Unsupervised Sequential Sensor Acquisition

Manjesh K. Hanawal
Dept. of IEOR
IIT Bombay, India
mhanawal@iitb.ac.in

Csaba Szepesvári
Dept. of Computing Sciences
University of Alberta, Canada
szepesva@cs.ualberta.ca

Venkatesh Saligrama
Dept. of ECE
Boston University, USA
srv@bu.edu

Abstract

In many security and healthcare systems a sequence of sensors/tests are used for detection and diagnosis. Each test outputs a prediction of the latent state, and carries with it inherent costs. Our objective is to *learn* strategies for selecting tests to optimize accuracy & costs. Unfortunately it is often impossible to acquire in-situ ground truth annotations and we are left with the problem of unsupervised sensor selection (USS). We pose USS as a version of stochastic partial monitoring problem with an *unusual* reward structure (even noisy annotations are unavailable). Unsurprisingly no learner can achieve sublinear regret without further assumptions. To this end we propose the notion of weak-dominance. This is a condition on the joint probability distribution of test outputs and latent state and says that whenever a test is accurate on an example, a later test in the sequence is likely to be accurate as well.We empirically verify that weak dominance holds on real datasets and prove that it is a maximal condition for achieving sublinear regret. We reduce USS to a special case of multi-armed bandit problem with side information and develop polynomial time algorithms that achieve sublinear regret.

1 Introduction

Sequential sensor selection arises in many security and healthcare diagnostic systems. In these applications we have a diverse collection of sensor-based-tests with differing costs and accuracy. In these applications (see Fig. 1) inexpensive tests are first conducted and based on their outcomes a decision for acquiring more (expensive) tests are made. The goal in these systems is to maximize overall accuracy for an available cost-budget. Generally, the components that can be optimized include sensor classifiers (to improve test accuracy), sensor ordering, and decision strategies for sequential sensor selection. Nevertheless, sensor classifiers and sensor ordering are typically part of the infrastructure and generally harder to control/modify in both security and medical systems. To this end we focus here on the sequential sensor selection problem and use the terms sensor and test interchangeably. The need for systematically learning optimal decision strategies for balancing accuracy & costs arises from the fact that these applications involve legacy systems where sensor/test selection strategies are local; often managed under institutional, rather than national guidelines [2]. While it is possible to learn such decision strategies given sufficient annotated training data, what makes these applications challenging is that it is often difficult to acquire in-situ ground truth labels.

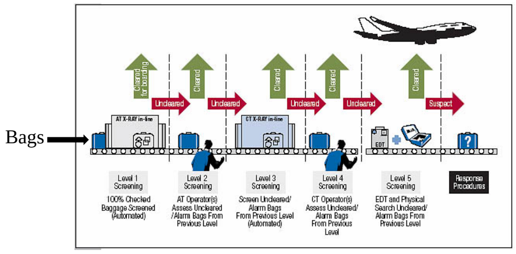
Bags
Cleared for boarding
Cleared
Cleared
Cleared
Cleared
Uncleared
Uncleared
Uncleared
Uncleared
Suspect
AT X-RAY in-line
CT X-RAY in-line
EDT
?
Level 1 Screening
100% Checked Baggage Screened (Automated)
Level 2 Screening
AT Operator(s) Assess Uncleared /Alarm Bags From Previous Level
Level 3 Screening
Screen Uncleared/ Alarm Bags From Previous Level (Automated)
Level 4 Screening
CT Operator(s) Assess Uncleared/ Alarm Bags From Previous Level
Level 5 Screening
EDT and Physical Search Uncleared/ Alarm Bags From Previous Level
Response Procedures

Figure 1: Sequential Sensor/Test Selection in Airport Security Systems. A number of different imaging and non-imaging tests are sequentially processed (see [1]). Costs can arise due to sensor availability and delay. Inexpensive tests are first conducted and based on their outcomes more expensive tests are conducted.

These observations motivate the problem of learning decision strategies for optimal sensor selection in situations where we do not have the benefit of ground-truth annotations, and what we refer to as the Unsupervised Sensor Selection (USS) problem. In Section 2 we pose our problem as a version of stochastic partial monitoring problem [3] with *atypical* reward structure, where tests are viewed as actions and sequential observations serve as side information. As is common, we pose

Appearing in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the authors.

the problem in terms of competitive optimality. We consider a competitor who can choose an optimal test with the benefit of hindsight. Our goal is to minimize cumulative regret, the extra cost incurred due to the initial ignorance of the learner.

In Section 3 we first show (unsurprisingly) that no learner can achieve sublinear regret without further assumptions. To this end we propose the notions of weak and strong dominance, which correspond to constraints on joint probability distributions over the latent state and test-outcomes. Strong Dominance (SD) is a property arising in many engineered systems and says that whenever a test is accurate on an example, a later test in the sequence is almost surely accurate on that example. Weak Dominance (WD) is a relaxed notion that allows for errors in these predictions. We empirically demonstrate that WD holds by evaluating it on several real datasets. We also show that WD is fundamental in the sense that while there exist learners that achieve sublinear regret over WD instances, no new instances can be added to this class without losing this property.

In Section 4 under SD we show that USS is reducible to a version of multi-armed bandit problem (MAB) with side-observation, a problem that is known to be learnable with sub-linear regret. In our reduction, we identify tests as bandit arms. The payoff of an arm is given by marginal losses relative to the root test, and the side observation structure is defined by the feedback graph induced by the directed graph. We then formally show that there is a one-to-one mapping between algorithms for USS and algorithms for MAB with side-observation. In particular, under SD, the regret bounds for MAB with side-observation then imply corresponding regret bounds for USS.

In Section 5 we give algorithms for the problem classes specified by either SD or WD. For SD instances the algorithm is based on the work of Wu et al. [4]. This algorithm is shown to enjoy an asymptotically optimal regret for problems satisfying SD.

1.1 Related Work

In contrast to our USS setup there exists a wide body of literature dealing with sensor selection (see [5]). Like us they also deal with cascade models with costs for features/tests but their method is based on training decision strategies with fully supervised data. There are also several methods that work in an online bandit setting and train prediction models with feature costs [6] but again they require true labels as reward-feedback. A somewhat more general version of [6] is developed in [7] where in addition the learner can choose to acquire true labels for a cost.

Our paper bears some similarity with the concept of active classification, which deals with learning stopping policies [8, 9] among a given sequence of tests. Like us these works also consider costs for utilizing tests and the goal is to learn when to stop to make decisions. Nevertheless, unlike our setup the loss associated with the decision is observed in their context.

Our paper is related to the framework of finite partial monitoring problems [3], which deals with how to infer unknown key information and where tests/actions reveal different types of information about the unknown information. In this context [10] consider special cases where payoff/rewards for a subset of actions are observed. This is further studied as a side-observation problem in [11] and as graph-structured feedback [12, 13, 4]. Our work is distinct from these setups because we are unable to observe rewards for our chosen actions or any other actions.

2 Unsupervised Sensor Selection

Preliminaries and Notation: Proofs for formal statements appear in the supplementary. All random variables are printed with upper case letters, while the reverse is not necessarily true. The set of real numbers is denoted by $\mathbb{R}$. For positive integer n, we let $[n] = \{1, \dots, n\}$. We let $M_1(\mathcal{X})$ to denote the set of probability distributions over some (measurable) set $\mathcal{X}$. When $\mathcal{X}$ is finite with a cardinality of $d \doteq |\mathcal{X}|$, $M_1(\mathcal{X})$ denotes the d-dimensional probability simplex. We let $\langle x, y \rangle = \sum_i x_i y_i$ denote the standard inner product of vectors x, y.

We first define the *unsupervised, stochastic, cascaded sensor selection* (USS) problem, a special subclass of stochastic partial monitoring problems (for completeness, we describe these problems in Appendix A). Later we will briefly describe extensions to tree-structures and contextual cases. An USS instance is specified by a pair $\theta = (P, c)$, where P is a distribution over the $K + 1$ dimensional binary hypercube and c is a K-dimensional, nonnegative valued vector of costs. While the learner knows c from the start, P is initially unknown. *Henceforth we identify problem instance θ and P* and e.g. will write $Y \sim \theta$ to denote $Y \sim P$. The instance parameters specify the learner-environment interaction as follows: In each round for $t = 1, 2, \dots$, the environment generates a $K + 1$-dimensional binary vector $Y = (Y_t, Y_t^1, \dots, Y_t^K)$ chosen at random from P. Here, Y_t^i is the output of sensor i, while Y_t is a (hidden) label to be guessed by the learner. Simultaneously, the learner chooses an index $I_t \in [K]$ and observes the sensor outputs $Y_t^1, \dots, Y_t^{I_t}$. The sensors are known to be ordered from least accurate to most accurate, i.e., $\gamma_k(\theta) \doteq \mathbb{P}\left(Y_t \neq Y_t^k\right)$ is decreasing with k increasing. Knowing this, the learner's choice of I_t also indicates that he/she chooses I_t to predict the unknown label

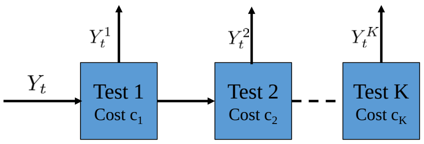

Figure 2: Cascaded Unsupervised Sequential Sensor Selection. Y_t is the hidden state of the instance and $Y_t^1, Y_t^2 \dots$ are test outputs. Not shown are features that a sensor could process to produce the output.

Y_t. Observing sensors is costly: The cost of choosing I_t is $C_{I_t} \doteq c_1 + \dots + c_{I_t}$. The total cost suffered by the learner in round t is thus $C_{I_t} + \mathbb{I}\{Y_t \neq Y_t^{I_t}\}$. The goal of the learner is to compete with the best choice given the hindsight of the values $(\gamma_k)_k$. Let $c(k, \theta) = \mathbb{E}\left[C_k + \mathbb{I}\{Y_t \neq Y_t^k\}\right] (= C_k + \gamma_k)$ and $c^*(\theta) = \min_k c(k, \theta)$. The expected regret of the learner up to the end of round T is $\mathfrak{R}_T = (\sum_{t=1}^T \mathbb{E}\left[c(I_t, \theta)\right]) - Tc^*(\theta)$.

Sublinear Regret: The quantification of the learning speed is given by the expected regret $\mathfrak{R}_T$, which, for brevity and when it does not cause confusion, we will just call regret. A sublinear regret, i.e., $\mathfrak{R}_T/T \to 0$ as $T \to \infty$ means that the learner in the long run collects almost as much reward on expectation as if the optimal action was known to it.

In what follows, we let $a^*(\theta)$ denote the optimal action that has the smallest index.[1] The optimality of actions can be captured in terms of marginal costs and marginal errors. In particular, an action i is optimal if for all $j > i$ the marginal increase in cost, $C_j - C_i$, is larger than the marginal decrease in error, $\gamma_i - \gamma_j$: $\forall\, j \geq i$

$$\underbrace{C_j - C_i}_{\text{Marginal Cost}} \geq \underbrace{E\left[\mathbb{I}\{Y_t \neq Y_t^i\} - \mathbb{I}\{Y_t \neq Y_t^j\}\right]}_{\text{Marginal Error } = \gamma_i - \gamma_j}. \quad (1)$$

3 When is USS Learnable?

Let Θ_{SA} be the set of all stochastic, cascaded sensor selection problems. Thus, $\theta \in \Theta_{\text{SA}}$ such that if $Y \sim \theta$ then $\gamma_k(\theta) \doteq \mathbb{P}\left(Y \neq Y^k\right)$ is a decreasing sequence. Given a subset $\Theta \subset \Theta_{\text{SA}}$, we say that Θ is *learnable* if there exists a learning algorithm $\mathfrak{A}$ such that for any $\theta \in \Theta$, the expected regret $\mathbb{E}\left[\mathfrak{R}_n(\mathfrak{A}, \theta)\right]$ of algorithm $\mathfrak{A}$ on instance θ is sublinear. A subset Θ is said to be a *maximal learnable problem class* if it is learnable and for any $\Theta' \subset \Theta_{\text{SA}}$ superset of Θ, Θ' is not learnable. In this section we study two special learnable problem classes, $\Theta_{\text{SD}} \subset \Theta_{\text{WD}}$, where the regularity properties of the instances in Θ_{SD} are more intuitive, while Θ_{WD} can be seen as a maximal extension of Θ_{SD}.

Let us start with some definitions. Given an instance $\theta \in \Theta_{\text{SA}}$, we can decompose θ (or P) into the joint distribution P_S of the sensor outputs $S = (Y^1, \dots, Y^k)$ and the conditional distribution of the state of the environment, given the sensor outputs, $P_{Y|S}$. Specifically, letting $(Y, S) \sim P$, for $s \in \{0,1\}^K$ and $y \in \{0,1\}$, $P_S(s) = \mathbb{P}\left(S = s\right)$ and $P_{Y|S}(y|s) = \mathbb{P}\left(Y = y | S = s\right)$. We denote this by $P = P_S \otimes P_{Y|S}$. A learner who observes the output of all sensors for long enough is able to identify P_S with arbitrary precision, while $P_{Y|S}$ remains hidden from the learner. This leads to the following statement:

Proposition 1. *A subset $\Theta \subset \Theta_{\text{SA}}$ is learnable if and only if there exists a map $a : M_1(\{0,1\}^K) \to [K]$ such that for any $\theta = (P, c) \in \Theta$ with decomposition $P = P_S \otimes P_{Y|S}$, $a(P_S)$ is an optimal action in θ. Following our previous convention, we also write $\theta = P_S \otimes P_{Y|S}$.*

An action selection map $a : M_1(\{0,1\}^K) \to [K]$ is said to be *sound* for an instance $\theta \in \Theta_{\text{SA}}$ with $\theta = P_S \otimes P_{Y|S}$ if $a(P_S)$ selects an optimal action in θ. With this terminology, the previous proposition says that a set of instances Θ is learnable if and only if there exists a sound action selection map for all the instances in Θ.

A class of sensor selection problems that contains instances that satisfy the so-called *strong dominance* condition will be shown to be learnable:

Definition 1 (Strong Dominance (SD))**.** *An instance $\theta \in \Theta_{\text{SA}}$ is said to satisfy the* strong dominance property *if it holds in the instance that if a sensor predicts correctly then all the sensors in the subsequent stages of the cascade also predict correctly, i.e., for any $i \in [K]$,*

$$Y^i = Y \;\Rightarrow\; Y^{i+1} = \dots = Y^K = Y \qquad (2)$$

almost surely (a.s.) where $(Y, Y^1, \dots, Y^K) \sim \theta$.

Before we develop this concept further we will motivate strong dominance based on experiments on a few real-world examples. First, SD naturally arises in the context of a cascade of error-correcting codes (see [14, 15]). On the other hand for "natural" systems SD holds "approximately" only. Table 1 lists the error probabilities of the classifiers (sensors) for the heart and diabetic datasets from UCI repository. We split features into two sets based on provided costs (cheap tests are based on patient history and costly tests include all the features). We then trained an SVM classifier with 5-fold cross-validation and report scores based on held-out test data. The last column lists the probability that second sensor misclassifies an instance that is correctly classified by the first sensor. SD is the notion that suggests that this probability is zero. We find in these datasets that δ_{12} is small thus justifying our notion. In general we have found this behavior is representative of other cost-associated datasets.

[1] Note that even if $i < j$ are optimal actions, there can be suboptimal actions in the interval $[i, j] (= [i, j] \cap \mathbb{N})$, e.g., $\gamma_1 = 0.3$, $C_1 = 0$, $\gamma_2 = 0.25$, $C_2 = 0.1$, $\gamma_3 = 0$, $C_3 = 0.3$.

Dataset	γ_1	γ_2	δ_{12}
PIMA Diabetes	0.32	0.23	0.065
Heart (Cleveland)	0.305	0.169	0.051

Table 1: Depicts approximate SD property on real datasets: $\gamma_1 \doteq \Pr(Y^1 \neq Y)$, $\gamma_2 \doteq \Pr(Y^2 \neq Y)$, $\delta_{12} \doteq \Pr(Y^1 = Y, Y^2 \neq Y)$

We next show that SD conditions ensures learnability. To this end, let $\Theta_{\text{SD}} = \{\theta \in \Theta_{\text{SA}} : \theta$ satisfies SD condition $\}$.

Theorem 2. *The set Θ_{SD} is learnable.*

In the following results we let $(Y, Y^1, \ldots, Y^K) \sim \theta$. The key to the proof of Theorem 2 is the following proposition:

Proposition 3. *Let $\gamma_i = \gamma_i(\theta)$ for $\theta \in \Theta_{\text{SA}}$, Then, for any $i, j \in [K]$, $\gamma_i - \gamma_j = \mathbb{P}\left(Y^i \neq Y^j\right) - 2\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$.*

Next we relax SD. The relaxation is developed through a series of smaller propositions. We start with a corollary to Proposition 3:

Corollary 4. *Let $i < j$. Then $0 \leq \gamma_i - \gamma_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$.*

The next two propositions consider the dual cases $i < j$ and $i > j$ and the decrease/increase of total costs:

Proposition 5. *Let $i < j$. Assume*

$$C_j - C_i \notin \left[\gamma_i - \gamma_j, \mathbb{P}\left(Y^i \neq Y^j\right)\right). \tag{3}$$

Then $\gamma_i + C_i \leq \gamma_j + C_j$ if and only if $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right)$.

Proposition 6. *Let $j < i$. Assume*

$$C_i - C_j \notin \left(\gamma_j - \gamma_i, \mathbb{P}\left(Y^i \neq Y^j\right)\right]. \tag{4}$$

Then, $\gamma_i + C_i \leq \gamma_j + C_j$ if and only if $C_i - C_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$.

These results motivate the following definition:

Definition 2 (Weak Dominance (WD))**.** *An instance $\theta \in \Theta_{\text{SA}}$ is said to satisfy the* weak dominance property *if for $i = a^*(\theta)$,*

$$\rho = \min_{j>i} \frac{C_j - C_i}{\mathbb{P}\left(Y^i \neq Y^j\right)} \geq 1. \tag{5}$$

We denote the set of all instances in Θ_{SA} that satisfies this condition by Θ_{WD}.

Note that $\Theta_{\text{SD}} \subset \Theta_{\text{WD}}$ since for any $\theta \in \Theta_{\text{SD}}$ and any $j > i = a^*(\theta)$, on the one hand $C_j - C_i \geq \gamma_i - \gamma_j$, while on the other hand, by SD, $\mathbb{P}\left(Y^i \neq Y^j\right) = \gamma_i - \gamma_j$.

We now relate WD to the optimality condition described in Eq. (1). WD can be viewed as a more stringent condition for optimal actions. For an action to be optimal we require that the marginal cost be larger than marginal *absolute* error, namely, for all $j > i$, with $i = a^*(\theta)$:

$$\underbrace{C_j - C_i}_{\text{Marginal Cost}} \geq \underbrace{E\left[\left|\mathbb{I}\{Y_t \neq Y_t^i\} - \mathbb{I}\{Y_t \neq Y_t^j\}\right|\right]}_{\text{Marginal Absolute Error}} \tag{6}$$

where we have re-written $\mathbb{P}\left(Y^i \neq Y^j\right)$ as the marginal absolute error.

We propose the following action selector $a_{\text{wd}} : M_1(\{0,1\}^K) \to [K]$:

Definition 3. *For $P_S \in M_1(\{0,1\}^K)$ let $a_{\text{wd}}(P_S)$ denote the smallest index $i \in [K]$ such that*

$$\forall j < i \ : \ C_i - C_j < \mathbb{P}\left(Y^i \neq Y^j\right), \tag{7a}$$

$$\forall j > i \ : \ C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right), \tag{7b}$$

where $C_i = c_1 + \cdots + c_i$, $i \in [K]$ and $(Y^1, \ldots, Y^K) \sim P_S$. (If no such index exists, a_{wd} is undefined, i.e., a_{wd} is a partial function.)

The action selector a_{wd} is sound for any $\theta \in \Theta_{\text{WD}}$ and is in in fact essentially the only sound action selector map defined for all instances of Θ_{WD}. Further, the set Θ_{WD} is essentially a maximal learnable set in the dom(a_{wd}), i.e., Θ_{WD} is learnable but not uniformly learnable (see Appendix B for formal statements and proofs).

4 Regret Equivalence

In this section we establish that USS with SD property is 'regret equivalent' to an instance of multi-armed-bandit (MAB) with side-information [11]. The corresponding MAB algorithm can then be suitably imported to solve USS efficiently. Recall that in a MAB with side-information when in some round a decision maker chooses an action a it receives noisy rewards for all actions in $\mathcal{N}(a)$. In the simplest case, $\mathcal{N}(a)$ is known ahead and $a \in \mathcal{N}(a)$. The challenge is to minimize regret as usual.

Let $\mathcal{P}_{\text{USS}}$ be the set of USSs with action set $\mathcal{A} = [K]$. The corresponding bandit problems will have the same action set, while for action $k \in [K]$ the neighborhood set is $\mathcal{N}(k) = [k]$. Take any instance $(P, c) \in \mathcal{P}_{\text{USS}}$ and let $(Y, Y^1, \ldots, Y^K) \sim P$ be the unobserved state of environment. We let the reward distribution for arm k in the corresponding bandit problem be a shifted Bernoulli distribution. In particular, the cost of arm k follows the distribution of $\mathbb{I}_{\{Y^k \neq Y^1\}} - C_k$ (we use costs here to avoid flipping signs).

The random costs of different arms are defined to be independent of each other. Let $\mathcal{P}_{\text{side}}$ denote the set of resulting bandit problems and let $f : \mathcal{P}_{\text{USS}} \to \mathcal{P}_{\text{side}}$ be the map that transforms USS instances to bandit instances by following the above transformation.

Now let $\pi \in \Pi(\mathcal{P}_{\text{side}})$ be a policy for $\mathcal{P}_{\text{side}}$, i.e., a map that describes what action to chosen given past information. Policy π can also be used on any (P, c) instance in $\mathcal{P}_{\text{USS}}$ in an obvious way: In particular, given the history of actions and states $A_1, U_1, \ldots, A_t, U_t$ in $\theta = (P, c)$ where $U_s = (Y_s, Y_s^1, \ldots, Y_s^K)$ such that the distribution of U_s given that $A_s = a$ is P marginalized to $\mathcal{Y}^a$, the next action to be taken is $A_{t+1} \sim \pi(\cdot|A_1, V_1, \ldots, A_t, V_t)$, where $V_s = (\mathbb{I}_{\{Y_s^1 \neq Y_s^1\}} - C_1, \ldots, \mathbb{I}_{\{Y_s^1 \neq Y_s^{A_s}\}} - C_{A_s})$. Let the resulting policy be denoted by π'. The following can be checked by simple direct calculation:

Proposition 7. *If $\theta \in \Theta_{\text{SD}}$, then the regret of π on $f(\theta) \in \mathcal{P}_{\text{side}}$ is the same as the regret of π' on θ.*

This implies that $\mathfrak{R}_T^*(\Theta_{\text{SD}}) \le \mathfrak{R}_T^*(f(\Theta_{\text{SD}}))$. Now note that this reasoning can also be repeated in the other "direction": For this, first note that the map f has a right inverse g (thus, $f \circ g$ is the identity over $\mathcal{P}_{\text{side}}$) and if π' is a policy for $\mathcal{P}_{\text{USS}}$, then π' can be "used" on any instance $\theta \in \mathcal{P}_{\text{side}}$ via the "inverse" of the above policy-transformation: Given the sequence $(A_1, V_1, \ldots, A_t, V_t)$ where $V_s = (B_s^1 + C_1, \ldots, B_s^K + C_s)$ is the vector of costs for round s with B_s^k being a Bernoulli with parameter γ_k, let $A_{t+1} \sim \pi'(\cdot|A_1, W_1, \ldots, A_t, W_t)$ where $W_s = (B_s^1, \ldots, B_s^{A_s})$. Denoting by π the resulting policy, we have the following proposition:

Proposition 8. *Let $\theta \in f(\Theta_{\text{SD}})$. Then the regret of policy π on $\theta \in f(\Theta_{\text{SD}})$ is the same as the regret of policy π' on instance $f^{-1}(\theta)$.*

Hence, $\mathfrak{R}_T^*(f(\Theta_{\text{SD}})) \le \mathfrak{R}_T^*(\Theta_{\text{SD}})$. In summary, we get the following result:

Theorem 9. $\mathfrak{R}_T^*(\Theta_{\text{SD}}) = \mathfrak{R}_T^*(f(\Theta_{\text{SD}}))$.

Lower Bounds: Note that as a consequence of the reduction and the one-to-one correspondence between the two problems, lower bounds for MAB with side-information is a lower bound for USS problem.

5 Algorithms

The reduction of the previous section suggests that one can utilize an algorithm developed for stochastic bandits with side-observation to learn on USS instances satisfying SD property. In this paper we make use of Algorithm 1 of [4] that was proposed for stochastic bandits with Gaussian side observations. As noted in the same paper, the algorithm is also suitable for problems where the payoff distributions are sub-Gaussian. As Bernoulli random variables are $\sigma^2 = 1/4$-sub-Gaussian (after centering), the algorithm is also applicable in our case.

Algorithm 1 Algorithm for USS under SD property

```
Play action $K$ and observe $Y^1, \ldots, Y^K$.
Set $\hat{\gamma}_i^1 \leftarrow \mathbb{I}_{\{Y^1 \neq Y^i\}}$ for all $i \in [K]$.
Initialize the exploration count: $n_e \leftarrow 0$.
Initialize the allocation counts: $N_K(1) \leftarrow 1$.
for $t = 2, 3, ...$ do
    if $\frac{N(t-1)}{4\alpha \log t} \in C(\hat{\gamma}^{t-1})$ then
        Set $I_t \leftarrow \text{argmin}_{k \in [K]} c(k, \hat{\gamma}^{t-1})$.
    else
        if $N_K(t-1) < \beta(n_e)/K$ then
            Set $I_t = K$.
        else
            Set $I_t$ to some $i$ for which
                $N_i(t-1) < u_i^*(\hat{\gamma}^{t-1}) 4\alpha \log t$.
        end if
        Increment exploration count: $n_e \leftarrow n_e + 1$.
    end if
    Play $I_t$ and observe $Y^1, \ldots, Y^{I_t}$.
    For $i \in [I_t]$, set
        $\hat{\gamma}_i^t \leftarrow (1 - \frac{1}{t})\hat{\gamma}_i^{t-1} + \frac{1}{t}\, \mathbb{I}_{\{Y^1 \neq Y^i\}}$.
end for
```

For the convenience of the reader, we give the algorithm resulting from applying the reduction to Algorithm 1 of [4] in an explicit form. To do this we need some extra notation. Recall that given a USS instance $\theta = (P, c)$, we let $\gamma_k = \mathbb{P}\left(Y \neq Y^k\right)$ where $(Y, Y^1, \ldots, Y^K) \sim P$ and $k \in [K]$. Let $k^* = \arg\min_k \gamma_k + C_k$ denote the optimal action and $\Delta_k(\theta) = \gamma_k + C_k - (\gamma_{k^*} + C_{k^*})$ the sub-optimality gap of arm k. Further, let $\Delta^*(\theta) = \min\{\Delta_k(\theta), k \neq k^*\}$ denote the smallest positive sub-optimality gap and define $\Delta_k^*(\theta) = \max\{\Delta_k(\theta), \Delta^*(\theta)\}$.

Since cost vector c is fixed, in the following we use parameter γ in place of θ to denote the problem instance. To explain the algorithm, we need to introduce some new concepts. A (fractional) allocation count $u \in [0, \infty)^K$ determines for each action i how many times the action is selected. Thanks to the cascade structure, using an action i implies observing the output of all the sensors with index j less than equal to i. Hence, a sensor j gets observed $u_j + u_{j+1} + \cdots + u_K$ times. We call an allocation count "sufficiently informative" if (with some level of confidence) it holds that *(i)* for each suboptimal choice, the number of observations for the corresponding sensor is sufficiently large to distinguish it from the optimal choice; and *(ii)* the optimal choice is also distinguishable from the second best choice. We collect these counts into the set $C(\gamma)$ for a given parameter γ: $C(\gamma) = \{u \in [0, \infty)^K : u_j + u_{j+1} + \cdots + u_K \ge \frac{2\sigma^2}{(\Delta_j^*(\theta))^2}, j \in [K]\}$ (note that

$\sigma^2 = 1/4$). Further, let $u^*(\gamma)$ be the allocation count that minimizes the total expected excess cost over the set of sufficiently informative allocation counts: In particular, we let $u^*(\gamma) = \operatorname{argmin}_{u \in C(\gamma)} \langle u, \Delta(\theta)\rangle$ with the understanding that for any optimal action k, $u^*_k(\gamma) = \min\{u_k : u \in C(\gamma)\}$. For an allocation count $u \in [0,\infty)^K$ let $m(u) \in \mathbb{N}^K$ denote total sensor observations, where $m_j(u) = \sum_{i=1}^{j} u_i$ corresponds to observations of sensor j.

The idea of our algorithm, shown as Algorithm 1, is as follows: The algorithm keeps track of an estimate $\hat{\gamma}^t \doteq (\hat{\gamma}^t_i)_{i\in[K]}$ of γ in each round, which is initialized by pulling arm K as this arm gives information about all the other arms. In each round, the algorithm first checks whether given the current estimate $\hat{\gamma}^t$ and the current confidence level (where the confidence level is gradually increased over time), the allocation count $N(t) \in \mathbb{N}^K$ is sufficiently informative (cf. line 6). If this holds, the action that is optimal under $\hat{\gamma}(t)$ is chosen (cf. line 7). If the check fails, we need to explore. The idea of the exploration is that it tries to ensure that the "optimal plan" – assuming $\hat{\gamma}$ is the "correct" parameter – is followed (line 12). However, this is only reasonable, if all components of γ are relatively well-estimated. Thus, first the algorithm checks whether any of the components of γ has a chance of being particularly poorly estimated (line 9). Note that the requirement here is that a significant, but still altogether diminishing fraction of the *exploration rounds* is spent on estimating each components: In the long run, the fraction of exploration rounds amongst all rounds itself is diminishing; hence the forced exploration of line 10 overall has a small impact on the regret, while it allows to stabilize the algorithm.

For $\theta \in \Theta_{\mathrm{SD}}$, let $\gamma(\theta)$ be the error probabilities for the various sensors. The following result follows from Theorem 6 of [4]:

Theorem 10. *Let $\epsilon > 0$, $\alpha > 2$ arbitrary and choose any non-decreasing $\beta(n)$ that satisfies $0 \le \beta(n) \le n/2$ and $\beta(m+n) \le \beta(m) + \beta(n)$ for $m,n \in \mathbb{N}$. Then, for any $\theta \in \Theta_{\mathrm{SD}}$, letting $\gamma = \gamma(\theta)$ the expected regret of Algorithm 1 after T steps satisfies*

$$\begin{aligned}\mathfrak{R}_T(\theta,c) \le{} & \left(2K + 2 + \frac{4K}{\alpha-2}\right) + 4K\sum_{s=0}^{T}\exp\left(\frac{-8\beta(s)\epsilon^2}{2K}\right)\\ & + 2\beta\left(4\alpha\log T\sum_{i\in[K]} u^*_i(\gamma,\epsilon) + K\right)\\ & + 4\alpha\log T\sum_{i\in[K]} u^*_i(\gamma,\epsilon)\Delta_i(\theta)\,,\end{aligned}$$

where $u^*_i(\gamma,\epsilon) = \sup\{u^*_i(\gamma') : \|\gamma'-\gamma\|_\infty \le \epsilon\}$.

Further specifying $\beta(n)$ and using the continuity of $u^*(\cdot)$ at θ, it immediately follows that Algorithm 1 achieves asymptotically optimal performance:

Algorithm 2 Algorithm for USS with WD property

```
1:  Play action $K$ and observe $Y^1,\ldots,Y^K$
2:  Set $\hat{\gamma}^1_{ij} \leftarrow \mathbb{I}_{\{Y^i \neq Y^j\}}$ for all $i,j \in [K]$ and $i<j$.
3:  $n_i(1) \leftarrow \mathbb{I}_{\{i=K\}}$ $\forall i \in [K]$.
4:  for $t = 2,3,\ldots$ do
5:      $U^t_{ij} = \hat{\gamma}^{t-1}_{ij} + \sqrt{\frac{1.5\log(t)}{n_j(t-1)}}$ $\forall\, i,j \in [K]$ and $i<j$
6:      $S_t = \{i \in [K-1] : C_j - C_i \geq U^t_{ij}\ \forall\, j > i\}$
7:      Set $I_t = \arg\min S_t \cup \{K\}$
8:      Play $I_t$ and observe $Y^1,\ldots,Y^{I_t}$.
9:      for $i \in [I_t]$ do
10:         $n_i(t) \leftarrow n_i(t-1) + 1$
11:         $\hat{\gamma}^t_{ij} \leftarrow \left(1 - \frac{1}{n_j(t)}\right)\hat{\gamma}^{t-1}_{ij} + \frac{1}{n_j(t)}\mathbb{I}_{\{Y^j \neq Y^i\}} \forall\ i < j \leq I_t$
12:     end for
13: end for
```

Corollary 11. *Suppose the conditions of Theorem 10 hold. Assume, furthermore, that $\beta(n)$ satisfies $\beta(n) = o(n)$ and $\sum_{s=0}^{\infty}\exp\left(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\right) < \infty$ for any $\epsilon > 0$, then for any θ such that $u^*(\theta)$ is unique, $\limsup_{T\to\infty}\mathfrak{R}_T(\theta,c)/\log T \le 4\alpha\inf_{u\in C_\theta}\langle u,\Delta(\theta)\rangle$, i.e., by the lower bound of [4] the algorithm is asymptotically optimal.*

Note that any $\beta(n) = an^b$ with $a \in (0,\frac{1}{2}]$, $b \in (0,1)$ satisfies the requirements in Theorem 10 and Corollary 11.

Algorithm 1 only estimates the disagreements $\mathcal{P}\{Y^1 \neq Y^j\}$ for all $j \in [K]$ which suffices to identify the optimal arm when SD property (see Section 4) holds. Clearly, one can estimate pairwise disagreements probabilities $\mathcal{P}\{Y^i \neq Y^j\}$ for $i \neq j$ and use them to order the arms. We next develop a heuristic algorithm that uses this information and works for USS under WD.

5.1 Algorithm for Weak Dominance

The reduction scheme described above is optimal for SD instances and can fail under the more relaxed WD property. This is because, while under SD, the marginal error is equal to marginal disagreement (see Prop. 2) implying that for any two sensors $i < j$, $\gamma_i - \gamma_j = (\gamma_i - \gamma_1) - (\gamma_j - \gamma_1) = \mathbb{P}\left(Y^i \neq Y^1\right) - \mathbb{P}\left(Y^i \neq Y^1\right)$ and that the marginal error between any two sensors can be computed by only keeping track of disagreements relative to sensor 1, under WD, the marginal error is a lower bound and keeping track only of disagreements relative to sensor 1 no longer suffices because $\mathbb{P}\left(Y^i \neq Y^1\right) - \mathbb{P}\left(Y^i \neq Y^1\right)$ is no longer a good estimate for $\mathbb{P}\left(Y^i \neq Y^j\right)$.

Our key insight is based on the fact that, under WD, for a given set of the disagreement probabilities, for all $i \neq j$, the set $\{i \in [K-1] : C_j - C_i \geq \mathcal{P}\{Y^i \neq Y^j\}$ for all $j > i\}$ includes the optimal arm. We use this idea in Algorithm 2 to identify the optimal arm when an instance of USS satisfies WD. We will experimentally validate Algorithm 2's performance on real datasets in the next section.

Algorithm 2 works as follows. In each round, t, based on history, we keep track of estimates, $\hat{\gamma}_{ij}^t$, of disagreements between sensors $i \neq j$. In the first round, the algorithm plays arm K and initializes its values. In each subsequent round, the algorithm computes the upper confidence value of $\hat{\gamma}_{ij}^t$ denoted as U_{ij}^t (5) for all pairs (i, j) and orders the arms: i is considered better than arm j if $C_j - C_i \geq U_{ij}^t$. Specifically, the algorithm plays an arm i that satisfies $C_j - C_i \geq U_{ij}^t$ for all $j > i$ (cf. line 6). If no such arm is found, then it plays arm K. Regret guarantees analogous to Theorem 10 under SD for this new scheme can also be derived but only under additional conditions. In particular, no similar guarantee can be provided under WD *alone* because the set of WD instances is not uniformly learnable (Theorem 19 of Appendix B).

5.2 Extensions

We describe briefly a few extensions here, which we will elaborate on in a longer version of the paper.

Tree Structures: The ideas presented can be extended to the case when sensors are organized as a tree with root node corresponding to sensor 1 and the decision maker can select any path starting from the root node. To see this, note that the marginal error condition of Eq. (1) still characterizes optimal sensor selection. Under a modified variant of SD, namely, when Eq. (2) applies to all children of any sensor, it follows that it is again sufficient to keep track of disagreements. In an analogous fashion Eq. (5) can be suitably modified as well. This leads to a sublinear regret algorithm for tree-structures.

Context-Dependent Sensor Selection: Before a decision is made about which sensor to pick, a context $X \in \mathcal{X} \subset \mathbb{R}^d$ is observed. It is assumed that the hidden state and the sensor measurements depend on the context in a stochastic fashion. Analogous to the context-independent case, we impose context dependent notions for SD and WD, namely, Eq. (2) and Eq. (5) hold conditioned on each $x \in \mathcal{X}$. To handle these cases we let $\gamma_i(x) \doteq \Pr(Y^i \neq Y \mid X = x)$ and $\gamma_{ij}(x) \doteq \Pr(Y^i \neq Y^j \mid X = x)$ denote the corresponding contextual error and disagreement probabilities. Our sublinear regret guarantees can be generalized for a parameterized GLM model for disagreement, namely, when the log-odds ratio $\log \frac{\gamma_{ij}(x)}{1-\gamma_{ij}(x)}$ can be written as $\theta_{ij}' x$ for some unknown $\theta_{ij} \in \mathbb{R}^d$.

6 Experiments

In this section we evaluate performance of Algorithms 1 and 2 on a completely synthetic example and two "real" examples derived from real datasets, PIMA-Diabetes and Heart Disease (Cleveland). Both of these datasets specify the costs for acquiring individual features.

Synthetic: We generate data as follows. The input, Y_t, is generated IID Ber(0.7). Outputs for sensors 1, 2, 3 have an overall error statistic, $\gamma_1 = 0.4$, $\gamma_2 = 0.1$, $\gamma_3 = 0.05$ respectively. To ensure SD we enforce Eq. (2) during the generation process. To relax SD, we introduce errors up to 10% during data generation for sensor outputs 2 and 3 when sensor 1 predicts correctly.

Real Datasets: We split the features into three "sensors" based on their costs. For PIMA-Diabetes dataset (size=768) the first sensor is associated with patient history/profile at a cost of \$6, the 2nd sensor in addition utilizes insulin test (cost \$ 22) and the 3rd sensor uses all attributes (cost \$46). For the Heart dataset (size=297) we use the first 7 attributes that includes cholesterol readings, blood-sugar, and rest-ECG (cost \$27), the 2nd sensor utilizes in addition the thalach, exang and oldpeak attributes that cost \$300 and the 3rd sensor utilizes more extensive tests at a total cost of \$601.

We train three linear SVM classifiers with 5-fold cross-validation and have verified that our results match known state-of-art. Note that Table 1 shows that the resulting classifiers/tests on these datasets approximately satisfies SD condition and thus our setup should apply to these datasets. The size of these datasets is relatively small and limits our ability to experiment. To run the online algorithm we therefore generate an instance randomly from the dataset (with replacement) in each round. We repeat the experiments 20 times and averages are shown with 95% confidence bounds.

Testing Learnability: We experiment with different USS algorithms on the synthetic dataset. Our purpose is twofold: *(a)* verify Algorithm 1 under SD; *(b)* verify that WD condition gives a maximal learnable set. Fig. 3 depicts the results of Algorithm 1 when SD condition is satisfied and shows that we obtain sublinear regret regardless of costs/probabilities. To test WD requirement we parameterize the problem by varying costs. Without loss of generality we fix the cost of sensor 1 to be zero and the total cost of the entire system to be C_{tot}. We then vary the costs of sensors 2 and 3. We test the hypothesis that WD is a maximal learnable set. We enforce Sensor 2 as the optimal sensor and vary the costs so that we continu-

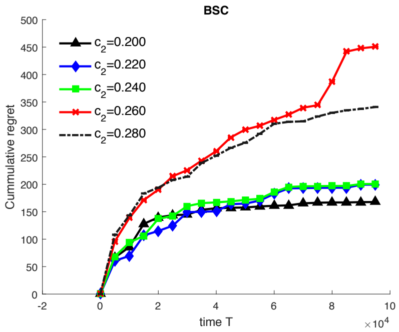

Figure 3: Regret under SD property

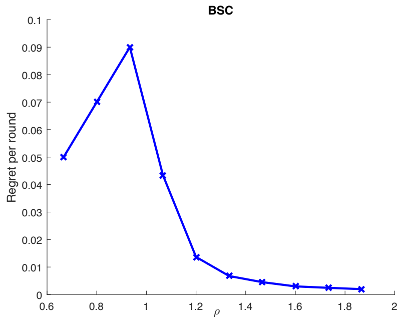

Figure 4: Regret under as ρ from Eq. (5) varies.

Fig. 3 depicts regret of Algorithm 1 on synthetic data under SD. Under SD the regret is always sublinear regardless of costs/probability. Fig. 4 demonstrates a phase-transition effect. (The regret of Algorithm 1 is not plotted here because this algorithm fails in this case.) As $\rho \to 1$ (from the right) regret-per-round drastically increases, which is an indirect indication of that WD is a maximal learnable set.

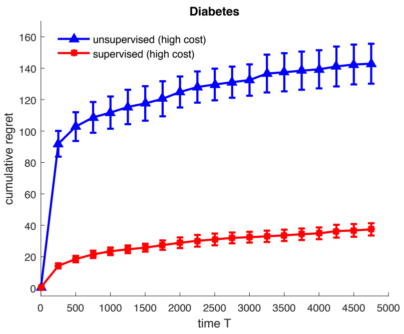

Figure 5: Regret Curves on PIMA Diabetes

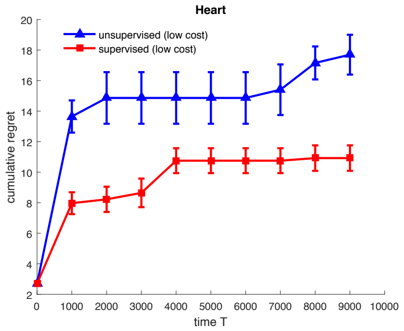

Figure 6: Regret Curves on Heart Disease

Figs. 5 and 6 depict performance for real-datasets and presents comparisons against supervised (bandit with ground-truth feedback) scenario.

ously pass from the situation where WD holds ($\rho \geq 1$) to the case where WD does not hold ($\rho < 1$). Fig. 4 depicts regret-per-round for Algorithm 2 and as we can verify there is indeed a transition at $\rho = 1$.

Unsupervised vs. Supervised Learning: The real datasets provide an opportunity to understand how different types of information impact performance. We compare Algorithm 2 against a corresponding bandit algorithm where the learner receives feedback. In particular, for each action in each round, in the bandit setting, the learner knows whether or not the corresponding sensor output is correct. We implement the "supervised bandit" setting by replacing Step 5 in Algorithm 2 with estimated marginal error rates.

We scale costs by means of a tuning parameter (since the costs of features are all greater than one) and consider minimizing a combined objective (λ Cost + Error) as stated in Section 2. High (low)-values for λ correspond to low (high)-budget constraint. If we set a fixed budget of \$50, this corresponds to high-budget (small λ) and low budget (large λ) for PIMA Diabetes (3rd test optimal) and Heart Disease (1st test optimal) respectively. Figs. 5 and 6 depict performance. We notice that for both high as well as low cost scenarios, while supervised does have lower regret, the USS cummulative regret is also sublinear and within a constant fraction of the supervised case. This is qualititively interesting because these plots demonstrate that (although under WD we do not have uniform learnability), in typical cases, we learn as well as the supervised setting.

7 Conclusions

The paper describes a novel approach for unsupervised sensor selection, which arises, e.g., in a number of healthcare and security applications. The main challenge in these applications is that ground-truth annotated examples are unavailable and it is often difficult to acquire them in-situ. We proposed a novel approach for sensor selection based on novel notions of weak- and strong-dominance. We showed that weak dominance property is maximal in that violation of this condition leads to loss of learnability. Our experiments demonstrate that weak dominance does hold in practice for real datasets and we also found that for these datasets sensor selection under no supervision can be as effective as under supervision.

Acknowledgments

M. K. Hanawal would like to thank support from INSPIRE faculty award (IFA-14/ENG-73, 16DSTINS002) from Department of Science and Technology, Government of India and SEED grant (16IRCCSG010) from IIT Bombay. Cs. Szepesvári greatly acknowledges the support of NSERC and the Alberta Innovates Technology Futures through the Alberta Machine Intelligence Institute (AMII).

References

[1] K. Trapeznikov, V. Saligrama, and D. A. Castanon, "Multi-stage classifier design," *Machine Learning*, vol. 39, pp. 1–24, 2014.

[2] A. H. Baghdanian, A. A. Baghdanian, A. Armetta, M. Krastev, T. Dechert, P. Burke, C. A. LeBedis, S. W. Anderson, and J. A. Soto, "Effect of an institutional triaging algorithm on the use of multidetector ct for patients with blunt abdominopelvic trauma over an 8-year period," *Radiology*, 2016.

[3] G. Bartók, D. Foster, D. Pál, A. Rakhlin, and C. Szepesvári, "Partial monitoring – classification, regret bounds, and algorithms," *Mathematics of Operations Research*, vol. 39, pp. 967–997, 2014.

[4] Y. Wu, A. György, and C. Szepesvári, "Online learning with gaussian payoffs and side observations," in *NIPS*, September 2015, pp. 1360–1368.

[5] K. Trapeznikov and V. Saligrama, "Supervised sequential classification under budget constraints," in *AISTATS*, 2013, pp. 235–242.

[6] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori, "Prediction with limited advice and multiarmed bandits with paid observations," in *Proceeding of International Conference on Machine Learning, ICML*, 2014, pp. 208–287.

[7] N. Zolghadr, G. Bartók, R. Greiner, A. György, and C. Szepesvári, "Online learning with costly features and labels," in *NIPS*, 2013, pp. 1241–1249.

[8] B. Póczos, Y. Abbasi-Yadkori, C. Szepesvári, R. Greiner, and N. Sturtevant, "Learning when to stop thinking and do something!" in *ICML*, 2009, pp. 825–832.

[9] R. Greiner, A. Grove, and D. Roth, "Learning cost-sensitive active classifiers," *Artificial Intelligence*, vol. 139, pp. 137–174, 2002.

[10] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space," *IEEE Transaction on Automatic Control*, vol. 34, pp. 258–267, 1989.

[11] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *NIPS*, 2011.

[12] N. Alon, N. Cesa-Biancbi, O. Dekel, and T. Koren, "Online learning with feedback graphs:beyond bandits," in *Proceeding of Conference on Learning Theory*, 2015, pp. 23–35.

[13] N. Alon, N. Cesa-Biancbi, C. Gentile, and Y. Mansour, "From bandits to experts: A tale of domination and independence," in *Proceeding of Conference on Neural Information Processing Systems, NIPS*, 2013, pp. 1610–1618.

[14] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications.* Prentice Hall., 1983.

[15] R. Ludwig and J. Taylor, "Voyager telecommunications manual, jpl descanso (design and performance summary series)," March 2002.

A Stochastic Partial Monitoring Problem

In Section 2 it was mentioned that our problem is a special case stochastic partial monitoring (SPM). The purpose of this short section is to formally define SPM problems. In an SPM a learner interacts with a stochastic environment in a sequential manner. In round $t = 1, 2, \dots$ the learner chooses an action A_t from an action set $\mathcal{A}$, and receives a feedback $Y_t \in \mathcal{Y}$ from a distribution p which depends on the action chosen and also on the environment instance identified with a "parameter" $\theta \in \Theta$: $Y_t \sim p(\cdot; A_t, \theta)$. The learner also incurs a reward R_t, which is a function of the action chosen and the unknown parameter θ: $R_t = r(A_t, \theta)$. The reward may or may not be part of the feedback for round t. The learner's goal is to maximize its total expected reward. The family of distributions $(p(\cdot; a, \theta))_{a,\theta}$ and the family of rewards $(r(a, \theta))_{a,\theta}$ and the set of possible parameters Θ are known to the learner, who uses this knowledge to judiciously choose its next action to reduce its uncertainty about θ so that it is able to eventually converge on choosing only an optimal action $a^*(\theta)$, achieving the best possible reward per round, $r^*(\theta) = \max_{a \in \mathcal{A}} r(a, \theta)$. Bandit problems are a special case of SPMs where $\mathcal{Y}$ is the set of real numbers, $r(a, \theta)$ is the mean of distribution $p(\cdot; a, \theta)$.

B Proofs for Section 3

Here we provide the missing proof for Section 3. For the convenience of the reader the statements of the various propositions are repeated. We start with proofs related to Strong Dominance.

B.1 Proofs Related to Strong Dominance

Proposition 1. *A subset $\Theta \subset \Theta_{\mathrm{SA}}$ is learnable if and only if there exists a map $a : M_1(\{0,1\}^K) \to [K]$ such that for any $\theta = (P, c) \in \Theta$ with decomposition $P = P_S \otimes P_{Y|S}$, $a(P_S)$ is an optimal action in θ. Following our previous convention, we also write $\theta = P_S \otimes P_{Y|S}$.*

Proof. $\Rightarrow$: Let $\mathfrak{A}$ be an algorithm that achieves sublinear regret and pick an instance $\theta \in \Theta$. Let $P = P_S \otimes P_{Y|S}$. The regret $\mathfrak{R}_n(\mathfrak{A}, \theta)$ of $\mathfrak{A}$ on instance θ can be written in the form

$$\mathfrak{R}_n(\mathfrak{A}, \theta) = \sum_{k \in [K]} \mathbb{E}_{P_S}\left[N_k(n)\right] \Delta_k(\theta)\,,$$

where $N_k(n)$ is the number of times action k is chosen by $\mathfrak{A}$ during the n rounds while $\mathfrak{A}$ interacts with θ, $\Delta_k(\theta) = c(k, \theta) - c^*(\theta)$ is the immediate regret and $\mathbb{E}_{P_S}[\cdot]$ denotes the expectation under the distribution induced by P_S. In particular, $N_k(n)$ hides dependence on the iid sequence $Y_1, \dots, Y_n \sim P_S$ that we are taking the expectation over here. Since the regret is sublinear, for any k suboptimal action, $\mathbb{E}_{P_S}[N_k(n)] = o(n)$. Define $a(P_S) = \min\{k \in [K]\,;\, \mathbb{E}_{P_S}[N_k(n)] = \Omega(n)\,\}$. Then, a is well-defined as the distribution of $N_k(n)$ for any k depends only on P_S (and c). Furthermore, $a(P_S)$ selects an optimal action.

$\Leftarrow$: Let a be the map in the statement and let $f : \mathbb{N}_+ \to \mathbb{N}_+$ be such that $1 \le f(n) \le n$ for any $n \in \mathbb{N}$, $f(n)/\log(n) \to \infty$ as $n \to \infty$ and $f(n)/n \to 0$ as $n \to \infty$ (say, $f(n) = \lceil \sqrt{n} \rceil$). Consider the algorithm that chooses $I_t = K$ for the first $f(n)$ steps, after which it estimates $\hat{P}_S$ by frequency counting and then uses $I_t = a(\hat{P}_S)$ in the remaining $n - f(n)$ trials. Pick any $\theta \in \Theta$ so that $\theta = P_S \otimes P_{Y|S}$. Note that by Hoeffding's inequality, $\sup_{y \in \{0,1\}^K} |\hat{P}_S(y) - P_S(y)| \le \sqrt{\frac{K \log(4n)}{2f(n)}}$ holds with probability $1 - 1/n$. Let n_0 be the first index such that for any $n \ge n_0$, $\sqrt{\frac{K \log(4n)}{2f(n)}} \le \Delta^*(\theta) \doteq \min_{k : \Delta_k(\theta) > 0} \Delta_k(\theta)$. Such an index n_0 exists by our assumptions that f grows faster than $n \mapsto \log(n)$. For $n \ge n_0$, the expected regret of $\mathfrak{A}$ is at most $n \times 1/n + f(n)(1 - 1/n) \le 1 + f(n) = o(n)$. □

Next, we present the proof of Proposition 3. We will need this proposition in the proof of Theorem 2.

Proposition 3. *Let $\gamma_i = \gamma_i(\theta)$ for $\theta \in \Theta_{\mathrm{SA}}$, Then, for any $i, j \in [K]$, $\gamma_i - \gamma_j = \mathbb{P}\left(Y^i \neq Y^j\right) - 2\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$.*

Proof. Recall that $\gamma_i = \mathbb{P}\left(Y^i \neq Y\right)$, $(Y, Y^1, \ldots, Y^K) \sim \theta$. We have

$$\begin{aligned}
\gamma_i - \gamma_j &= \mathbb{P}\left(Y^i \neq Y\right) - \mathbb{P}\left(Y^j \neq Y\right) \\
&= \cancel{\mathbb{P}\left(Y^i \neq Y, Y^i = Y^j\right)} + \mathbb{P}\left(Y^i \neq Y, Y^i \neq Y^j\right) - \left\{\cancel{\mathbb{P}\left(Y^j \neq Y, Y^i = Y^j\right)} + \mathbb{P}\left(Y^j \neq Y, Y^i \neq Y^j\right)\right\} \\
&= \mathbb{P}\left(Y^i \neq Y, Y^i \neq Y^j\right) + \mathbb{P}\left(Y^i = Y, Y^i \neq Y^j\right) - \left\{\mathbb{P}\left(Y^j \neq Y, Y^i \neq Y^j\right) + \mathbb{P}\left(Y^i = Y, Y^i \neq Y^j\right)\right\} \\
&\overset{(a)}{=} \mathbb{P}\left(Y^j \neq Y^i\right) - 2\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right) ,
\end{aligned}$$

where in (a) we used that $\mathbb{P}\left(Y^j \neq Y, Y^i \neq Y^j\right) = \mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$ and also $\mathbb{P}\left(Y^i = Y, Y^i \neq Y^j\right) = \mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$ which hold because Y, Y^i, Y^j only take on two possible values. □

With this, we are ready to give the proof of Theorem 2:

Theorem 2. *The set Θ_{SD} is learnable.*

Proof of Theorem 2. We construct a map as required by Proposition 1. Take an instance $\theta \in \Theta_{\mathrm{SD}}$ and let $\theta = P_S \otimes P_{Y|S}$ be its decomposition as before. Let $\gamma_i = \mathbb{P}\left(Y^i \neq Y\right)$, $(Y, Y^1, \ldots, Y^K) \sim \theta$, $C_i = c_1 + \cdots + c_i$. For identifying an optimal action in θ, it clearly suffices to know the sign of $\gamma_i + C_i - (\gamma_j + C_j) = \gamma_i - \gamma_j + (C_i - C_j)$ for all pairs $i, j \in [K]^2$. Without loss of generality (WLOG) let $i < j$. By Proposition 3, $\gamma_i - \gamma_j = \mathbb{P}\left(Y^i \neq Y^j\right) - 2\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$. Now, since θ satisfies the strong dominance condition, $\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right) = 0$. Thus, $\gamma_i - \gamma_j = \mathbb{P}\left(Y^i \neq Y^j\right)$ which is a function of P_S only. Since $(C_i)_i$ are known, a map as required by Proposition 1 exists. □

Let us now turn to proofs and statements related to Weak Dominance.

B.2 Proofs and Statements Related to Weak Dominance

We start with two statements that prepare for the definition of weak dominance.

Proposition 5. *Let $i < j$. Assume*

$$C_j - C_i \notin \left[\gamma_i - \gamma_j, \mathbb{P}\left(Y^i \neq Y^j\right)\right). \tag{3}$$

Then $\gamma_i + C_i \leq \gamma_j + C_j$ if and only if $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right)$.

Proof. ⇒: From the premise, it follows that $C_j - C_i \geq \gamma_i - \gamma_j$. Thus, by (3), $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. ⇐: We have $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right) \geq \gamma_i - \gamma_j$, where the last inequality is by Corollary 4. □

Proposition 6. *Let $j < i$. Assume*

$$C_i - C_j \notin \left(\gamma_j - \gamma_i, \mathbb{P}\left(Y^i \neq Y^j\right)\right]. \tag{4}$$

Then, $\gamma_i + C_i \leq \gamma_j + C_j$ if and only if $C_i - C_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$.

Proof. ⇒: The condition $\gamma_i + C_i \leq \gamma_j + C_j$ implies that $\gamma_j - \gamma_i \geq C_i - C_j$. By Corollary 4 we get $\mathbb{P}\left(Y^i \neq Y^j\right) \geq C_i - C_j$. ⇐: Let $C_i - C_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$. Then, by (4), $C_i - C_j \leq \gamma_j - \gamma_i$. □

Recall the definition of a_{wd}:

Definition 3. *For $P_S \in M_1(\{0,1\}^K)$ let $a_{\mathrm{wd}}(P_S)$ denote the smallest index $i \in [K]$ such that*

$$\forall j < i \ : \ C_i - C_j < \mathbb{P}\left(Y^i \neq Y^j\right) , \tag{7a}$$

$$\forall j > i \ : \ C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right) , \tag{7b}$$

where $C_i = c_1 + \cdots + c_i$, $i \in [K]$ and $(Y^1, \ldots, Y^K) \sim P_S$. (If no such index exists, a_{wd} is undefined, i.e., a_{wd} is a partial function.)

This definition makes sense:

Proposition 12. *The action-selector a_{wd} is sound over Θ_{WD}: For any $\theta \in \Theta_{\mathrm{WD}}$ with $\theta = P_S \otimes P_{Y|S}$, $a_{\mathrm{wd}}(P_S)$ is well-defined, i.e., the domain of a_{wd} includes all of Θ_{WD}.*

Proof. Let $\theta \in \Theta_{\mathrm{WD}}$, $i = a^*(\theta)$. It suffices to show that i satisfies both (7a) and (7b). Obviously, (7b) holds by the definition of Θ_{WD}. Thus, the only question is whether (7a) also holds. We prove this by contradiction: In particular if (7a) does not hold then for some $j < i$, $C_i - C_j \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. Then, by Corollary 4, $\mathbb{P}\left(Y^i \neq Y^j\right) \geq \gamma_j - \gamma_i$, hence $\gamma_j + C_j \leq \gamma_i + C_i$, which contradicts the definition of i, thus finishing the proof. □

We can now prove that a_{wd} is sound:

Proposition 13. *The map a_{wd} is sound over Θ_{WD}: In particular, for any $\theta \in \Theta_{\mathrm{WD}}$ with $\theta = P_S \otimes P_{Y|S}$, $a_{\mathrm{wd}}(P_S) = a^*(\theta)$.*

Proof. Take any $\theta \in \Theta_{\mathrm{WD}}$ with $\theta = P_S \otimes P_{Y|S}$, $i = a_{\mathrm{wd}}(P_S)$, $j = a^*(\theta)$. If $i = j$, there is nothing to be proven. Hence, first assume that $j > i$. Then, by (7b), $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. By Corollary 4, $\mathbb{P}\left(Y^i \neq Y^j\right) \geq \gamma_i - \gamma_j$. Combining these two inequalities we get that $\gamma_i + C_i \leq \gamma_j + C_j$, which contradicts with the definition of j. Now, assume that $j < i$. Then, by (5), $C_i - C_j \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. However, by (7a), $C_i - C_j < \mathbb{P}\left(Y^i \neq Y^j\right)$, thus $j < i$ cannot hold either and we must have $i = j$. □

A corollary of the previous result is that Θ_{WD} is also learnable:

Theorem 14. *The set Θ_{WD} is learnable.*

Proof. By Proposition 12, a_{wd} is well-defined over Θ_{WD}, while by Proposition 13, a_{wd} is sound over Θ_{WD}. □

Now, we will prove that a_{wd} is the only sound action selector over Θ_{WD}.

Proposition 15. *Let $\theta \in \Theta_{\mathrm{SA}}$ and $\theta = P_S \otimes P_{Y|S}$ be such that a_{wd} is defined for P_S and $a_{\mathrm{wd}}(P_S) = a^*(\theta)$. Then $\theta \in \Theta_{\mathrm{WD}}$.*

The proof follows from the definitions. An immediate corollary of the previous proposition is as follows:

Corollary 16. *Let $\theta \in \Theta_{\mathrm{SA}}$ and $\theta = P_S \otimes P_{Y|S}$. Assume that a_{wd} is defined for P_S and $\theta \notin \Theta_{\mathrm{WD}}$. Then $a_{\mathrm{wd}}(P_S) \neq a^*(\theta)$.*

The next result states that a_{wd} is essentially the only sound action selector map defined for all instances derived from instances of Θ_{WD}:

Theorem 17. *Take any action selector map $a : M_1(\{0,1\}^K) \to [K]$ which is sound over Θ_{WD}. Then, for any P_S such that $\theta = P_S \otimes P_{Y|S} \in \Theta_{\mathrm{WD}}$ with some $P_{Y|S}$, $a(P_S) = a_{\mathrm{wd}}(P_S)$.*

Proof. Pick any $\theta = P_S \otimes P_{Y|S} \in \Theta_{\mathrm{WD}}$. If $A^*(\theta)$ is a singleton, then clearly $a(P_S) = a_{\mathrm{wd}}(P_S)$ since both are sound over Θ_{WD}. Hence, assume that $A^*(\theta)$ is not a singleton. Let $i = a^*(\theta) = \min A^*(\theta)$ and let $j = \min A^*(\theta) \setminus \{i\}$. We argue that $P_{Y|S}$ can be changed so that on the new instance i is still an optimal action, while j is not an optimal action, while the new instance $\theta' = P_S \otimes P'_{Y|S}$ is in Θ_{WD}.

The modification is as follows: Consider any $y^{-j} \doteq (y^1, \ldots, y^{j-1}, y^{j+1}, \ldots, y^K) \in \{0,1\}^{K-1}$. For $y, y^j \in \{0,1\}$, define $q(y|y^j) = P_{Y|S}(y|y^1, \ldots, y^{j-1}, y^j, y^{j+1}, \ldots, y^K)$ and similarly let $q'(y|y^j) = P'_{Y|S}(y|y^1, \ldots, y^{j-1}, y^j, y^{j+1}, \ldots, y^K)$ Then, we let $q'(0|0) = 0$ and $q'(0|1) = q(0|0)+q(0|1)$, while we let $q'(1|1) = 0$ and $q'(1|0) = q(1|1)+q(1|0)$. This makes $P'_{Y|S}$ well-defined ($P'_{Y|S}(\cdot|y^1, \ldots, y^K)$ is a distribution for any $y^1, \ldots, y^K$). Further, we claim that the transformation has the property that it leaves γ_p unchanged for $p \neq j$, while γ_j is guaranteed to decrease. To see why γ_p is left unchanged for $p \neq j$ note that $\gamma_p = \sum_{y^p} P_{Y^p}(y^p) P_{Y|Y^p}(1-y^p|y^p)$. Clearly, P_{Y^p} is left unchanged. Introducing y^{-k} to denote a tuple where the kth component is left out, $P_{Y|Y^p}(1-y^p|y^p) = \sum_{y^{-p,-j}} P_{Y|Y^1,\ldots,Y^K}(1-y^p|y^1, \ldots, y^{j-1}, 0, y^{j+1}, \ldots, y^K) + P_{Y|Y^1,\ldots,Y^K}(1-y^p|y^1, \ldots, y^{j-1}, 1, y^{j+1}, \ldots, y^K)$ and by definition,

$$\begin{aligned}&P_{Y|Y^1,\dots,Y^K}(1-y^p|y^1,\dots,y^{j-1},0,y^{j+1},\dots,y^K)\\&\quad+P_{Y|Y^1,\dots,Y^K}(1-y^p|y^1,\dots,y^{j-1},1,y^{j+1},\dots,y^K)\\&=P'_{Y|Y^1,\dots,Y^K}(1-y^p|y^1,\dots,y^{j-1},0,y^{j+1},\dots,y^K)\\&\quad+P'_{Y|Y^1,\dots,Y^K}(1-y^p|y^1,\dots,y^{j-1},1,y^{j+1},\dots,y^K)\,,\end{aligned}$$

where the equality holds because "$q'(y|0)+q'(y|1)=q(y|0)+q(y|1)$". Thus, $P_{Y|Y^p}(1-y^p|y^p)=P'_{Y|Y^p}(1-y^p|y^p)$ as claimed. That γ_j is non-increasing follows with an analogue calculation. In fact, this shows that γ_j is strictly decreased if for any $(y^1,\dots,y^{j-1},y^{j+1},\dots,y^K)\in\{0,1\}^{K-1}$, either $q(0|0)$ or $q(1|1)$ was positive. If these are never positive, this means that $\gamma_j=1$. But then j cannot be optimal since $c_j>0$. Since j was optimal, γ_j is guaranteed to decrease.

Finally, it is clear that the new instance is still in Θ_{WD} since $a^*(\theta)$ is left unchanged. □

The next result shows that the set Θ_{WD} is essentially a maximal learnable set in $\mathrm{dom}(a_{\mathrm{wd}})$:

Theorem 18. *Let $a: M_1(\{0,1\}^K)\to[K]$ be an action selector map such that a is sound over the instances of Θ_{WD}. Then there is no instance $\theta=P_S\otimes P_{Y|S}\in\Theta_{\mathrm{SA}}\setminus\Theta_{\mathrm{WD}}$ such that $P_S\in\mathrm{dom}(a_{\mathrm{wd}})$, the optimal action of θ is unique and $a(P_S)=a^*(\theta)$.*

Proof. Let a as in the theorem statement. By Theorem 17, a_{wd} is the unique sound action-selector map over Θ_{WD}. Thus, for any $\theta=P_S\otimes P_{Y|S}\in\Theta_{\mathrm{WD}}$, $a_{\mathrm{wd}}(P_S)=a(P_S)$. Hence, the result follows from Corollary 16. □

Note that $\mathrm{dom}(a_{\mathrm{wd}})\setminus\{P_S:\exists P_{Y|S}\text{ s.t. }P_S\otimes P_{Y|S}\in\Theta_{\mathrm{WD}}\}\neq\emptyset$, i.e., the theorem statement is non-vacuous. In particular, for $K=2$, consider (Y,Y^1,Y^2) such that Y and Y^1 are independent and $Y^2=1-Y^1$, we can see that the resulting instance gives rise to P_S which is in the domain of a_{wd} for any $c\in\mathbb{R}_+^K$ (because here $\gamma_1=\gamma_2=1/2$, thus $\gamma_1-\gamma_2=0$ while $\mathbb{P}\left(Y^1\neq Y^2\right)=1$). While Θ_{WD} is learnable, it is not uniformly learnable, i.e., the minimax regret $\mathfrak{R}_n^*(\Theta_{\mathrm{WD}})=\inf_{\mathfrak{A}}\sup_{\theta\in\Theta_{\mathrm{WD}}}\mathfrak{R}_n(\mathfrak{A},\theta)$ over Θ_{WD} grows linearly:

We close by showing that while Θ_{WD} is learnable and maximal, the price is that it is not uniformly learnable:

Theorem 19. *Θ_{WD} is not uniformly learnable: $\mathfrak{R}_n^*(\Theta_{\mathrm{WD}})=\Omega(n)$.*

Proof. We first consider the case when $K=2$ and arbitrarily choose $C_2-C_1=1/4$. We will consider two instances, $\theta,\theta'\in\Theta_{\mathrm{WD}}$ such that for instance θ, action $k=1$ is optimal with an action gap of $c(2,\theta)-c(1,\theta)=1/4$ between the cost of the second and the first action, while for instance θ', $k=2$ is the optimal action and the action gap is $c(1,\theta)-c(2,\theta)=\epsilon$ where $0<\epsilon<3/8$. Further, the entries in $P_S(\theta)$ and $P_S(\theta')$ differ by at most ϵ. From this, a standard reasoning gives that no algorithm can achieve sublinear minimax regret over Θ_{WD} because any algorithm is only able to identify P_S.

The constructions of θ and θ' are shown in Table 2: The entry in a cell gives the probability of the event as specified by the column and row labels. For example, in instance θ, 3/8 is the probability of $Y=Y^1=Y^2$, while the probability of $Y^1=Y\neq Y^2$ is 1/8. Note that the cells with zero actually correspond to impossible events, i.e., these cannot be assigned a positive probability. The rationale of a redundant (and hence sparse) table is so that probabilities of certain events of interest, such as $Y^1\neq Y^2$ are easier to determine based on the table. The reader should also verify that the positive probabilities correspond to events that are possible.

We need to verify the following: *(i)* $\theta,\theta'\in\Theta_{\mathrm{WD}}$; *(ii)* the optimality of the respective actions in the respective instances; *(iii)* the claim concerning the size of the action gaps; *(iv)* that $P_S(\theta)$ and $P_S(\theta')$ are close. Details of the calculations to support *(i)–(iii)* can be found in Table 3. The row marked by $(*)$ supports that the instances are proper USS instances. In the row marked by $(**)$, there is no requirement for θ' because in θ' action two is optimal, and hence there is no action with larger index than the optimal action, hence $\theta'\in\Theta_{\mathrm{WD}}$ automatically holds.

To verify the closeness of $P_S(\theta)$ and $P_S(\theta')$ we actually would need to first specify P_S (the tables do not fully specify these). However, it is clear the only restriction we put on P_S is the value of $\mathbb{P}\left(Y^1\neq Y^2\right)$ (and that of

Instance θ		$Y^1 = Y^2$	$Y^1 \neq Y^2$
$Y^1 = Y$	$Y^2 = Y$	$\frac{3}{8}$	0
	$Y^2 \neq Y$	0	$\frac{1}{8}$
$Y^1 \neq Y$	$Y^2 = Y$	0	$\frac{1}{8}$
	$Y^2 \neq Y$	$\frac{3}{8}$	0

Instance θ'		$Y^1 = Y^2$	$Y^1 \neq Y^2$
$Y^1 = Y$	$Y^2 = Y$	$\frac{3}{8} - \epsilon$	0
	$Y^2 \neq Y$	0	0
$Y^1 \neq Y$	$Y^2 = Y$	0	$\frac{2}{8} + \epsilon$
	$Y^2 \neq Y$	$\frac{3}{8}$	0

Table 2: The construction of two problem instances for the proof of Theorem 19.

	θ	θ'
$\gamma_1 = \mathbb{P}\left(Y^1 \neq Y\right)$	$\frac{1}{4}$	$\frac{5}{8} + \epsilon$
$\gamma_2 = \mathbb{P}\left(Y^2 \neq Y\right)$	$\frac{1}{4}$	$\frac{3}{8}$
$\gamma_2 \leq \gamma_1{}^{(*)}$	✓	✓
$c(1, \cdot)$	$\frac{1}{4}$	$\frac{5}{8} + \epsilon$
$c(2, \cdot)$	$\frac{2}{4}$	$\frac{5}{8}$
$a^*(\cdot)$	$k = 1$	$k = 2$
$\mathbb{P}\left(Y^1 \neq Y^2\right)$	$\frac{1}{4}$	$\frac{1}{4} + \epsilon$
$\theta \in \Theta_{\text{WD}}{}^{(**)}$	$\frac{1}{4} \geq \frac{1}{4}$ ✓	✓
$\lvert c(1, \cdot) - (2, \cdot) \rvert$	$\frac{1}{4}$	ϵ

Table 3: Calculations for the proof of Theorem 19.

Y^1	Y^2	Y	θ	θ'
0	0	0	$\frac{3}{8}$	$\frac{3}{8} - \epsilon$
0	0	1	$\frac{3}{8}$	$\frac{3}{8} - \epsilon$
0	1	0	0	0
0	1	1	0	0
1	0	0	$\frac{1}{8}$	$\frac{2}{8} + \epsilon$
1	0	1	$\frac{1}{8}$	0
1	1	0	0	0
1	1	1	0	0

Y^1	Y^2	θ	θ'
0	0	$\frac{6}{8}$	$\frac{6}{8} - \epsilon$
0	1	0	0
1	0	$\frac{2}{8}$	$\frac{2}{8} + \epsilon$
1	1	0	0

Table 4: Probability distributions for instances θ and θ'. On the left are shown the joint probability distributions, while on the right are shown their marginals for the sensors.

$\mathbb{P}\left(Y^1 = Y^2\right))$ and these values are within an ϵ distance of each other. Hence, P_S can also be specified to satisfy this. In particular, one possibility for P and P_S are given in Table 4.

□

C Proofs for Section 4

Proposition 7. *If $\theta \in \Theta_{\text{SD}}$, then the regret of π on $f(\theta) \in \mathcal{P}_{\text{side}}$ is the same as the regret of π' on θ.*

Proof. First note that the mapping of the policies is such that number of pull of arm k after n rounds by policy π on problem instance $f(\theta)$ is the same as the number of pulls of arm k by π' on problem instance θ. Recall that mean value of arm k in problem instance θ $is \gamma_k + C_k$ and that of corresponding arm in problem instance $f(\theta)$ is $\gamma_1 - (\gamma_i + C_i)$. We have

$$\mathfrak{R}_n(\pi', \theta) = \sum_{k \in [K]} \mathbb{E}_{P_S}\left[N_k(n)\right]\left(\gamma_k + C_k - \gamma_{k^*} - C_{k^*}\right),$$

and

$$\begin{aligned}\mathfrak{R}_n(\pi, f(\theta)) &= \sum_{k \in [K]} \mathbb{E}_{P_S}\left[N_k(n)\right]\left(\max_{i \in [K]}\{\gamma_1 - \gamma_i - C_i\} - (\gamma_1 - \gamma_k - C_k)\right)\\ &= \sum_{k \in [K]} \mathbb{E}_{P_S}\left[N_k(n)\right]\left(\gamma_k + C_k - \min_{i \in [K]}\{\gamma_i + C_i\}\right)\\ &= \mathfrak{R}_n(\pi', \theta).\end{aligned}$$

□