{\eta}\frac{1}{1-2^{-m+i+1}}=\frac{2}{\eta}\sum_{i=1}^{m-2}\frac{1}{1-2^{-i}}\ . \tag{37}$$

Now, observe that

$$\begin{aligned}
\frac{1}{1-2^{-i}} &= 1+\frac{1}{2^i-1}\\
&\leq 1+\frac{1}{2^{i-1}}\ . && (2^i-1\geq 2^{i-1}\text{ for }i\geq 1)
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\sum_{i=1}^{m-2}k_i &\leq \frac{2}{\eta}\sum_{i=1}^{m-2}\left(1+\frac{1}{2^{i-1}}\right)\\
&\leq \frac{2}{\eta}\left[m-2+\sum_{i=0}^{m-2}\frac{1}{2^i}\right]\\
&\leq \frac{2}{\eta}\left[m-2+2\right] && (\text{geometric series})\\
&= \frac{2}{\eta}m.
\end{aligned}$$

□

B.3.3 High-probability lemma for event $\mathcal{E}_2$

Proof of Lemma 27. Recall the definition of $\mathcal{E}_1$ from Definition 22 and $\mathcal{E}_2$ from Definition 23 as follows:

$$\mathcal{E}_1 = \left\{\pi_{T_1+1,1} \geq 2^{-M}\right\}$$
$$\mathcal{E}_2 = \left\{\pi_{T_1+T_2+1,1} \geq \frac{1}{4}\right\},$$

where $M = \frac{1}{\ln 2}\left(2(1+\varepsilon_1)(1+\frac{\eta K}{\gamma})\eta T_1 + \ln \frac{2K}{\gamma}\right)$. We then show the following two statements.

(a) For $\delta = \frac{1}{T^2}$, we prove that with probability at least $1-\delta_1$, $\mathcal{E}_1$ happens, i.e., $\Pr(\mathcal{E}_1) \geq 1-\delta_1$

(b) Given $\mathcal{E}_1$ happened, for $\delta_2 = \frac{1}{T^2}$, we prove that with probability at least $1-\delta_2$, $\mathcal{E}_2$ happens, i.e., $\Pr(\mathcal{E}_2 \mid \mathcal{E}_1) \geq 1-\delta_2$.

Having both (a) and (b) implies that the lemma holds true, as

$$\Pr(\mathcal{E}_2) \geq \Pr(\mathcal{E}_1 \text{ and } \mathcal{E}_2) = \Pr(\mathcal{E}_1)\,\Pr(\mathcal{E}_2 \mid \mathcal{E}_1) = (1-\delta_1)\,(1-\delta_2)$$
$$\geq 1-\delta_1-\delta_2 = 1-\frac{2}{T^2}.$$

Now, we prove (a) and (b).

Proof of (a) By Lemma 24, with probability at least $1-\frac{1}{T^2}$, we have

$$\pi_{T_1+1,1} \geq 2^{-M},$$

for $M = \frac{1}{\ln 2}\left(2(1+\varepsilon_1)(1+\frac{\eta K}{\gamma})\eta T_1 + \ln \frac{2K}{\gamma}\right)$.

Proof of (b) We show that for any history $\mathcal{F}_{T_1}$ such that $\mathcal{E}_1$ happened, we have

$$\Pr(\mathcal{E}_2 \mid F_{T_1}) \geq 1-\delta_2. \tag{38}$$

This implies that $\Pr(\mathcal{E}_2 \mid \mathcal{E}_1) \geq 1-\delta_2$.

Consider a fixed history $\mathcal{F}_{T_1}$ such that $\mathcal{E}_1$ happened. Event $\mathcal{E}_1$ implies that for some $M' \leq M$, we have

$$2^{-M'} \leq \pi_{T_1+1,1} \leq 2^{-(M'-1)}.$$

Now, Lemma 26 states that $\Gamma = \frac{2}{\eta}M'$ pulls is sufficient to get

$$\pi_{t',1} \geq \frac{1}{4}, \tag{39}$$

where t' is round number after Γ-th pull. We define

$$X_t := \mathbb{1}\left[I_t = 2 \text{ or } \pi_{t,1} \geq \frac{1}{4}\right].$$

Next, observe that if

$$\sum_{t\in\text{phase 2.1}} X_t = \sum_{t=T_1+1}^{T_1+T_2} X_t \geq \frac{2}{\eta}M', \tag{40}$$

then this implies that $\pi_{T_1+T_2+1} \geq 1/4$, (i.e. $\mathcal{E}_2$ happens.) To see why, note that if for any for t in $T_1+1 \leq t \leq T_1+T_2$, we have $\pi_{t,1} \geq 1/4$, this implies $\pi_{T_1+T_2+1} \geq 1/4$ since $\pi_{t,1}$ can only increase in phase 2.1. If for all t in $T_1+1 \leq t \leq T_1+T_2$, we have $\pi_{t,1} < 1/4$, then (40) implies that

$$\sum_{t=T_1+1}^{T_1+T_2} X_t = \sum_{t=T_1+1}^{T_1+T_2} \mathbb{1}[I_t = 2] \geq \frac{2}{\gamma}M'.$$

Therefore, Lemma 26 implies that $\pi_{T_1+T_2+1} \geq 1/4$.

Now it remains to show that with probability at least $1-\frac{1}{T^2}$, (40) happens. We use a martingale concentration argument to show this. Indeed, we define $Z_{T_1} = 0$ and for any t in $T_1 + 1 \leq t \leq T_1 + T_2$, we define $W_t := X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ and $Z_t := \sum_{s=T_1+1}^{t} W_s$. Observe that

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = Z_{t-1} + \mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = Z_{t-1},$$

and hence $(Z_t)_{t\in\{T_1,\dots,T\}}$ is a martingale. Since $X_t \in \{0,1\}$, we have $Z_t - Z_{t-1} = W_t \in [-A_t, B_t]$ for $A_t = \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right]$ and $B_t = 1 - A_t$. Consequently, we have $A_t + B_t = 1 := c$ for all $t \geq t'$. Define $q := \frac{3-\gamma}{4}$. Observe that we have $\mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] \geq q$. It is because for any $\mathcal{F}_{t-1}$ such that $\pi_{t,1} \geq 1/4$, we have $\mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{1}\left[I_t = 2 \text{ or } \pi_{t,1} \geq \frac{1}{4}\right] \mid \mathcal{F}_{t-1}\right] = 1 > q$. Moreover, for any $\mathcal{F}_{t-1}$ such that $\pi_{t,1} < 1/4$, we have $\pi_{t,2} \geq 3/4$ and hence $\mathbb{E}\left[\mathbb{1}\left[I_t = 2\right] | \mathcal{F}_{t-1}\right] = \tilde{\pi}_{t,2} = (1-\gamma)\pi_{t,2} + \frac{\gamma}{2} \geq \frac{3-\gamma}{4}$, therefore

$$\mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{1}\left[I_t = 2\right] \text{ or } \mathbb{1}[\pi_{t,1} \geq \frac{1}{4}] \,\middle|\, \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{1}\left[I_t = 2\right] | \mathcal{F}_{t-1}\right] \geq \frac{3-\gamma}{4}.$$

Note that we have

$$\sum_{t=T_1+1}^{T_1+T_2} A_t = \sum_{t=T_1+1}^{T_1+T_2} \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] \leq \sum_{t=T_1+1}^{T_1+T_2} q = qT_2 =: \mu.$$

We now apply Theorem 15 from Kuszmaul and Qi (2021) to get for any $\varepsilon > 0$,

$$\Pr(Z_{T_1+T_2} - Z_{T_1} \leq -\varepsilon\mu \mid \mathcal{F}_{T_1}) \leq \exp\left(-\frac{\varepsilon^2\mu}{2c}\right)$$

for $\mathcal{F}_{T_1}$ where $\mathcal{E}_1$ holds.

Plugging in our setting of c and using our definition of the martingale sequence gives

$$\Pr\left(\sum_{t=T_1+1}^{T_1+T_2} X_t \leq \sum_{t=T_1+1}^{T_1+T_2} \mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] - \varepsilon q T_2 \,\middle|\, \mathcal{F}_{T_1}\right) \leq \exp\left(-\frac{\mu\varepsilon^2}{2}\right).$$

Using $\sum_{t=T_1+1}^{T_1+T_2} \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] \leq \mu$, we have all $\varepsilon > 0$,

$$\Pr\left(\sum_{t=T_1+1}^{T_1+T_2} X_t \leq (1-\varepsilon)\mu \,\middle|\, \mathcal{F}_{T_1}\right) \leq \exp\left(-\frac{\mu\varepsilon^2}{2}\right).$$

This implies that, for any given $\mathcal{F}_{T_1}$ such that $\mathcal{E}_1$ holds, with probability at least $1 - \frac{1}{T^2}$, we have

$$\sum_{t=T_1+1}^{T_1+T_2} X_t \geq \mu = \frac{3-\gamma}{4}(1-\varepsilon_2)T_2,$$

where $\varepsilon_2 = \sqrt{\frac{4\ln T}{\frac{3-\gamma}{4}T_2}}$. Now, recall T' from Definition 14. For large enough T, by (17), we have $T_2 \geq T'$. Therefore,

$$(\frac{3-\gamma}{4})(1-\varepsilon_2)T_2 \geq (\frac{3-\gamma}{4})(1-\varepsilon_2)T'.$$

Also by definition of T', we get

$$\frac{3-\gamma}{4}(1-\varepsilon_2)T' = M\frac{2}{\eta}.$$

Finally by definition of M', we have

$$M\frac{2}{\eta} \geq M'\frac{2}{\eta},$$

which means with probability at least $1 - \frac{1}{T^2}$, (40) happens.

□

B.3.4 Proof of Lemmas 28 and 29

We first prove Lemma 28.

Proof of Lemma 28. Consider any round t where $t \geq t_0+1$. Consider any history $\mathcal{F}_{t-1}$ where Event $\mathcal{E}_2$ happened. By applying Lemma 12 for $i=2$, we get

$$\mathbb{E}\left[\pi_{t+1,2} \mid \mathcal{F}_{t-1}\right] = (1-\eta)\,\pi_{t,2} + \eta\,\pi_{t,2}^2.$$

Now, note that since $\mathcal{E}_2$ happened we have $\pi_{t_0,1} \geq 1/4$. Since $\pi_{t,1} = \pi_{t_0+\tau}$ can only increase, we have $\pi_{t,1} \geq 1/4$. This implies $\pi_{t,2} \leq 3/4$, therefore

$$\pi_{t,2}^2 \leq \frac{3}{4}\,\pi_{t,2}.$$

Therefore, we get

$$\begin{aligned}\mathbb{E}\left[\pi_{t+1,2} \mid \mathcal{F}_{t-1}\right] &\leq (1-\eta)\,\pi_{t,2} + \frac{3}{4}\eta\,\pi_{t,2}\\ &\leq (1-\frac{\eta}{4})\,\pi_{t,2}.\end{aligned} \tag{41}$$

We now can show an upper bound on $\mathbb{E}\left[\pi_{t+1,2} \,|\, \mathcal{E}_2\right]$ by noting that

$$\begin{aligned}\mathbb{E}\left[\pi_{t+1,2} \,|\, \mathcal{E}_2\right] &= \mathbb{E}\left[\mathbb{E}\left[\pi_{t+1,2} \,|\, \mathcal{E}_2, \mathcal{F}_{t-1}\right] \,|\, \mathcal{E}_2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\pi_{t+1,2} \,|\, \mathcal{F}_{t-1}\right] \,|\, \mathcal{E}_2\right].\end{aligned}$$

This means that we can take the conditional expectation on both sides of (41) to get

$$\mathbb{E}[\pi_{t+1,2}|\mathcal{E}_2] \leq (1-\frac{\eta}{4})\mathbb{E}\left[\pi_{t,2}|\mathcal{E}_2\right]. \tag{42}$$

Moreover, by definition of $\mathcal{E}_2$ we have $\mathbb{E}[\pi_{t_0,2}|\mathcal{E}_2] = \pi_{t_0,2} \leq 3/4$. Therefore, we get

$$\begin{aligned}\mathbb{E}[\pi_{t,2} \mid \mathcal{E}_2] = \mathbb{E}[\pi_{t_0,2} \mid \mathcal{E}_2] &\prod_{s=t_0}^{t_0+\tau-1} \frac{\mathbb{E}[\pi_{s+1,2} \mid \mathcal{E}_2]}{\mathbb{E}[\pi_{s,2} \mid \mathcal{E}_2]}\\ \leq \mathbb{E}[\pi_{t_0,2} \mid \mathcal{E}_2] &\prod_{s=t_0}^{t_0+\tau-1} (1-\frac{\eta}{4}) && \text{by (42)}\\ \leq \frac{3}{4}\,(1-\frac{\eta}{4})^{\tau}&\\ \leq \frac{3}{4}\exp\,(-\frac{\eta}{4}\tau).& && (1-x \leq e^{-x})\end{aligned}$$

□

We now prove Lemma 29.

Proof of Lemma 29. Let $\tau = T_3 + T_4 = cT$ for $c > 0$. Define random variable $X := \pi_{T+1,1} \in [0,1]$. Clearly Lemma 28 implies that $\mathbb{E}[X|\mathcal{E}_2] = \mathbb{E}[\pi_{T+1,1}|\mathcal{E}_2] = 1 - \mathbb{E}[\pi_{T+1,2}|\mathcal{E}_2] \geq 1 - \frac{3}{4}e^{-\frac{\eta}{4}cT}$. Therefore, using Chebyshev's inequality, we get

$$\begin{aligned}
\Pr\left(X \le \frac{3}{4} \,\middle|\, \mathcal{E}_2\right) &= \Pr\left(X - \mathbb{E}\left[X|\mathcal{E}_2\right] + \mathbb{E}\left[X|\mathcal{E}_2\right] \le \frac{3}{4} \,\middle|\, \mathcal{E}_2\right) \\
&= \Pr\left(X - \mathbb{E}\left[X|\mathcal{E}_2\right] \le \frac{3}{4} - \mathbb{E}\left[X|\mathcal{E}_2\right] \,\middle|\, \mathcal{E}_2\right) \\
&\le \Pr\left(X - \mathbb{E}\left[X|\mathcal{E}_2\right] \le \frac{3}{4} - \left(1 - \frac{3}{4}e^{-\frac{\eta}{4}cT}\right) \,\middle|\, \mathcal{E}_2\right) && \left(\mathbb{E}\left[X|\mathcal{E}_2\right] \ge 1 - \frac{3}{4}e^{-\frac{\eta}{4}cT}\right) \\
&= \Pr\left(X - \mathbb{E}\left[X|\mathcal{E}_2\right] \le \frac{3}{4}e^{-\frac{\eta}{4}cT} - \frac{1}{4} \,\middle|\, \mathcal{E}_2\right) \\
&\le \Pr\left(X - \mathbb{E}\left[X|\mathcal{E}_2\right] \le -\frac{1}{5} \,\middle|\, \mathcal{E}_2\right) && \left(\text{for large } T \text{ we have} \frac{3}{4}e^{-\frac{c}{4}\eta T} \le \frac{1}{20}\right) \\
&\le \Pr\left(\left|X - \mathbb{E}\left[X|\mathcal{E}_2\right]\right| \ge \frac{1}{5} \,\middle|\, \mathcal{E}_2\right) \\
&\le 25 \operatorname{Var}(X|\mathcal{E}_2) && (\text{Chebyshev inequality}) \\
&= 25\left(\mathbb{E}\left[X^2|\mathcal{E}_2\right] - \mathbb{E}\left[X|\mathcal{E}_2\right]^2\right) \\
&\le 25\left(\mathbb{E}\left[X|\mathcal{E}_2\right] - \mathbb{E}\left[X|\mathcal{E}_2\right]^2\right) && (E[X] \ge E[X^2] \text{ for X } \in [0,1]) \\
&= 25\,\mathbb{E}\left[X|\mathcal{E}_2\right]\left(1 - \mathbb{E}\left[X \mid \mathcal{E}_2\right]\right) \\
&\le 25\left(1 - \mathbb{E}\left[X|\mathcal{E}_2\right]\right) && (\mathbb{E}\left[X|\mathcal{E}_2\right] \le 1) \\
&\le \frac{75}{4}e^{-c\frac{\eta}{4}T}. && (\text{from Lemma 28})
\end{aligned}$$

□

B.3.5 Proof of Claim 1

Now, we are ready to prove Claim 1. We recall Claim 1.

Claim 1 (Concentration on best arm at the end)**.** *For large enough T, there exists $c_1 > 0$ such that*

$$\mathbb{E}\left[\ln \pi_{T+1,1} + \ln K\right] \ge c_1. \tag{11}$$

First, we have the following simple observation.

Observation 31. When running WSU-UX with any valid hyperparameter η, γ on the loss sequence $\{\ell_t\}_{t=1}^T$ defined in Definition 9, we have with probability 1, that

$$\pi_{T+1,1} \ge \left(\frac{1}{2}\right)^{\frac{T}{100}+1}.$$

Proof of Observation 31. The probability of $\pi_{t,1}$ can only decrease in the first $\frac{T}{100}$ rounds and only if arm 1 is pulled in those rounds. It is easy to see that the value drops by at most a factor of 2 each time it is pulled as for $1 \le s \le T_1$ we have

$$\frac{\pi_{s+1,1}}{\pi_{s,1}} = 1 - \eta\Big(\hat{\ell}_{s,1} - \sum_{j\in[2]} \pi_{s,j}\hat{\ell}_{s,j}\Big) \ge 1 - \frac{\eta K}{\gamma}\mathbb{1}[I_s = 1] \ge 1/2.$$

□

Proof of Claim 1. Define Event $A := \mathbb{1}\left[\pi_{T+1,1} \ge 3/4\right]$. Using conditional expectation, we have

$$\begin{aligned}
\mathbb{E}[\ln \pi_{T+1,1} \mid \mathcal{E}_2] &= \Pr(A \mid \mathcal{E}_2)\,\mathbb{E}[\ln \pi_{T+1,1} \mid \mathcal{E}_2, A] + \Pr(A^c \mid \mathcal{E}_2)\,\mathbb{E}[\ln \pi_{T+1,1} \mid \mathcal{E}_2, A^c] \\
&\ge \left(1 - \frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\mathbb{E}[\ln \pi_{T+1,1}|\mathcal{E}_2, A] + \left(\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\mathbb{E}[\ln \pi_{T+1,1}|\mathcal{E}_2, A^c],
\end{aligned}$$

where the inequality comes from Lemma 29. This can be further lower bounded by

$$\begin{aligned}
&\left(1-\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\ln\frac{3}{4}+\left(\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\mathbb{E}[\ln\pi_{T+1,1}|\mathcal{E}_2,A^c]\\
&\geq\left(1-\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\ln\frac{3}{4}+\left(\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\min\left[\ln\pi_{T+1,1}\right]\\
&\geq\left(1-\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\ln\frac{3}{4}+\underbrace{\left(\frac{75}{4}e^{-\frac{c\eta}{4}T}\right)\left(\frac{T}{100}+1\right)\ln\frac{1}{2}}_{\text{second term}} && \text{(Observation 31)}\\
&\geq\ln\frac{11}{16}. && \text{second term } \to 0 \text{ as T} \to\infty
\end{aligned}$$

Therefore, we have

$$\mathbb{E}[\ln\pi_{T+1,1}\mid\mathcal{E}_2]\geq\ln\frac{11}{16}. \tag{43}$$

Now, we can lower bound $\mathbb{E}[\ln\pi_{T+1,1}]$ using conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\ln\pi_{T+1,1}\right]&=\Pr(\mathcal{E}_2)\,\mathbb{E}\left[\ln\pi_{T+1,1}\,|\,\mathcal{E}_2\right]+\Pr(\mathcal{E}_2^c)\,\mathbb{E}\left[\ln\pi_{T+1,1}\,|\,\mathcal{E}_2^c\right]\\
&\geq(1-2\delta)\;\mathbb{E}\left[\ln\pi_{T+1,1}\,|\,\mathcal{E}_2\right]+(2\delta)\,\mathbb{E}\left[\ln\pi_{T+1,1}\,|\,\mathcal{E}_2^c\right]\\
&\geq(1-2\delta)\,\mathbb{E}\left[\ln\pi_{T+1,1}\,|\,\mathcal{E}_2\right]+(2\delta)\;\left[\min\ln\pi_{T+1,1}\right]\\
&\geq(1-2\delta)\,(\ln\frac{11}{16})+(2\delta)\,\left[\min\ln\pi_{T+1,1}\right] && \text{(By inequality 43)}\\
&\geq(1-2\delta)\,(\ln\frac{11}{16})+(2\delta)\,(\frac{T}{100}+1)\ln\frac{1}{2} && \text{(Observation 31)}\\
&\geq(\ln\frac{11}{16})+\underbrace{(2\delta)\,(\frac{T}{100}+1)\ln\frac{1}{2}}_{\text{second term}}\\
&\geq(\ln\frac{5}{8}). && \delta=\frac{1}{T^2}\text{, therefore second term}\to 0\text{ as T}\to\infty
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
\mathbb{E}[\ln\pi_{T+1,1}+\ln K]&\geq(\ln\frac{5}{8})+\ln K\\
&=\ln\frac{5}{4}. && (K=2)
\end{aligned}$$

Therefore, for $c_1=\ln\frac{5}{4}>0$,

$$\mathbb{E}[\ln\pi_{T+1,1}+\ln K]\geq c_1.$$

□

B.3.6 Proof of Claim 3

We first prove Lemma 30 and then we prove Claim 3.

Proof of Lemma 30. Set $t=\frac{31T}{100}=T_1+T_2+T_3+1$. We can use conditional expectation for $\pi_{2,t}$ on event $\mathcal{E}_2$ defined in Definition 23 to get

$$\begin{aligned}
\mathbb{E}[\pi_{t,2}]&=\Pr(\mathcal{E}_2)\mathbb{E}[\pi_{t,2}|\mathcal{E}_2]+\Pr(\mathcal{E}_2^c)\mathbb{E}[\pi_{t,2}|\mathcal{E}_2^c]\\
&\leq\mathbb{E}[\pi_{t,2}|\mathcal{E}_2]+\Pr(\mathcal{E}_2^c)\cdot 1. && (\pi_{t,2}\leq 1)
\end{aligned}$$

Now, by setting $\tau = T_3$ in Lemma 28, one would get $\mathbb{E}\left[\pi_{t,2}|\mathcal{E}_2\right] \leq \frac{3}{4}\exp\left(-\frac{\eta}{4}T_3\right) = \frac{3}{4}\exp\left(\frac{-c\eta}{4}T\right)$ for $c = \frac{1}{10} > 0$. Moreover, by Lemma 27 we have that $\Pr(\mathcal{E}_2^c) \leq \frac{2}{T^2}$. Therefore, for large enough T, we can further upper bound $\mathbb{E}[\pi_{t,2}]$ by

$$\mathbb{E}\left[\pi_{t,2} \mid \mathcal{E}_2\right] + \Pr(\mathcal{E}_2^c) \cdot 1 \leq \frac{3}{4}e^{\frac{-c\eta T}{4}} + \frac{2}{T^2} \leq \frac{1}{4}.$$

□

Proof of Claim 3. By expanding $\tilde{\pi}_{t,j} = \pi_{t,j}(1-\gamma) + \frac{\gamma}{2}$, we get

$$\left[\sum_{j=1}^{2} \tilde{\pi}_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right] = \sum_{j=1}^{2}\left(\pi_{t,j}(1-\gamma) + \frac{\gamma}{2}\right)\hat{\ell}_{t,j} - \hat{\ell}_{t,1} = \left[\sum_{j=1}^{2} \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right] - \sum_{j=1}^{2}\gamma\pi_{t,j}\hat{\ell}_{t,j} + \sum_{j=1}^{2}\frac{\gamma}{2}\hat{\ell}_{t,j}$$

Taking the expectation from both sides, we get

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=1}^{2} \tilde{\pi}_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{2} \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] + \left(-\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\hat{\ell}_{t,j}\right] + \sum_{j=1}^{2}\mathbb{E}\left[\frac{\gamma}{2}\hat{\ell}_{t,j}\right]\right) \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{2} \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] - \sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\,\mathbb{E}_{t-1}\left[\hat{\ell}_{t,j}\right]\right] + \sum_{j=1}^{2}\mathbb{E}\left[\frac{\gamma}{2}\,\mathbb{E}_{t-1}\left[\hat{\ell}_{t,j}\right]\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{2} \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] + \left(-\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\ell_{t,j}\right] + \sum_{j=1}^{2}\frac{\gamma}{2}\ell_{t,j}\right).
\end{aligned}$$

Summing over T rounds, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{2} \tilde{\pi}_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{j=1}^{2} \pi_{t,j}\hat{\ell}_{t,j} - \hat{\ell}_{t,1}\right)\right] + \underbrace{\left(-\sum_{t=1}^{T}\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\ell_{t,j}\right] + \sum_{t=1}^{T}\sum_{j=1}^{2}\frac{\gamma}{2}\ell_{t,j}\right)}_{\Delta},$$

where we define

$$\Delta := -\sum_{t=1}^{T}\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\ell_{t,j}\right] + \sum_{t=1}^{T}\sum_{j=1}^{2}\frac{\gamma}{2}\ell_{t,j}. \tag{44}$$

Note that to prove Claim 3, we need to show that for large enough T,

$$\Delta \geq c_3\gamma T \tag{45}$$

holds true.

Note that in loss sequence $\{\ell_t\}_{t=1}^{T}$ for all rounds t, we have $\ell_{t,1} + \ell_{t,2} = 1$; therefore

$$\sum_{t=1}^{T}\sum_{j=1}^{2}\frac{\gamma}{2}\ell_{t,j} = \frac{\gamma T}{2}. \tag{46}$$

Moreover, for large enough T we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\ell_{t,j}\right] &= \underbrace{\sum_{t\in\mathcal{T}_1}\mathbb{E}\left[\gamma\pi_{t,1}\right]+}_{\text{phase 1}}\underbrace{\sum_{t\in\mathcal{T}_2\cup\mathcal{T}_3}\mathbb{E}\left[\gamma\pi_{t,2}\right]}_{\text{phase 2.1 and phase 2.2}}+\underbrace{\sum_{t\in\mathcal{T}_4}\mathbb{E}\left[\gamma\pi_{t,2}\right]}_{\text{phase 2.3}} \\
&\le T_1\frac{\gamma}{2}+\sum_{t\in\mathcal{T}_2\cup\mathcal{T}_3}\mathbb{E}\left[\gamma\pi_{t,2}\right]+\sum_{t\in\mathcal{T}_4}\mathbb{E}\left[\gamma\pi_{t,2}\right] && (\pi_{t,1}\le 1/2\in\mathcal{T}_1) \\
&\le T_1\frac{\gamma}{2}+\gamma\left(T_2+T_3\right)+\sum_{t\in\mathcal{T}_4}\mathbb{E}\left[\gamma\pi_{t,2}\right] && (\pi_{t,2}\le 1) \\
&\le T_1\frac{\gamma}{2}+\gamma\left(T_2+T_3\right)+T_4\,\frac{\gamma}{4}, && (\forall t\in\mathcal{T}_4,\mathbb{E}[\pi_{t,2}]\le 1/4 \text{ when } T \text{ is large})
\end{aligned} \tag{47}$$

where the first inequality comes from the fact that $\pi_{t,1}\le 1/2$ when $1\le t\le T_1$. The third inequality comes from the fact that by Lemma 30, for large enough T, we have $\mathbb{E}[\pi_{t,2}]\le\frac{1}{4}$ when $t\in\mathcal{T}_4$.

By using (46) and (47), we have

$$\begin{aligned}
\Delta = -\sum_{t=1}^{T}\sum_{j=1}^{2}\mathbb{E}\left[\gamma\pi_{t,j}\ell_{t,j}\right]+\sum_{t=1}^{T}\sum_{j=1}^{2}\frac{\gamma}{2}\ell_{t,j} &\ge -\left(T_1\frac{\gamma}{2}+\left(T_2+T_3\right)\gamma+T_4\,\frac{\gamma}{4}\right)+T\frac{\gamma}{2} \\
&= -T\,(\frac{1}{100}\frac{\gamma}{2}+\frac{3}{10}\gamma+\frac{69}{100}\frac{\gamma}{4})+T\frac{\gamma}{2} \\
&= \frac{9}{400}\gamma T = c_3\gamma T,
\end{aligned}$$

for large enough T, i.e., (45) holds. This proves the claim. □

C Potential Analysis

C.1 WSU as a Linear approximation of Hedge Update

As mentioned in the main text, the WSU update can be viewed as a linear approximation to the Hedge update. In this section, we briefly show this approximation argument.

Observe that by applying the linear approximation $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ for $f(x) = \exp(-x)$ and for $x = \eta\ell_{t,i}$ and $x_0 = \eta\bar{\ell}_t$, where $\bar{\ell}_t := \sum_j \pi_{t,j}\ell_{t,j}$, we get

$$\exp\left(-\eta\ell_{t,i}\right) \approx \exp\left(-\eta\bar{\ell}_t\right) \cdot \left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right). \tag{48}$$

Note that Hedge updates weights by the LHS of (48). Now, if we instead update the weights by RHS of (48), we get

$$w_{t+1,i} = w_{t,i} \cdot \exp\left(-\eta\bar{\ell}_t\right) \cdot \left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right).$$

By defining $\pi_{t,i} := \frac{w_{t,i}}{\sum_{j\in[K]} w_{t,j}}$, we get

$$\begin{aligned}
\pi_{t+1,i} = \frac{w_{t+1,i}}{\sum_j w_{t+1,j}} &= \frac{\exp\left(-\eta\bar{\ell}_t\right)\left[w_{t,i} \cdot \left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right)\right]}{\exp\left(-\eta\bar{\ell}_t\right)\left[\sum_{j\in[K]} w_{t,j} \cdot \left(1 - \eta\left(\ell_{t,j} - \bar{\ell}_t\right)\right)\right]} \\
&= \frac{w_{t,i} \cdot \left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right)}{\sum_{j\in[K]} w_{t,j} \cdot \left(1 - \eta\left(\ell_{t,j} - \bar{\ell}_t\right)\right)} \\
&= \frac{w_{t,i} \cdot \left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right)}{\sum_{j\in[K]} w_{t,j}} \\
&= \pi_{t,i}\left(1 - \eta\left(\ell_{t,i} - \bar{\ell}_t\right)\right).
\end{aligned}$$

Note that this recovers the WSU update.[10]

C.2 Completed version of Potential Argument of Subsection 3.1

In this subsection, for the convenience of the reader, we give a comprehensive explanation of the derivation of (7) and (5).

In the potential analysis of Hedge which can be found in Hazan et al. (2016), for any $i \in [K]$ and $t \in [T]$, we define $\Phi^{\text{HEDGE}}_{t,i} := w_{t,i}$ with $w_{t,i}$ and $\pi_{t,i}$ as in Definition 5. Moreover, assume that $w_{1,i} = 1$.[11] We also define define $\Phi^{\text{HEDGE}}_{t} := \sum_{j\in[K]} w_{t,j}$. By non-negativity of $w_{t,i}$, we have

$$\frac{1}{\eta}\ln\left(\Phi^{\text{HEDGE}}_{T+1,i}\right) \leq \frac{1}{\eta}\ln\left(\Phi^{\text{HEDGE}}_{T+1}\right), \tag{49}$$

It is easy to see that for any $t \in [T]$ we can write

$$\Phi^{\text{HEDGE}}_{t+1} = \Phi^{\text{HEDGE}}_{t}\left(\sum_{j\in[K]} \pi_{t,j}\exp\left(-\eta\ell_{t,j}\right)\right).$$

Note that we have

$$\begin{aligned}
\sum_{j\in[K]} \pi_{t,j}\exp\left(-\eta\ell_{t,j}\right) &\leq 1 - \eta\sum_{j\in[K]} \pi_{t,j}\ell_{t,j} + \eta^2\sum_{j\in[K]} \pi_{t,j}\left(\ell_{t,j}\right)^2 && (\exp\left(-x\right) \leq 1 - x + x^2 \text{ for } x \geq 0) \\
&\leq \exp\left(-\eta\sum_{j\in[K]} \pi_{t,j}\ell_{t,j} + \eta^2\sum_{j\in[K]} \pi_{t,j}\left(\ell_{t,j}\right)^2\right). && (\exp\left(x\right) \leq 1 + x)
\end{aligned}$$

[10] The idea of linear approximation of hedge was noted by Kivinen and Warmuth (1997) for a slightly different setting.

[11] This is slightly different than Definition 5 where $w_{1,i} = 1/K$. We can view it as dividing all weights by the same constant. This does not impact the behaviour of Hedge at all.

Therefore, we have

$$\Phi_{t+1}^{\text{HEDGE}} \le \Phi_t^{\text{HEDGE}} \exp\left(-\eta \sum_{j\in[K]} \pi_{t,j}\ell_{t,j} + \eta^2 \sum_{t\in[T]}\sum_{j\in[K]} \pi_{t,j}\left(\ell_{t,j}\right)^2\right).$$

By applying (50) recursively, we get

$$\begin{aligned}\Phi_{T+1}^{\text{HEDGE}} &\le \Phi_1^{\text{HEDGE}} \exp\left(-\eta \sum_{t\in[T]}\sum_{j\in[K]} \pi_{t,j}\ell_{t,j} + \eta^2 \sum_{t\in[T]}\sum_{j\in[K]} \pi_{t,j}\left(\ell_{t,j}\right)^2\right)\\ &= \exp\left(\ln K - \eta \sum_{t\in[T]}\sum_{j\in[K]} \pi_{t,j}\ell_{t,j} + \eta^2 \sum_{t\in[T]}\sum_{j\in[K]} \pi_{t,j}\left(\ell_{t,j}\right)^2\right), \end{aligned} \tag{50}$$

since $\Phi_1^{\text{HEDGE}} = \sum_{j\in[K]} \frac{1}{K} = K$.

On the other hand, we have

$$\Phi_{T+1,i}^{\text{HEDGE}} = \Phi_{T,i}^{\text{HEDGE}} \exp\left(-\eta\ell_{T,i}\right) = \Phi_{1,i}^{\text{HEDGE}} \exp\left(-\eta\sum_{t\in[T]}\ell_{t,i}\right) = \exp\left(-\eta\sum_{t\in[T]}\ell_{t,i}\right). \tag{51}$$

We can upper bound the RHS of (49) by (50) and lower bound the LHS of (49) by (51) to get

$$-\sum_{t\in[T]}\ell_{t,i} \le \frac{1}{\eta}\ln\left(\Phi_{T+1,i}^{\text{HEDGE}}\right) \le \frac{1}{\eta}\ln\left(\Phi_{T+1}^{\text{HEDGE}}\right) \le -\sum_{t\in[T]}\sum_{j\in[K]}\pi_{t,j}\ell_{t,j} + \frac{\ln K}{\eta} + \eta\sum_{t\in[T]}\left[\sum_{j\in[K]}\pi_{t,j}\left(\ell_{t,j}\right)^2\right].$$

Note that the above is the full version of (5). Rearranging, we get [12]

$$\sum_{t\in[T]}\sum_{j\in[K]}\pi_{t,j}\ell_{t,j} - \sum_{t\in[T]}\ell_{t,i} \le \underbrace{\frac{\ln K}{\eta}}_{\text{exploration term}} + \eta\sum_{t\in[T]}\underbrace{\left[\sum_{j}\pi_{t,j}\left(\ell_{t,j}\right)^2\right]}_{\text{Second order error}}.$$

For WSU, the potential is defined as $\Phi_{t,i}^{\text{WSU}} := \pi_{t,i}$ and $\Phi_t^{\text{WSU}} := \sum_{j\in[K]}\pi_{t,i} = 1$. By non-negativity of $\pi_{t,i}$ we have

$$\frac{1}{\eta}\ln\left(\Phi_{T+1,i}^{\text{WSU}}\right) \le \frac{1}{\eta}\ln\left(\Phi_{T+1}^{\text{WSU}}\right) = 0. \tag{52}$$

Now, the RHS of (6) (which is 0) does not involve any second-order error term. In fact, since WSU is normalized, the RHS does not give us information about the regret. However, we can extract the difference between the cumulative loss of the algorithm and expert i from the LHS of (6).

[12] Note that the exploration term is an inevitable error incurred by both Hedge and WSU when they move toward the optimal point in the simplex Δ_K by the learning rate η. We call it exploration term as the algorithm is trying to explore and find the optimal point in the domain of the simplex.

Indeed, we have

$$\begin{aligned}
\Phi^{\mathrm{WSU}}_{T+1,i} &= \Phi^{\mathrm{WSU}}_{T,i}\left(1-\eta\left[\ell_{T,i}-\sum_j \pi_{T,j}\ell_{T,j}\right]\right) \\
&= \Phi^{\mathrm{WSU}}_{1,i}\prod_{t\in[T]}\left(1-\eta\left[\ell_{t,i}-\sum_j \pi_{t,j}\ell_{t,j}\right]\right) \\
&\geq \frac{1}{K}\prod_{t\in[T]}\exp\left(-\eta\left[\ell_{t,i}-\sum_j \pi_{t,j}\ell_{t,j}\right]-\eta^2\left[\ell_{t,i}-\sum_j \pi_{t,j}\ell_{t,j}\right]^2\right) \\
&= \frac{1}{K}\exp\left(-\eta\sum_{t\in[T]}\left[\ell_{t,i}-\sum_j \pi_{t,j}\ell_{t,j}\right]-\eta^2\sum_{t\in[T]}\left[\ell_{t,i}-\sum_j \pi_{t,j}\ell_{t,j}\right]^2\right), \qquad (53)
\end{aligned}$$

where the inequality comes from $1-x \geq \exp(-x-x^2)$ for $0 \leq x \leq 1/2$.

Using (53), we can lower bound the LHS of (52) as

$$\sum_{t\in[T]}\left[\sum_j \pi_{t,j}\ell_{t,j}-\ell_{t,i}\right]-\frac{\ln K}{\eta}-\eta\sum_{t\in[T]}\left[\sum_j \pi_{t,j}\ell_{t,j}-\ell_{t,i}\right]^2 \leq \frac{1}{\eta}\ln\left(\Phi^{\mathrm{WSU}}_{T+1,i}\right) \leq \frac{1}{\eta}\ln\left(\Phi^{\mathrm{WSU}}_{T+1}\right)=0.$$

Note that the above is the full version of (7). Rearranging, we get

$$\sum_{t\in[T]}\sum_{j\in[K]}\pi_{t,j}\ell_{t,j}-\sum_{t\in[T]}\ell_{t,i} \leq \underbrace{\frac{\ln K}{\eta}}_{\text{exploration term}} + \eta\underbrace{\sum_{t\in[T]}\left[\sum_j \pi_{t,j}\ell_{t,j}-\ell_{t,i}\right]^2}_{\text{Second order error}}.$$

Implication for Bandit Case In the bandit setting, when we use WSU-UX, we can show that we get a second-order term in (7) which is upper bounded by

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_j \pi_{t,j}\hat{\ell}_{t,j}-\hat{\ell}_{t,i}\right)^2\right] &\leq \mathbb{E}\left[\left(\sum_j \pi_{t,j}\hat{\ell}_{t,j}\right)^2+\left(\hat{\ell}_{t,i}\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_j \pi_{t,j}\left(\hat{\ell}_{t,j}\right)^2+\left(\hat{\ell}_{t,i}\right)^2\right]. \qquad \text{(Jensen's inequality for } f(x)=x^2\text{)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in[K]}\pi_{t,j}\left(\hat{\ell}_{t,j}\right)^2\right] &= \mathbb{E}\left[\sum_{j\in[K]}\pi_{t,j}\left(\frac{\ell_{t,j}\mathbb{1}[I_t=j]}{\tilde{\pi}_{t,j}}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j\in[K]}\pi_{t,j}\left(\frac{\ell_{t,j}}{\tilde{\pi}_{t,j}}\right)^2\mathbb{E}_{t-1}\left[\mathbb{1}[I_t=j]^2\right]\right] \\
&= \mathbb{E}\left[\sum_{j\in[K]}\frac{\pi_{t,j}}{\tilde{\pi}_{t,j}}\right] \qquad (\ell_{t,j}\leq 1) \qquad (54) \\
&\leq 2K, \qquad (55)
\end{aligned}$$

where the last inequality holds since $\frac{\pi_{t,j}}{\tilde{\pi}_{t,j}} \leq 2$ as we have

$$2\tilde{\pi}_{t,i} - \pi_{t,i} = 2\left((1-\gamma)\pi_{t,i} + \frac{\gamma}{K}\right) - \pi_{t,i} = (1-2\gamma)\,\pi_{t,i} + 2\gamma\frac{1}{K} \geq \min\{\pi_{t,i}, \frac{1}{K}\} \geq 0.$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\ell}_{t,i}\right)^2\right] &= \mathbb{E}\left[\left(\frac{\ell_{t,i}\mathbb{1}[I_t = i]}{\tilde{\pi}_{t,i}}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\ell_{t,i}}{\tilde{\pi}_{t,i}}\right)^2 \mathbb{E}_{t-1}\left[\mathbb{1}[I_t = i]^2\right]\right] \\
&= \mathbb{E}\left[\frac{\ell_{t,i}^2}{\tilde{\pi}_{t,i}}\right] \\
&\leq \mathbb{E}\left[\frac{1}{\tilde{\pi}_{t,i}}\right]. && (\ell_{t,i} \leq 1)
\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[\sum_{j\in[K]} \pi_{t,j}\left(\hat{\ell}_{t,j}\right)^2\right] \leq 2K + \mathbb{E}\left[\frac{1}{\tilde{\pi}_{t,i}}\right] \leq 2K + \frac{K}{\gamma} = O(\frac{K}{\gamma}),$$

where the last inequality holds because we have $\tilde{\pi}_{t,i} = (1-\gamma)\pi_{t,i} + \frac{\gamma}{K} \geq \frac{\gamma}{K}$.