Proceedings of Machine Learning Research 129:545–560, 2020 ACML 2020

Thompson Sampling for Unsupervised Sequential Selection

Arun Verma V.ARUN@IITB.AC.IN
Manjesh K. Hanawal MHANAWAL@IITB.AC.IN
Nandyala Hemachandra NH@IITB.AC.IN
Indian Institute of Technology Bombay, India

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Thompson Sampling has generated significant interest due to its better empirical performance than upper confidence bound based algorithms. In this paper, we study Thompson Sampling based algorithm for *Unsupervised Sequential Selection* (USS) problem. The USS problem is a variant of the stochastic multi-armed bandits problem, where the loss of an arm can not be inferred from the observed feedback. In the USS setup, arms are associated with fixed costs and are ordered, forming a cascade. In each round, the learner selects an arm and observes the feedback from arms up to the selected arm. The learner's goal is to find the arm that minimizes the expected total loss. The total loss is the sum of the cost incurred for selecting the arm and the stochastic loss associated with the selected arm. The problem is challenging because, without knowing the mean loss, one cannot compute the total loss for the selected arm. Clearly, learning is feasible only if the optimal arm can be inferred from the problem structure. As shown in the prior work, learning is possible when the problem instance satisfies the so-called 'Weak Dominance' (WD) property. Under WD, we show that our Thompson Sampling based algorithm for the USS problem achieves near-optimal regret and has better numerical performance than existing algorithms.

Keywords: Sequential Decision Making, Partial Monitoring System, Thompson Sampling

1. Introduction

Many variants of sequential decision-making problems are considered in the literature depending on the type of feedback and the amount of information they reveal about the rewards. The multi-armed bandits and the expert setting (Auer et al., 2002; Bubeck et al., 2012) are well-studied problems where feedback provides direct information about the rewards. In the multi-armed bandit setting, feedback observed from an action reveals only the reward associated with that action. However, in the expert setting, the feedback observed from an action reveals reward associated with the action played as well as all other actions. The settings that span in between these two extreme cases are also studied, namely, bandits with side-information (Mannor and Shamir, 2011; Alon et al., 2013, 2015; Wu et al., 2015). In many problems, the actions can be indirectly tied to the rewards. Such setting is referred as partial monitoring setting (Cesa-Bianchi et al., 2006; Bartók and Szepesvári, 2012; Bartók et al., 2014). It includes all the previously described setups as special cases.

Most of the previous work on partial monitoring is restricted to cases where feedback from the actions allows the learner to identify the rewards of the actions. However, in many areas like crowd-sourcing (Bonald and Combes, 2017; Kleindessner and Awasthi, 2018), medical diagnosis (Hanawal et al., 2017), resource allocation (Verma et al., 2019a), and many others, feedback from actions may not even be sufficient to identify their rewards.

Such reward structures can be found in many prediction problems, where one may have to predict labels for instances whose associated ground-truth cannot be obtained. Such problems arise naturally in medical diagnosis, crowd-sourcing, security system (Hanawal et al., 2017), and unsupervised features selection (Verma et al., 2020). In the medical diagnosis problem, the true state of the patients may not be known; hence, the test's effectiveness cannot be known. Whereas in the crowd-sourcing systems, the expertise level of self-listed-agents (workers) is unknown; therefore, the quality of their work cannot be known. In these prediction problems, we can observe prediction from test/worker, but we cannot ascertain their reliability due to the absence of ground truth.

In many of the real-world situations like those found in medical diagnosis, airport security, and manufacturing, a set of tests or classifiers is used to monitor patients, people, and products. Tests have cost with the more informative ones resulting in higher monetary costs and higher latency. Thus, they are often organized as a cascade (Chen et al., 2012; Trapeznikov and Saligrama, 2013), so that a new input is first probed by an inexpensive test then more expensive one. We refer to such cascaded systems as *Unsupervised Sequential Selection* (USS) problem[1], where an arm represents a test/ worker. A learner's goal in the USS problem is to select the most cost-effective arm so that the overall system maintains high accuracy at low average costs.

In this paper, we draw upon several concepts introduced in prior work (Hanawal et al., 2017; Verma et al., 2019b). Specifically, we use the notion of weak dominance (Verma et al., 2019b) that helps to find optimal arm using observed disagreements between arms. We propose a Thompson Sampling (Agrawal and Goyal, 2012; Kaufmann et al., 2012; Agrawal and Goyal, 2013) based algorithm for the USS problem and show that it is a near-optimal algorithm. We then validate its performance on several problem instances derived from synthetic and real datasets. Our contributions can be summarized as follows:

- We develop a Thompson Sampling based algorithm named USS-TS for the USS problem. This algorithm uses a one-sided test to find the optimal arm, whereas the state-of-the-art algorithm proposed in Verma et al. (2019b) uses a two-sided test to identify the optimal arm. The new one-sided test leads to a simpler algorithm.

- In Section 4, we characterize the regret of USS-TS in terms of how well the problem instance satisfies the WD property and show that it has sub-linear regret under WD property. We also give problem independent regret bound and establish that the regret bounds are near-optimal using results from the partial monitoring system.

- We demonstrate empirical performance of USS-TS on synthetic and real datasets in Section 5. Our experimental results show that regret of USS-TS is always lower than USS-UCB (Verma et al., 2019b) and heuristic algorithm given in Hanawal et al. (2017).

1. Note that the unsupervised sequential selection problem is referred to as the unsupervised sensor selection problem in the prior work (Hanawal et al., 2017; Verma et al., 2019b).

2. Problem Setting

We consider a stochastic K-armed bandits problem. The set of arms is denoted by $[K]$ where $[K] \doteq \{1, 2, \ldots, K\}$. In each round t, the environment generates a binary $K+1$-dimensional vector $\left(Y_t, \{Y_t^i\}_{i \in [K]}\right)$. The variable Y_t denotes the best binary feedback for round t, which is hidden from the learner. The vector $\left(\{Y_t^i\}_{i \in [K]}\right) \in \{0,1\}^K$ represents observed feedback at time t, where Y_t^i denote the feedback[2] observed after playing arm i. We denote the cost for using arm $i \in [K]$ as $c_i \geq 0$ that is known to learner and the same for all rounds.

In the USS setup, the arms are assumed to be ordered and form a cascade. When the learner selects an arm $i \in [K]$, the feedback from all arms till arm i in the cascade is observed. The expected loss of playing the arm i is denoted as $\gamma_i \doteq \mathbb{E}\left[\mathbb{1}_{\{Y^i \neq Y\}}\right] = \mathbb{P}\left\{Y^i \neq Y\right\}$, where $\mathbb{1}_{\{A\}}$ denotes indicator of event A. The *expected total cost* incurred by playing arm i is defined as $\gamma_i + \lambda_i C_i$, where $C_i \doteq c_1 + \ldots + c_i$ and λ_i is a trade-off parameter that normalizes the loss and the incurred cost of playing arm i.

Since the best binary feedback are hidden from the learner, the expected loss of an arm cannot be inferred from the observed feedback. We thus have a version of the stochastic partial monitoring problem, and we refer to it as unsupervised sequential selection (USS) problem. Let $\boldsymbol{Q}$ be the unknown joint distribution of $(Y, Y^1, Y^2 \ldots, Y^K)$. Henceforth we identify an USS instance as $P \doteq (\boldsymbol{Q}, \boldsymbol{c})$ where $\boldsymbol{c} \doteq (c_1, c_2, \ldots, c_K)$ is the known cost vector of arms. We denote the collection of all USS instances as $\mathcal{P}_{\text{USS}}$. For instance $P \in \mathcal{P}_{\text{USS}}$, the optimal arm is given by

$$i^\star \in \max\left\{\arg\min_{i \in [K]} \left(\gamma_i + \lambda_i C_i\right)\right\} \tag{1}$$

where the 'max' operator selects the arm with the largest index among the minimizers. The choice of $i^\star$ in Eq. (1) is risk-averse as we prefer the arm with lower error among the good arms. The interaction between the environment and a learner is given in Algorithm 1.

Algorithm 1 Learning with USS instance $(\boldsymbol{Q}, \boldsymbol{c})$

For each round t:

1. **Environment** chooses a vector $(Y_t, \{Y_t^i\}_{i \in [K]}) \sim \boldsymbol{Q}$.
2. **Learner** selects an arm $I_t \in [K]$ to stop in cascade.
3. **Feedback and Loss:** The learner observes feedback $(Y_t^1, Y_t^2, \ldots, Y_t^{I_t})$ and incurs a total loss $\mathbb{1}_{\{Y^{I_t} \neq Y_t\}} + \lambda_{I_t} C_{I_t}$.

The learner's goal is to learn a policy that find an arm such that the cumulative expected loss is minimized. Specifically, for T rounds, we measure the performance of a policy that selects an arm I_t in round t in terms of regret given by

$$\mathfrak{R}_T = \sum_{t=1}^{T} \left(\gamma_{I_t} + \lambda_{I_t} C_{I_t} - \left(\gamma_{i^\star} + \lambda_{i^\star} C_{i^\star}\right)\right). \tag{2}$$

2. In the USS setup, an arm i could represent a classifier. After using the first i classifiers, the final label can be a function of labels predicted by the first i classifiers, $i \in [K]$.

A good policy should have sub-linear regret, i.e., $\lim_{T\to\infty} \mathfrak{R}_T/T = 0$. The sub-linear regret implies that the learner collects almost as much reward in expectation in the long run as an oracle that knew the optimal arm from the first round. We say that a problem instance $P \in \mathcal{P}_{\text{USS}}$ is learnable if there exists a policy with sub-linear regret.

3. Conditions for Learning Optimal Arm

Next, we define the strong and weak dominance property of the USS problem instance that makes the learning of the optimal arm possible.

Definition 1 (Strong Dominance (SD) (Hanawal et al., 2017)). *A problem instance is said to satisfy* SD *property if*

$$Y^i = Y \textit{ for some } i \in [K] \implies Y^j = Y, \quad \forall j > i.$$

We represent the set of all instances in $\mathcal{P}_{USS}$ that satisfy SD *property by $\mathcal{P}_{\text{SD}}$.*

The SD property implies that if the feedback of an arm is same as the true reward, then the feedback of all the arms in the subsequent stages of the cascade is also same as the true reward. Hanawal et al. (2017) show that the set of all instances satisfying SD property is learnable by mapping such instances to stochastic multi-armed bandits problem with side information (Wu et al., 2015). A weaker version of the SD property is defined as follows:

Definition 2 (Weak Dominance (WD) (Verma et al., 2019b)). *Let $i^\star$ denote the optimal arm. Then an instance $P \in \mathcal{P}_{USS}$ is said to satisfy weak dominance property if*

$$\forall j > i^\star : C_j - C_{i^\star} > \mathbb{P}\left\{Y^{i^\star} \neq Y^j\right\}. \tag{3}$$

We denote the set of all instances in $\mathcal{P}_{USS}$ that satisfy WD *property by $\mathcal{P}_{\text{WD}}$.*

The set of problems satisfying the WD property is maximally learnable, and any relaxation of WD property makes the problem unlearnable (Verma et al., 2019b, Theorem 1). In the following equation, we use an alternative characterization of the WD property, given as

$$\xi \doteq \min_{j>i^\star} \left\{C_j - C_{i^\star} - \mathbb{P}\left\{Y^{i^\star} \neq Y^j\right\}\right\} > 0. \tag{4}$$

The larger the value of ξ, 'stronger' is the WD property, and easier to identify an optimal arm. We later characterize the regret upper bound of our algorithm in terms of ξ.

3.1. Optimal Arm Selection

Without loss of generality, we set $\lambda_i = 1$ for all $i \in [K]$ as their value can be absorbed into the costs. Since $i^\star = \max\left\{\arg\min_{i\in[K]} (\gamma_i + C_i)\right\}$, it must satisfy following equation:

$$\forall j < i^\star \, : \, C_{i^\star} - C_j \leq \gamma_j - \gamma_{i^\star} \, , \tag{5a}$$

$$\forall j > i^\star \, : \, C_j - C_{i^\star} > \gamma_{i^\star} - \gamma_j \, . \tag{5b}$$

As the loss of an arm is not observed, the above equations can not lead to a sound arm selection criteria. We thus have to relate the unobservable quantities in terms of the quantities that can be observed. In our setup, we can compare the feedback of two arms, which can be used to estimate their disagreement probability. For notation convenience, we define $p_{ij} \doteq \mathbb{P}\left\{Y^i \neq Y^j\right\}$. The value of p_{ij} can be estimated as it is observable. We use the following result from Hanawal et al. (2017) that relates the differences in the unobserved error rates in terms of their observable disagreement probability.

Proposition 1 (Proposition 3 in Hanawal et al. (2017))**.** *For any two arms i and j, $\gamma_i - \gamma_j = p_{ij} - 2\mathbb{P}\left\{Y^i = Y, Y^j \neq Y\right\}$.*

Now, using Proposition 1, we can replace Eq. (5a) by

$$\forall j < i^\star : C_{i^\star} - C_j \leq p_{ji^\star}, \tag{6}$$

which only has observable quantities. For $j > i^\star$, we can replace Eq. (5b) by using the WD property as follows:

$$\forall j > i^\star : C_j - C_{i^\star} > p_{i^\star j}. \tag{7}$$

Using Eq. (6) and Eq. (7), our next result gives the optimal arm for a problem instance.

Lemma 1. *Let $P \in \mathcal{P}_{\text{WD}}$ and $\mathcal{B} = \{i : \forall j > i, C_j - C_i > p_{ij}\} \cup \{K\}$. Then the arm $I_t = \min(\mathcal{B})$ is the optimal arm for the problem instance P.*

Proof. Let $i^\star$ be an optimal arm for the problem instance P. Since $p_{i^\star j} \doteq \mathbb{P}\left\{Y^{i^\star} \neq Y^j\right\}$, we have $\forall j < i^\star : C_{i^\star} - C_j \leq \mathbb{P}\left\{Y^{i^\star} \neq Y^j\right\} \implies C_{i^\star} - C_j \ngtr \mathbb{P}\left\{Y^{i^\star} \neq Y^j\right\} \implies j \notin \mathcal{B}, \forall j < i^\star$. If any sub-optimal arm $h \in \mathcal{B}$ then the index of arm h must be larger than the index of optimal arm $i^\star$ in the cascade. Hence the element of the set $\mathcal{B}$ in round t is given as follows:

$$\mathcal{B} = \{i^\star, h_1, \ldots, h_t, K\},$$

where $i^\star < h_1 < \cdots < h_t < K$. By construction of set $\mathcal{B}$, the minimum indexed arm in set $\mathcal{B}$ is the optimal arm. □

Remark 1. *The* WD *property holds trivially for the problem instances that satisfy* SD *property as the difference of mean losses is the same as the disagreement probability between two arms due to $\mathbb{P}\left\{Y_t^i = Y_t, Y_t^j \neq Y_t\right\} = 0$ for $j > i$. Also, by definition, the* WD *property holds for all problem instances where the last arm of the cascade is an optimal arm.*

4. Thompson Sampling based Algorithm for USS

Upper Confidence Bound (UCB) based methods are useful for dealing with the trade-off between exploration and exploitation in bandit problems (Auer et al., 2002; Garivier and Cappé, 2011). UCB has been widely used for solving various sequential decision-making problems. On the other hand, Thompson Sampling (TS) is an online algorithm based on Bayesian updates. TS selects an arm to play according to its probability of being the best arm, and it is shown that TS is empirically superior then UCB based algorithms for various MAB problems (Chapelle and Li, 2011). TS also achieves lower bound for MAB when rewards of arms have Bernoulli distribution, as shown by Kaufmann et al. (2012).

4.1. Algorithm: USS-TS

We develop a Thompson Sampling based algorithm, named USS-TS, that uses Lemma 1 to select optimal arm. The algorithm works as follows: It sets the prior distribution of disagreement probability for each pair of arms as the Beta distribution, Beta$(1, 1)$, which is the same as Uniform distribution on $[0, 1]$. The variable S_{ij} represents the number of rounds when a disagreement is observed between arm i and j. Whereas, the variable F_{ij} represents the number of rounds when an agreement is observed. The variables $S_{ij}^{(t)}$ and $F_{ij}^{(t)}$ denote the values of S_{ij} and F_{ij} at the beginning of round t.

USS-TS Thompson Sampling based Algorithm for Unsupervised Sequential Selection

```
Set ∀1 ≤ i < j ≤ K : S_ij^(1) ← 1, F_ij^(1) ← 1
for t = 1, 2, ... do
  Set i = 1 and I_t = 0
  while I_t = 0 do
    Play arm i
    ∀j ∈ [i+1, K] : compute p̃_ij^(t) ← Beta(S_ij^(t), F_ij^(t))
    If ∀j ∈ [i+1, K] : C_j − C_i > p̃_ij^(t) or i = K then set I_t = i else set i = i + 1
  end while
  Select arm I_t and observe Y_t^1, Y_t^2, ..., Y_t^{I_t}
  ∀1 ≤ i < j ≤ I_t : update S_ij^(t+1) ← S_ij^(t) + 1{Y_t^i ≠ Y_t^j}, F_ij^(t+1) ← F_ij^(t) + 1{Y_t^i = Y_t^j}
end for
```

In round t, the learner plays the arm $i = 1$ and then observe its feedback. For each (i, j) pair, a sample $\tilde{p}_{ij}^{(t)}$ is independently drawn from Beta$(S_{ij}^{(t)}, F_{ij}^{(t)})$. Then algorithm checks whether the arm i is the best arm using Eq. (7) with $\tilde{p}_{ij}^{(t)}$ in place of $p_{ij}^{(t)}$. If the arm i is not the best, then the algorithm plays the next arm, and the same process is repeated. If the arm i is the best arm for the round t, then the algorithm stops at arm $I_t = i$ in the round t.

After selecting arm I_t, the feedback from arms $1, \ldots, I_t$ are observed, which is used to update the values of $S_{ij}^{(t+1)}$ and $F_{ij}^{(t+1)}$. The same process is repeated in the subsequent rounds.

Remark 2. *USS-TS is adapted for the USS problem from the Thompson Sampling algorithm for stochastic multi-armed bandits. However, the feedback structure and the way arms are selected in the USS setup differ from that in the stochastic multi-armed bandits.*

4.2. Analysis

The following definitions and results are useful in subsequent proof arguments.

Definition 3. *For the optimal arm $i^\star$ and $j \in [K]$, define*

$$\xi_j \doteq \begin{cases} p_{i^\star j} - (C_{i^\star} - C_j), & \text{if } j < i^\star \quad (8a) \\ C_j - C_{i^\star} - p_{i^\star j}, & \text{if } j > i^\star \quad (8b) \end{cases}$$

where $p_{i^\star j} = \mathbb{P}\left\{Y^{i^\star} = Y^j\right\}$.

Note that the values of ξ_j for all $j \in [K]$ is positive under the WD property.

Definition 4 (Action Preference ($\succ_t$)). *USS-TS prefers the arm i over arm j in round t if:*

$$i \succ_t j \doteq \begin{cases} \tilde{p}_{ji}^{(t)} \geq C_i - C_j & \text{if } j<i \quad (9a) \\ \tilde{p}_{ij}^{(t)} < C_j - C_i & \text{if } j>i \quad (9b) \end{cases}$$

Definition 5 (Transitivity Property). *If $i \succ_t j$ and $j \succ_t k$ then $i \succ_t k$.*

Definition 6. *Let $\mathcal{H}_t$ denote the σ-algebra generated by the history of selected arms and observations at the beginning of the time t and given as follows:*

$$\mathcal{H}_t \doteq \left\{ I_s, \left\{Y_s^i\right\}_{i \leq I_s}, s = 1, \ldots, t-1 \right\},$$

where I_s denotes the arm selected and set $\left\{Y_s^i\right\}_{i \leq I_s}$ denotes the observations from arm 1 to I_s in the round s. Define $\mathcal{H}_1 \doteq \{\}$.

Fact 1 (Beta-Binomial equality, Fact 1 in Agrawal and Goyal (2012)). *Let $F_{\alpha,\beta}^{beta}(y)$ be the cumulative distribution function (cdf) of the beta distribution with integer parameters α and β. Let $F_{n,p}^{B}(\cdot)$ be the cdf of the binomial distribution with parameters n and p. Then,*

$$F_{\alpha,\beta}^{beta}(y) = 1 - F_{\alpha+\beta-1,y}^{B}(\alpha - 1).$$

Lemma 2 (Lemma 2 in Agrawal and Goyal (2013)). *Let $n \geq 0$ and $\hat{\mu}_n$ be the empirical average of n samples from Bernoulli(μ). Let $x < \mu$ and $q_n(x) \doteq 1 - F_{n\hat{\mu}_n+1, n(1-\hat{\mu}_n)+1}^{beta}(x)$ be the probability that the posterior sample from the Beta distribution with its parameter $n\hat{\mu}_n + 1, n(1-\hat{\mu}_n) + 1$ exceeds x. Then,*

$$\mathbb{E}\left[\frac{1}{q_n(x)} - 1\right] \leq \begin{cases} \frac{3}{\Delta(x)} & \text{if } n < 8/\Delta(x) \\ \Theta\left(\exp^{-\frac{n\Delta(x)^2}{2}} + \frac{\exp^{-nd(x,\mu)}}{(n+1)\Delta(x)^2} + \frac{1}{\exp^{\frac{n\Delta(x)^2}{4}} - 1}\right) & \text{if } n \geq 8/\Delta(x), \end{cases}$$

where $\Delta(x) \doteq \mu - x$ and $d(x, \mu) \doteq x \log\left(\frac{x}{\mu}\right) + (1-x)\log\left(\frac{1-x}{1-\mu}\right)$.

Recall that $p_{i^\star j}$ is the disagreement probability between arm $i^\star$ and j and $\tilde{p}_{i^\star j}^{(t)}$ is the sample of $p_{i^\star j}$ using Beta distribution with the t samples. Next, we bound the probability by which USS-TS selects the sub-optimal arm whose index is smaller than the optimal arm.

Definition 7. *For any $j < i^\star$, define $q_{j,t}$ as the probability*

$$q_{j,t} \doteq \mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} \geq p_{i^\star j} - \xi_j | \mathcal{H}_t\right\}.$$

Lemma 3. *Let $P \in \mathcal{P}_{\text{WD}}$ and satisfies the transitivity property. If $j < i^\star$ then the probability by which USS-TS selects any sub-optimal arm j over the optimal arm is given by*

$$\mathbb{P}\left\{I_t = j, j < i^\star | \mathcal{H}_t\right\} \leq \frac{(1 - q_{j,t})}{q_{j,t}} \mathbb{P}\left\{I_t \geq i^\star | \mathcal{H}_t\right\}.$$

Proof. If the sub-optimal arm j is selected then arm j is preferred over the arms whose indexed is larger than j (Lemma 1). Hence we have

$$\mathbb{P}\left\{I_t = j, j < i^\star | \mathcal{H}_t\right\} = \mathbb{P}\left\{j \succ_t k, \forall k > j, j < i^\star | \mathcal{H}_t\right\} \leq \mathbb{P}\left\{j \succ_t k, \forall k \geq i^\star, j < i^\star | \mathcal{H}_t\right\}.$$

Since the feedback from an arm is independent of the feedback of other arms,

$$= \mathbb{P}\left\{j \succ_t i^\star, j < i^\star | \mathcal{H}_t\right\} \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\}.$$

If arm j is preferred over the arm $i^\star$ then $\tilde{p}_{i^\star j}^{(t)} < C_{i^\star} - C_j$. As $C_{i^\star} - C_j = p_{i^\star j} - \xi_j$ for $j < i^\star$,

$$\begin{aligned}
&= \mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} < p_{i^\star j} - \xi_j | \mathcal{H}_t\right\} \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\} \\
&= \left(1 - \mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} \geq p_{i^\star j} - \xi_j | \mathcal{H}_t\right\}\right) \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\} \\
\implies \mathbb{P}\left\{I_t = j, j < i^\star | \mathcal{H}_t\right\} &\leq (1 - q_{j,t}) \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\}. \quad \text{(Definition 7)} \quad (10)
\end{aligned}$$

Similarly, the probability of selecting an arm whose index is larger than the optimal arm can be lower bounded as follows:

$$\begin{aligned}
\mathbb{P}\left\{I_t \geq i^\star | \mathcal{H}_t\right\} \geq \mathbb{P}\left\{I_t = i^\star | \mathcal{H}_t\right\} &\geq \mathbb{P}\left\{I_t = i^\star, i^\star \succ_t j, j < i^\star | \mathcal{H}_t\right\} \\
&= \mathbb{P}\left\{i^\star \succ_t k, \forall k > i^\star, i^\star \succ_t j, j < i^\star | \mathcal{H}_t\right\} \quad \text{(Lemma 1)} \\
&\geq \mathbb{P}\left\{i^\star \succ_t j, j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\} \quad \text{(Definition 5)} \\
&= \mathbb{P}\left\{i^\star \succ_t j, j < i^\star | \mathcal{H}_t\right\} \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star | \mathcal{H}_t\right\}.
\end{aligned}$$

If arm $i^\star$ is preferred over the arm j then $\tilde{p}_{i^\star j}^{(t)} \geq C_{i^\star} - C_j$. As $C_{i^\star} - C_j = p_{i^\star j} - \xi_j$ for $j < i^\star$,

$$\begin{aligned}
&= \mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} \geq p_{i^\star j} - \xi_j | \mathcal{H}_t\right\} \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star\right\} \\
\implies \mathbb{P}\left\{I_t \geq i^\star | \mathcal{H}_t\right\} &\geq q_{j,t} \mathbb{P}\left\{j \succ_t k, \forall k > i^\star, j < i^\star\right\}. \quad \text{(Definition 7)} \quad (11)
\end{aligned}$$

Combining the Eq. (10) and Eq. (11), we get

$$\mathbb{P}\left\{I_t = j, j < i^\star | \mathcal{H}_t\right\} \leq \frac{(1 - q_{j,t})}{q_{j,t}} \mathbb{P}\left\{I_t \geq i^\star | \mathcal{H}_t\right\}. \qquad \square$$

Lemma 4. *Let $P \in \mathcal{P}_{\mathrm{WD}}$ and satisfies the transitivity property. If s be the number of times the sub-optimal arm j is selected by USS-TS then, for any $j < i^\star$,*

$$\sum_{t=1}^{T} \mathbb{P}\left\{I_t = j, j < i^\star\right\} \leq \frac{24}{\xi_j^2} + \sum_{s \geq 8/\xi_j} \Theta\left(\exp^{-s\xi_j^2/2} + \frac{\exp^{-sd(p_{i^\star j} - \xi_j, p_{i^\star j})}}{(s+1)\xi_j^2} + \frac{1}{\exp^{s\xi_j^2/4} - 1}\right).$$

Proof. **(sketch)** Using Lemma 3 and property of conditional expectations, we can have $\sum_{t=1}^{T} \mathbb{P}\left\{I_t = j, j < i^\star\right\} = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{P}\left\{j I_t = j, j < i^\star | \mathcal{H}_t\right\}\right]$. By using some simple algebraic manipulations on quantity $\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{P}\left\{I_t = j, j < i^\star | \mathcal{H}_t\right\}\right]$ with Lemma 2, we can get the above stated upper bound. $\square$

The detailed proof of Lemma 4 and all other missing proofs appear in the supplementary material. Our next result is useful to bound the probability by which USS-TS prefers the sub-optimal arms whose index is larger than the optimal arm.

Lemma 5. *Let $\hat{p}_{i^\star j}^{(t)}$ be the empirical estimate of $p_{i^\star j}$ and $j > i^\star$. Then, for any $x_j > p_{i^\star j}$ and $y_j > x_j$,*

$$\sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j\right\} \leq \frac{\ln T}{d(x_j, y_j)} + 1.$$

Proof. Define $L_j(T) = \frac{\ln T}{d(x_j,y_j)}$. Let $N_j(t)$ be the number of times the output from arm j is observed in t rounds. Then, the given probability term can be decomposed into two parts:

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j\right\} &= \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) \leq L_j(T)\right\} + \\ &\qquad \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) > L_j(T)\right\} \\ &\leq L_j(T) + \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) > L_j(T)\right\}. \end{aligned} \tag{12}$$

The first term of the above decomposition is bounded trivially by $L_j(T)$. To bound the second term, we demonstrate that if $N_j(t)$ is large enough and event $\hat{p}_{i^\star j}^{(t)} \leq x_j$ is satisfied, then the probability that the event $\tilde{p}_{ij}^{(t)} > y_j$ happens, is small. Then,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) > L_j(T)\right\} \\ &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}_{\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) > L_j(T)\right\}}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}_{\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j, N_j(t) > L_j(T)\right\}} | \mathcal{H}_t\right]\right].\end{aligned}$$

Since $N_j(t)$ and $\hat{p}_{i^\star j}^{(t)}$ are determined by the history $\mathcal{H}_t$,

$$= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, N_j(t) > L_j(T)\right\}} \mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} > y_j | \mathcal{H}_t\right\}\right]. \tag{13}$$

Now, by definition, $\mathcal{S}_{i^\star j}(t) = \hat{p}_{i^\star j}^{(t)} N_j(t)$, and therefore, $\tilde{p}_{i^\star j}^{(t)}$ is a Beta$(\hat{p}_{i^\star j}^{(t)} N_j(t) + 1, (1 - \hat{p}_{i^\star j}^{(t)}) N_j(t) + 1)$ distributed random variable. A Beta(α, β) random variable is stochastically dominated by Beta(α', β') if $\alpha' \geq \alpha, \beta' \leq \beta$. Therefore, if $\hat{p}_{i^\star j}^{(t)} \leq x_j$, the distribution of $\tilde{p}_{i^\star j}^{(t)}$ is stochastically dominated by Beta$(x_j N_j(t) + 1, (1 - x_j) N_j(t))$. Therefore, given a history $\mathcal{H}_t$ such that $\hat{p}_{i^\star j}^{(t)} \leq x_j$ and $N_j(t) > L_j(T)$, we have

$$\mathbb{P}\left\{\tilde{p}_{i^\star j}^{(t)} > y_j | \mathcal{H}_t\right\} = 1 - F^{beta}_{x_j N_j(t)+1, (1-x_j) N_j(t)}(y_j).$$

Now, using Beta-Binomial equality (Fact 1), we obtain that for any fixed $N_j(t) > L_j(T)$,

$$1 - F^{beta}_{x_j N_j(t)+1,(1-x_j)N_j(t)}(y_j) = F^{B}_{N_j(t),y_j}(x_j N_j(t)) \qquad \text{(using Fact 1)}$$

Here $F^{B}_{N_j(t),y_j}(x_j N_j(t))$ is the cdf of Binomial distribution with parameter y_j and $N_j(T)$ observations. Let $\mathcal{S}'_t$ be the number of successes observed in $N_j(T)$ observations. Then,

$$\begin{aligned} 1 - F^{beta}_{x_j N_j(t)+1,(1-x_j)N_j(t)}(y_j) &= \mathbb{P}\left\{\mathcal{S}'_t \leq x_j N_j(t)\right\} \\ &= \mathbb{P}\left\{\frac{\mathcal{S}'_t}{N_j(t)} \leq x_j\right\} \\ &= \mathbb{P}\left\{\hat{y}_j \leq x_j\right\} && \text{(using } \hat{y}_j = \mathcal{S}'_t/N_j(t)) \\ &\leq \exp^{-N_j(t)d(x_j,y_j)} && \text{(using Chernoff-Hoeffding bound)} \\ &\leq \exp^{-L_j(t)d(x_j,y_j)}, && \text{(as } N_j(t) > L_j(T)) \end{aligned}$$

which is smaller than $1/T$ because $L_j(T) = \frac{\log(T)}{d(x_j,y_j)}$. Substituting, we get that for a history $\mathcal{H}_t$ such that $\hat{p}^{(t)}_{i^\star j} \leq x_j$ and $N_j(t) > L_j(T)$,

$$\mathbb{P}\left\{\tilde{p}^{(t)}_{i^\star j} > y_j | \mathcal{H}_t\right\} \leq \frac{1}{T}.$$

For other history $\mathcal{H}_t$, the indicator term $\mathbb{1}_{\left\{\hat{p}^{(t)}_{i^\star j} \leq x_j, N_j(t) > L_j(T)\right\}}$ in Eq. (13) will be 0 as either event $\hat{p}^{(t)}_{i^\star j} \leq x_j$ or event $N_j(t) > L_j(T)$ is violated. Summing over t, this bounds the right hand side term in Eq. (13) as follows:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}^{(t)}_{i^\star j} \leq x_j, \tilde{p}^{(t)}_{i^\star j} > y_j, N_j(t) > L_j(T)\right\} &\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\mathbb{1}_{\left\{\hat{p}^{(t)}_{i^\star j} \leq x_j, N_j(t) > L_j(T)\right\}}}{T}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{T}\right] \\ &= 1. \end{aligned}$$

Replacing the second term in Eq. (12) by its upper bound and $L_j(T)$ with its value,

$$\sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}^{(t)}_{i^\star j} \leq x_j, \tilde{p}^{(t)}_{i^\star j} > y_j\right\} \leq \frac{\ln T}{d(x_j, y_j)} + 1. \qquad \square$$

Lemma 6. *For any $x_j > p_{i^\star j}$,*

$$\sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}^{(t)}_{i^\star j} > x_j\right\} \leq \frac{1}{d(x_j, p_{i^\star j})}.$$

Proof. **(sketch)** This result is easily proved by using Chernoff-Hoeffding bound. See details in the supplementary material. $\square$

Lemma 7. *Let $P \in \mathcal{P}_{\text{WD}}$. For any $\varepsilon > 0$ and $j > i^\star$,*

$$\sum_{t=1}^{T} \mathbb{P}\left\{j \succ_t i^\star, j > i^\star\right\} \leq (1+\varepsilon)\frac{\ln T}{d(p_{i^\star j}, p_{i^\star j} + \xi_j)} + O\left(\frac{1}{\varepsilon^2}\right).$$

Proof. **(sketch)** Let $p_{i^\star j} < x_j < y_j < p_{i^\star j} + \xi_j$ where $j > i^\star$. Then, it can be easily shown that $\sum_{t=1}^{T} \mathbb{P}\left\{j \succ_t i^\star, j > i^\star\right\} \leq \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} \leq x_j, \tilde{p}_{i^\star j}^{(t)} > y_j\right\} + \sum_{t=1}^{T} \mathbb{P}\left\{\hat{p}_{i^\star j}^{(t)} > x_j\right\}$. The upper bound on first term of right hand side quantity is given by Lemma 5 and the upper bound of the second term of right hand side quantity is given by Lemma 6. Then, for $\varepsilon \in (0,1)$ with suitable values of x_j and y_j, we can get the above stated upper bound. □

Let $\Delta_j = C_j + \gamma_j - (C_{i^\star} + \gamma_{i^\star})$ be the sub-optimality gap for arm j. Now we state the problem dependent regret upper bound of USS-TS.

Theorem 1 (Problem Dependent Bound)**.** *Let $P \in \mathcal{P}_{\text{WD}}$ and satisfies the transitivity property. If $\varepsilon > 0$ then, the expected regret of USS-TS in T rounds is bounded by*

$$\mathfrak{R}_T \leq \sum_{j > i^\star} \frac{(1+\varepsilon)\ln T}{d(p_{i^\star j}, p_{i^\star j} + \xi_j)}\Delta_j + O\left(\frac{K - i^\star}{\varepsilon^2}\right),$$

Proof. **(sketch)** Let $M_j(T)$ is the number of times arm j is selected by USS-TS. Then, the regret of USS-TS is given by $\mathfrak{R}_T = \sum_{j \in [K]} \mathbb{E}\left[M_j(T)\right]\Delta_j = \sum_{j \in [K]} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}_{\{I_t = j\}}\right]\Delta_j = \sum_{j \in [K]} \sum_{t=1}^{T} \mathbb{P}\left\{I_t = j\right\}\Delta_j$. We divide the regret into two parts and it can be re-written as $\mathfrak{R}_T \leq \sum_{j < i^\star} \sum_{t=1}^{T} \mathbb{P}\left\{I_t = j, j < i^\star\right\}\Delta_j + \sum_{j > i^\star} \sum_{t=1}^{T} \mathbb{P}\left\{I_t = j, j > i^\star\right\}\Delta_j$. The first part of the regret is upper bounded by using Lemma 4. For the second part, when arm $I_t > i^\star$ is selected, then there exists at least one arm $k > i^\star$, which must be preferred over $i^\star$. Using transitivity property and a recursive argument, we can show that the selected arm is preferred over the optimal arm. Hence, $\sum_{j > i^\star} \sum_{t=1}^{T} \mathbb{P}\left\{I_t = j, j > i^\star\right\}\Delta_j$ can be upper bounded by $\sum_{j > i^\star} \sum_{t=1}^{T} \mathbb{P}\left\{j \succ_t i^\star, j > i^\star\right\}\Delta_j$. We can upper bound $\sum_{j > i^\star} \sum_{t=1}^{T} \mathbb{P}\left\{j \succ_t i^\star, j > i^\star\right\}\Delta_j$ by using Lemma 7 to get the above stated regret upper bound for USS-TS. □

Next we present problem independent bounds on the regret of USS-TS.

Theorem 2 (Problem Independent Bound)**.** *Let $P \in \mathcal{P}_{\text{WD}}$ and satisfies the transitivity property. Then the expected regret of USS-TS in T rounds*

- *for any instance in $\mathcal{P}_{\text{SD}}$ is bounded as*

$$\mathfrak{R}_T \leq O\left(\sqrt{KT\ln T}\right).$$

- *for any instance in $\mathcal{P}_{\text{WD}}$ is bounded as*

$$\mathfrak{R}_T \leq O\left((K\ln T)^{1/3} T^{2/3}\right).$$

Proof. **(sketch)** To get the above problem independent regret upper bound, we maximize the problem-dependent regret of USS-TS with respect to the value of ξ_j. □

Corollary 1. *Let $P \in \mathcal{P}_{\text{WD}}$ and satisfies the transitivity property. Then the expected regret of USS-TS on $\mathcal{P}_{\text{SD}}$ is $\tilde{O}(T^{1/2})$ and on $\mathcal{P}_{\text{WD}}$ it is $\tilde{O}(T^{2/3})$, where $\tilde{O}$ hides K and the logarithmic terms that are having T in them.*

Discussion on optimality of USS-TS: Stochastic partial monitoring problems can be classified as an 'easy,' 'hard,' or 'hopeless' problem with expected regret bounds of the order $\Theta(T^{1/2}), \Theta(T^{2/3})$, or $\Theta(T)$, respectively. And there exists no other class of problems in between (Bartók et al., 2014). The class $\mathcal{P}_{\text{SD}}$ is regret equivalent to a stochastic multi-armed bandit with side observations (Hanawal et al., 2017), for which regret scales as $\Theta(T^{1/2})$, hence $\mathcal{P}_{\text{SD}}$ resides in the easy class and our bound on it is near-optimal. Since $\mathcal{P}_{\text{WD}} \supsetneq \mathcal{P}_{\text{SD}}$, $\mathcal{P}_{\text{WD}}$ is not easy problem. Since $\mathcal{P}_{\text{WD}}$ is also learnable, it cannot be a hopeless problem. Therefore, the class $\mathcal{P}_{\text{WD}}$ is hard. We thus conclude that the regret bound of USS-TS is also near-optimal in T up to a logarithmic term.

5. Experiments

We evaluate the performance of USS-TS on different problem instances derived from synthetic and two real datasets: PIMA Indians Diabetes (Kaggle, 2016) and Heart Disease (Cleveland) (Detrano, 1998). The details of the used problem instances are given as follows.

Synthetic Dataset: We generate synthetic Bernoulli Symmetric Channel (BSC) dataset (Hanawal et al., 2017) as follows: The true binary feedback Y_t is generated from i.i.d. Bernoulli random variable with mean 0.7. The problem instance used in the experiment has three arms. We fix feedback as true binary feedback for the first arm with probability 0.6, second arm with probability 0.7, and third arm with probability 0.8. To ensure strong dominance, we impose the condition during data generation. When the feedback of arm 1 matches the true binary feedback, we introduce error up to 10% to the feedback of arm 2 and 3. We use five problem instances of the BSC dataset by varying the cumulative cost of playing the arms as given in Table 1.

Values/Arms	Arm 1	Arm 2	Arm 3	WD Property
Error-rate (γ_i)	0.3937	0.2899	0.1358	
Instance 1 Costs	**0.05**	0.285	0.45	✓
Instance 2 Costs	0.05	**0.1**	0.53	✓
Instance 3 Costs	**0.05**	0.3	0.45	✓
Instance 4 Costs	0.05	0.25	**0.29**	✓
Instance 5 Costs	0.1	**0.2**	0.41	×

Table 1: WD propoerty doesn't hold for Instance 5. Optimal arm's cost is in **red bold** font.

Real Datasets: An arm i represents a classifier whose prediction is treated as the feedback of the arm i. The disagreement label for (i, j) pair is computed using the labels of classifier (Clf.) i and j. In Heart Disease dataset, each sample has 12 features. We split the features into three subsets and train a logistic classifier on each subset. We associate 1st classifier with the first 6 features as input, including cholesterol readings, blood sugar, and rest-ECG. The 2nd classifier, in addition to the 6 features, utilizes the thalach, exang and oldpeak features, and the 3rd classifier uses all the features. In PIMA Indians Diabetes dataset, each sample has 8 features related to the conditions of the patient. We split the features into three subsets and train a logistic classifier on each subset. We associate 1st classifier with the first 6 features as input. These features include patient profile. The 2nd classifier, in addition to the 6 features, utilizes the feature on the glucose tolerance test, and the 3rd classifier uses all the previous features and the feature that gives values of insulin test. The PIMA Indians Diabetes dataset has 768 samples, whereas the Heart Disease dataset has only 297 samples. As 10000 rounds are used in our experiments, we select a sample from the original dataset in a round-robin fashion and give it as input to the algorithm. The details about the different costs used in five problem instances of the real datasets are given in Table 2.

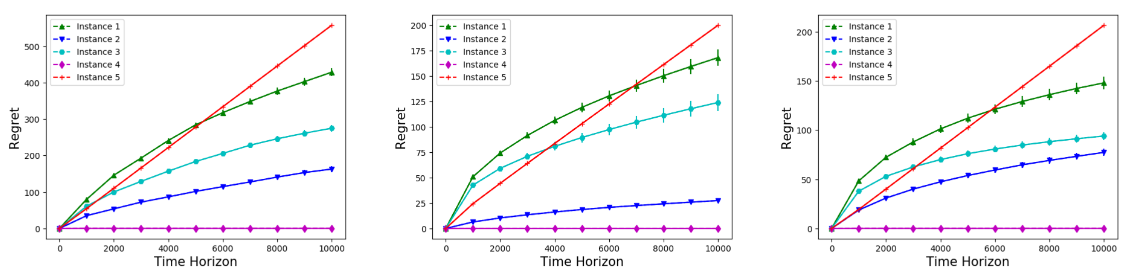
Values/ Classifiers (Arms)	**PIMA Indians Diabetes**			**Heart Disease**			WD Property
	Clf. 1	Clf. 2	Clf. 3	Clf. 1	Clf. 2	Clf. 3	
Error-rate (γ_i)	0.3098	0.233	0.2278	0.2929	0.2025	0.1483	
Instance 1 Costs	**0.05**	0.28	0.45	**0.02**	0.32	0.45	✓
Instance 2 Costs	0.2	**0.25**	0.269	0.2	**0.25**	0.395	✓
Instance 3 Costs	**0.05**	0.309	0.45	**0.02**	0.34	0.45	✓
Instance 4 Costs	0.2	0.25	**0.255**	0.2	0.25	**0.3**	✓
Instance 5 Costs	**0.05**	0.146	0.3	0.2	**0.25**	0.325	×

Table 2: Costs of different problem instances which are derived from real datasets. WD property doesn't hold for Instance 5 and cost of optimal arm is in **red bold** font.

Verifying WD **property:** The error-rate associated with each arm is known to us as given in Table 1 and Table 2 (but note that the error-rates are unknown to the algorithm); hence we can find an optimal arm for a given problem instance. After knowing optimal arm, WD property is verified by using the disagreement probability estimates after 10000 rounds.

5.1. Experimental Results

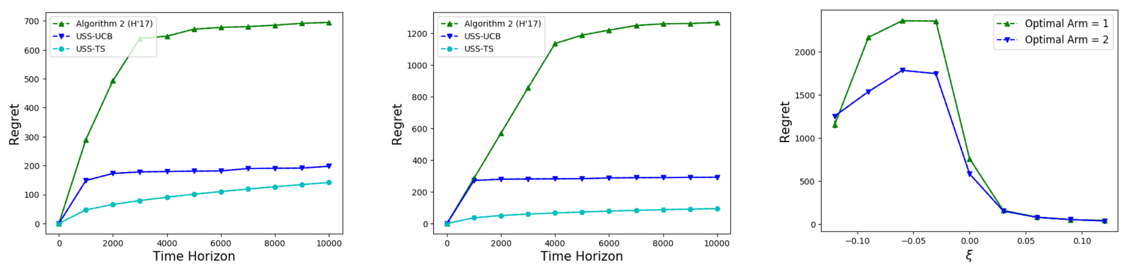
We fix the time horizon to 10000 in all experiments and repeat each experiment 500 times. The average regret is presented with a 95% confidence interval. The vertical line on each plot shows the confidence interval.



(*a*) BSC Dataset (*b*) PIMA Indians Diabetes (*c*) Heart Disease

Figure 1: Regret of USS-TS for different problem instances derived from synthetic and real datasets.

Expected Cumulative Regret v/s Time Horizon: The *Regret* of USS-TS versus *Time Horizon* plots for the different problem instances derived from BSC Dataset and two real datasets are shown in Figure 1. These plots verify that any instance that satisfies WD property has sub-linear regret. Note that USS-TS has linear regret for the Instance 5 as it does not satisfy WD property. We also compare the performance of USS-TS with existing UCB based algorithm USS-UCB algorithm of Verma et al. (2019b) with value of $\alpha = 0.5$ (best possible parameter value mentioned in the paper) and Algorithm 2 of Hanawal et al. (2017) with value of $\alpha = 1.5$ (as used in the paper) on Heart Disease and PIMA Indians Diabetes datasets. As expected, USS-TS outperforms other algorithms with large margins as shown in Fig. 2(a) (PIMA Indians Diabetes dataset) and Fig. 2(b) (Heart Disease dataset).



(a) PIMA Indians Diabetes (b) Heart Disease (c) BSC Dataset

Figure 2: Comparing regret of USS-TS with USS-UCB (Verma et al., 2019b) and Algorithm 2 (Hanawal et al., 2017) for real datasets (Fig. 2(a) and Fig. 2(b)). Regret behavior of USS-TS versus WD property for BSC Dataset is shown in Fig. 2(c).

Learnability v/s WD **Property:** We experiment with different problem instances of the BSC dataset to know the relationship between regret of USS-TS and WD property. We fixed an optimal arm and vary the cumulative cost of using arms in such a way that we pass from the case where WD property does not hold ($\xi \leq 0$ or $C_j - C_{i^\star} \in (\gamma_{i^\star} - \gamma_j, p_{i^\star j}]$ for any $j > i^\star$ where $\xi := \min_{j > i^\star} \xi_j$) to the situation where WD property holds ($\xi > 0$). When WD property does not hold for any problem instance, USS-TS treats a sub-optimal arm as the optimal arm. In such problem instances, as $C_j - C_{i^\star}$ increases, the regret will also increase due to selection of sub-optimal arm by USS-TS until WD property does not satisfy for that problem instance. When WD property does not satisfy for a problem instance then $C_j - C_{i^\star} \in (\gamma_{i^\star} - \gamma_j, p_{i^\star j}]$ holds in such cases, hence, it is easy to verify that ξ can not be smaller than $-\max(p_{i^\star j} - (\gamma_{i^\star} - \gamma_j))$.

We consider the problem instances with the minimum possible value of ξ for which problem instance satisfies WD property. Then we increase the value of ξ by increasing the cumulative cost of the arm. The regret versus ξ plots for BSC Dataset is shown in Fig. 2(c). It can be observed that there is a transition at $\xi = 0$. Through our experiments, we show that the stronger the WD property (large value of ξ) for the problem instance, it is easier to identify the optimal arm and, hence the less regret is incurred by USS-TS.

6. Conclusion

We studied the unsupervised sequential selection (USS) problem, where both accuracy and cost of using arms are important. It is a variant of the stochastic partial monitoring problem, where the losses are not observed. Still, one can compare the feedback of two arms to see if they agree or disagree. We estimate the disagreement probability between each pair of the arms and develop an algorithm named USS-TS that achieves near-optimal regret. We demonstrate our algorithms' performance on two real datasets and empirically show that any problem instance satisfying WD property has sub-linear regret. We ignored the inherent side observations due to the arms' cascade structure. By using these side observations, one can tighten the regret bounds. Another interesting future direction is to develop algorithms that relax the cascade structure assumption and selects the best subset of arms.

Acknowledgments

Manjesh K. Hanawal would like to thank the support from INSPIRE faculty fellowship from DST and Early Career Research (ECR) Award from SERB, Govt. of India.

References

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial intelligence and statistics*, pages 99–107, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Annual Conference on Learning Theory*, volume 40. Microtome Publishing, 2015.

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

Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pages 305–319. Springer, 2012.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Thomas Bonald and Richard Combes. A minimax optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, pages 4352–4360, 2017.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Minmin Chen, Zhixiang Xu, Kilian Weinberger, Olivier Chapelle, and Dor Kedem. Classifier cascade for minimizing feature evaluation cost. In *Artificial Intelligence and Statistics*, pages 218–226, 2012.

Robert Detrano. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, MD, Ph.D., Donor: David W. Aha, 1998. URL https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

Manjesh Hanawal, Csaba Szepesvari, and Venkatesh Saligrama. Unsupervised sequential sensor acquisition. In *Artificial Intelligence and Statistics*, pages 803–811, 2017.

UCI Machine Learning, Kaggle. Pima Indians Diabetes Database, 2016. URL https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Matthäus Kleindessner and Pranjal Awasthi. Crowdsourcing with arbitrary adversaries. In *International Conference on Machine Learning*, pages 2713–2722, 2018.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

Kirill Trapeznikov and Venkatesh Saligrama. Supervised sequential classification under budget constraints. In *Artificial Intelligence and Statistics*, pages 581–589, 2013.

Arun Verma, Manjesh Hanawal, Arun Rajkumar, and Raman Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. In *Advances in Neural Information Processing Systems*, pages 14499–14509, 2019a.

Arun Verma, Manjesh Hanawal, Csaba Szepesvari, and Venkatesh Saligrama. Online algorithm for unsupervised sensor selection. In *Artificial Intelligence and Statistics*, pages 3168–3176, 2019b.

Arun Verma, Manjesh K Hanawal, and Nandyala Hemachandra. Unsupervised online feature selection for cost-sensitive medical diagnosis. In *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 1–6. IEEE, 2020.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.