Unsupervised Sequential Sensor Acquisition

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Sequential sensor acquisition problems (SAP) arise in many application domains including medical-diagnostics, security and surveillance. SAP architecture is organized as a cascaded network of "intelligent" sensors that produce decisions upon acquisition. Sensors must be acquired sequentially and comply with the architecture. Our task is to identify the sensor with optimal accuracy-cost tradeoff. We formulate SAP as a version of the stochastic partial monitoring problem with side information and *unusual* reward structure. Actions correspond to choice of sensor and the chosen sensor's parents decisions are available as side information. Nevertheless, what is atypical, is that we do not observe the reward/feedback, which a learner often uses to reject suboptimal actions. Unsurprisingly, with no further assumptions, we show that no learner can achieve sublinear regret. This negative result leads us to introduce the notion of weak dominance on cascade structures. Weak dominance supposes that a child node in the cascade has higher accuracy whenever its parent's predictions are correct. We then empirically verify this assumption on real datasets. We show that weak dominance is a maximal learnable set in the sense that we must suffer linear regret for any non-trivial expansion of this set. Furthermore, by reducing SAP to a special case of multi-armed bandit problem with side information we show that for any instance in the weakly dominant we only suffer a sublinear regret.

Cs: The story is a bit more complicated. The abstract will need a rewrite once we settle on the results.

1 Introduction

Sequential sensor acquisition arises in many scenarios where we have a diverse collection of sensors with differing costs and accuracy. In these applications, to minimize costs, one often chooses inexpensive sensors first; and based on their outcomes, one sequentially decides whether or not to acquire more expensive sensors. For instance, in security systems(see Trapeznikov et al. (2014) and other medically oriented examples), costs can arise due to sensor availability and delay. A suite of sensors/tests including inexpensive ones such as magnetometers, video feeds, to more expensive ones such as millimeter wave imagers are employed. These sensors are typically organized in a hierarchical architecture with low-cost sensors at the top of the hierarchy. The task is to determine which sensor acquisitions lead to maximizing accuracy for the available cost-budget.

These scenarios motivate us to propose the unsupervised sequential sensor acquisition problem (SAP). Our SAP architecture is organized as a cascaded network of intelligent sensors. The sensors when utilized to probe an instance, outputs a prediction of the underlying state of the instance (anomaly or normal, threat or no-threat etc.). Sensors are ordered with respect to increasing cost and accuracy. While the costs are assumed to be known a priori, the exact misclassification rate of a sensor is unknown. This setup is realistic in security and surveillance scenarios because sensors are often required to be deployed in new domains/environments with little or no opportunity for re-calibration.

We assume that the scenario is played over multiple rounds with an instance associated with each round. Sensors must be acquired sequentially and comply with the cascade architecture in each round.

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

The learner's goal is to figure out the hidden, stochastic state of the instance based on the sensor outputs. Since the learner knows that the sensors are ordered from least to most accurate he/she can use the most accurate sensor among his/her acquired sensors for prediction. Nevertheless, since the learner does not know the sensor accuracy he/she faces the dilemma of as to which sensor to use for predicting this state.

We frame our problem as a version of stochastic partial monitoring problem (Bartók et al., 2014) with *atypical* reward structure. As is common, we pose the problem in terms of competitive optimality. We consider a competitor who can choose an optimal action with the benefit of hindsight. Our goal is to minimize cummulative regret based on learning the optimal action based on observations that are observed during multiple rounds of play.

Stochastic partial monitoring problem is itself a generalization of multi-armed bandit problems, the latter going back to Thompson (1933). In our context, we view sensors choices as actions. The availability of predictions of parent sensors of a chosen sensor is viewed as side observation. Recall that in a stochastic partial monitoring problem a decision maker needs to choose the action with the lowest expected cost by repeatedly trying the actions and observing some feedback. The decision maker lacks the knowledge of some key information, such as in our case, the misclassification error rates of the classifiers, but had this information been available, the decision maker could calculate the expected costs of all the actions (sensor acquisitions) and could choose the best action (sensor). The feedback received by the decision maker in a given round depends stochastically on the unknown information and the action chosen. Bandit problems are a special case of partial monitoring, where the key missing information is the expected cost for each action (or arm), and the feedback is simply the noisy version of the expected cost of the action chosen. In the *unsupervised* version considered here and which we call the unsupervised *sequential sensor acquisition problem* (SAP), the learner only observes the outputs of the classifiers, but not the label to be predicted over multiple rounds in a stochastic, stationary environment.

This leads us to the following question: Can a learner still achieve the optimal balance in this case? We first show that, unsurprisingly, with no further assumptions, no learner can achieve sublinear regret. This negative result leads us to introduce the notion of weak dominance on tests. It is best described as a relaxed notion of strong dominance. Strong dominance states that a sensor's predictions are almost surely correct whenever the parent nodes in the cascade are correct. We empirically demonstrate that weak dominance appears to hold by evaluating it on several real datasets. We also show that in a sense weak dominance is fundamental, namely, without this condition there exist problem instances that result in linear regret. On the other hand whenever this condition is satisfied there exist polynomial time algorithms that lead to sublinear ($O(\sqrt{T})$) cummulative regret.

Cs: Has regret been defined?

Cs: Weak dominance has not been introduced yet.

Our proof of sublinear regret is based on reducing SAP to a version of multi-armed bandit problem (MAB) with side-observation. The latter problem has already been shown to have sub-linear regret in the literature. In our reduction, we identify sensor nodes in the cascade as the bandit arms. The payoff of an arm is given by loss from the corresponding stage, and the side observation structure is defined by the feedback graph induced by the cascade. We then formally show that there is a one-to-one mapping between algorithms for SAP and algorithms for MAB with side-observation. In particular, under weak dominance, the regret bounds for MAB with side-observation then imply corresponding regret bounds for SAP.

2 Related Work

In contrast to our SAP setup there exists a wide body of literature dealing with fully supervised sensor acquisition. Like us Trapeznikov & Saligrama (2013) Wang et al. (2015) Nan et al. (2015) also deal with cascade models. However, unlike us these works focus on prediction-time cost/accuracy tradeoffs. In particular they assume that a fully labeled training dataset is provided for test-time use. This dataset has sensor feature data, sensor decisions as well as annotated ground-truth labels. The goal for the learner is to learn a policy for acquiring sensors based on training data to optimize cost/accuracy during test-time. The work of Póczos et al. (2009) decide when to quit a cascade that leads to better decisions to maximize throughput against error rates. Full feedback about classification accuracy is assumed.

Active classification: Greiner et al. (2002) considers the problem of PAC learning the best "active classifier", a classifier that decides about what tests to take given the results of previous tests to minimize total cost when both tests and misclassification errors are priced. Unlike us they only consider the batch, supervised learning. The same setting is also studied under hard budget constraints in Kapoor & Greiner (2005) and its applications in imaging and computer vision systems are explored in (Draper et al., 1999; Isukapalli & Greiner, 2001)).

Cs: I suspect they assume more than this: In our previous paper we had a sentence that said that "their model requires knowing a model of the actions in advance" (this would mean knowing the joint probabilities, I think).

Online learning: In Seldin et al. (2014), the decision maker can opt to pay for additional observations of the costs associated with other arms. Unlike ours this setting is not unsupervised. In Zolghadr et al. (2013), online learning with costly features and labels is studied. In each round, learner has to decide which features to observe, where each feature costs some money. The learner can also decide not to observe the label, but the learner always has the option to observe the label. Again this setting is not unsupervised.

Cs: Actually, much work exists, need to google this

Partial monitoring: General theory of Bartók et al. (2014) applies to the so-called finite problems (unknown "key information") is an element of the probability simplex. Agrawal et al. (1989) considers special case when the payoff is also observed (akin to the side-observation problem of Mannor & Shamir (2011)Alon et al. (2015),Alon et al. (2013)).

The paper is organized as follows: in Section 3 we give brief background on online learning setups that will be helpful to setup the problem. In Section 4 we introduce the SAP and in Sectjon 5 conditions under which optimal action can be learned are established. In Section 6 we establish SAP is regret equivalent to stochastic multi-armed bandits with side-observations and give an algorithm to solve SAP in Section 7. We conclude in Section 9 with a discussion on further extensions.

3 Background

The purpose of this section is to present some necessary background material that will prove to be useful later. In particular, we introduce a number of sequential decision making problems, namely stochastic partial monitoring, bandits and bandits with side-observations, which we will build upon later.

First, a few words about our notation: We will use upper case letters to denote random variables. 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 set $\mathcal{X}$. When $\mathcal{X}$ is finite with a cardinality of $d \doteq |\mathcal{X}|$, $M_1(\mathcal{X})$ can be identified with the d-dimensional probability simplex.

In a *stochastic partial monitoring problem* 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, $r^*(\theta) = \max_{a \in \mathcal{A}} r(a, \theta)$. The quantification of the learning speed is given by the expected regret $\mathfrak{R}_n = nr^*(\theta) - \mathbb{E}\left[\sum_{t=1}^{n} R_t\right]$, which, for brevity and when it does not cause confusion, we will just call regret. A sublinear expected regret, i.e., $\mathfrak{R}_n/n \to 0$ as $n \to \infty$ means that the learner in the long run collects almost as much reward on expectation as if the optimal action was known to it. Such a learner is called Hannan consistent. In some cases it is more natural to define the problems in terms of costs as opposed to rewards; in such cases the definition of regret is modified appropriately. Transforming between costs and rewards is trivial by flipping the sign of the rewards and costs.

Cs: Not sure whether Hannan consistency is this, or when the random average regret converges to zero with probability one.

A wide range of interesting sequential learning scenarios can be cast as partial monitoring. One special case is bandit problems when $\mathcal{Y}$ is the set of real numbers and $r(a, \theta)$ is the mean of distribution $p(\cdot; a, \theta)$: Thus, in a bandit problem in evert round the learner chooses an action A_t based on its past observations and receives the noisy reward $Y_t \sim p(\cdot; A_t, \theta)$ as feedback. A bandit problem is special in that the observation Y_t and the reward are directly tied. Another special case is finite-armed *bandits with side-observations* Mannor & Shamir (2011), where each action $a \in \mathcal{A}$ is associated with a neighbor-set $\mathcal{N}(a) \subset \mathcal{A}$ and the set of neighborhoods is known to the learner from the beginning. The learner upon choosing action $A_t \in \mathcal{A}$ receives noisy reward observations for each action in $\mathcal{N}(A_t)$: $Y_t = (Y_{t,a})_{a \in N(A_t)}$, where $Y_{t,a} \sim p_r(\cdot; a, \theta)$, and $\mathbb{E}[Y_{t,a}] = r(a, \theta)$. (The action chosen may or may not be an element of $N(A_t)$.) The reader can readily verify that this problem can also be cast as a partial monitoring problem by defining $\mathcal{Y}$ as the set $\cup_{i=0}^{K} \mathbb{R}^i$ and defining the family of distributions $(p(\cdot; a, \theta))_{a,\theta}$ such that $Y_t \sim p(\cdot; A_t, \theta)$. Finally, we note in passing that while we called Θ a parameter set, we have not equipped Θ with any structure. As such, the framework is able to model both bona fide parametric settings (e.g., Bernoulli rewards) and the so-called non-parametric settings. For example, K-armed bandits with reward distributions supported over $[0, 1]$ can be modelled by choosing Θ as the set of all K-tuples $\theta := (\theta_1, \ldots, \theta_K)$ of distributions over $[0, 1]$ and setting $p(\cdot; a, \theta) = \theta_a(\cdot)$. More generally, we can identify Θ with set of instances $(p(\cdot; a, \theta), r(a, \theta))_{\theta \in \Theta}$. In what follows, when convenient, we will use this identification and will view elements of Θ as a pair p, r where $p(\cdot; a)$ is a probability distribution over $\mathcal{Y}$ for each $a \in \mathcal{A}$ and r is a map from $\mathcal{A}$ to the reals.

4 Unsupervised Sensor Acquisition Problem

Cs: I compressed the problem spec. We don't want the reader to get bored.

The formal problem specification of the unsupervised, stochastic, cascaded sensor acquisition problem is as follows: A problem instance is specified by a pair $\theta = (P, c)$, where P is a distribution over the $K + 1$ dimensional hypercube, and c is a K-dimensional, nonnegative valued vector of costs. While c is known to the learner from the start, P is initially unknown. The instance parameters specify the learner-environment interaction as follows: In each round for $t = 1, 2, \ldots$, the environment generates a $K + 1$-dimensional binary vector $Y = (Y_t, Y_t^1, \ldots, Y_t^K)$ 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 $Y_t^1, \ldots, Y_t^{I_t}$. The sensors are known to be ordered from least accurate to most accurate, i.e., $\gamma_k \doteq \mathbb{P}\left(Y_t \neq Y_t^k\right)$ 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 Y_t. Observing sensors is costly: The cost of choosing I_t is $C_{I_t} \doteq c_1 + \cdots + c_{I_t}$. 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 $(\gamma_k)_k$. The expected regret of learner up to the end of round n is $\mathfrak{R}_n = (\sum_{t=1}^{n} \mathbb{E}\left[C_{I_t} + \mathbb{I}\{Y_t \neq Y_t^{I_t}\}\right]) - n \min_k (C_k + \gamma_k)$. For future reference, we let $c(k, \theta) = \mathbb{E}\left[C_k + \mathbb{I}\{Y_t \neq Y_t^k\}\right] (= C_k + \gamma_k)$ and $c^*(\theta) = \min_k c(k, \theta)$. Thus, $\mathfrak{R}_n = (\sum_{t=1}^{n} \mathbb{E}[c(I_t, \theta)]) - nc^*(\theta)$. In what follows, we shall denote by $\mathcal{A}^*(\theta)$ the set of optimal actions of θ and we let $a^*(\theta)$ denote the optimal action that has the smallest index. Thus, in particular, $a^*(\theta) = \min \mathcal{A}^*(\theta)$. 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., $\gamma_1 = 0.3$, $C_1 = 0$, $\gamma_2 = 0.25$, $C_2 = 0.1$, $\gamma_3 = 0$, $C_3 = 0.3$. Next, for future reference note that one can express optimal actions from the viewpoint of marginal costs and marginal error. 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, $\gamma_i - \gamma_j$:

Cs: I added stochastic and cascaded. Later we may want to consider alternatives, thus it will be useful to have these so that we can distinguish between the problem defined here and those future alternatives.

Cs: Add proper figure.

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

5 When is SAP Learnable?

Let Θ_{SA} be the set of all stochastic, cascaded sensor acquisition problems. Thus, $\theta = (P, c) \in \Theta_{\mathrm{SA}}$ such that if $Y \sim P$ then $\gamma_k(\theta) := \mathbb{P}\left(Y \neq Y^k\right)$ is a decreasing sequence. Given a subset $\Theta \subset \Theta_{\mathrm{SA}}$, we say that Θ is *learnable* if there exists a learning algorithm $\mathfrak{A}$ such that for any $\theta \in \Theta$, the expected regret $\mathbb{E}[\mathfrak{R}_n(\mathfrak{A}, \theta)]$ 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 $\Theta' \subset \Theta_{\mathrm{SA}}$ superset of Θ, Θ' is not learnable. In this section we study two special learnable problem classes, $\Theta_{\mathrm{SD}} \subset \Theta_{\mathrm{WD}}$, 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 $\theta = (P, c) \in \Theta_{\mathrm{SA}}$, we can decompose P into the joint distribution P_S of the sensor outputs $S = (Y^1, \ldots, Y^k)$ 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 $y \in \{0,1\}$, $P_S(s) = \mathbb{P}(S=s)$ and $P_{Y|S}(y|s) = \mathbb{P}(Y=y|S=s)$. 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 $\Theta \subset \Theta_{\mathrm{SA}}$ is learnable if and only if there exists a map $a : M_1(\{0,1\}^K) \times \mathbb{R}_+^K \to [K]$ 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.*

Proof. $\Rightarrow$: Let $\mathfrak{A}$ be an algorithm that achieves sublinear regret and pick an instance $\theta = (P,c) \in \Theta$. 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 $a(P_S,c) = \min\{k \in [K]\,;\, \mathbb{E}_{P_S}[N_k(n)] = \Omega(n)\,\}$. Then, a is well-defined as the distribution of $N_k(n)$ for any k depends only on P_S and c. Furthermore, $a(P_S,c)$ 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 $n \in \mathbb{N}$, $f(n)/\log(n) \to n$ 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 $I_t = a(\hat{P}_S,c)$ in the remaining $n - f(n)$ trials. Pick any $\theta = (P,c) \in \Theta$ so that $P = P_S \otimes P_{Y|S}$. 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 $n \times 1/n + f(n)(1-1/n) \le 1 + f(n) = o(n)$. $\square$

An action selection map $a : M_1(\{0,1\}^K) \times \mathbb{R}_+^K \to [K]$ is said to be *sound* for an instance $\theta \in \Theta_{\mathrm{SA}}$ with $\theta = (P_S \otimes P_{Y|S}, c)$ if $a(P_S,c)$ 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 acquisition problems that contains instances that satisfy the so-called *strong dominance* condition will be shown to be learnable:

Definition 1 (Strong Dominance)**.** *An instance $\theta = (P,c) \in \Theta_{\mathrm{SA}}$ 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 $i \in [K]$,*

$$Y^i = Y \;\Rightarrow\; Y^{i+1} = \cdots = Y^K = Y \tag{2}$$

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

dataset	γ_1	γ_2	δ_{12}
diabetic	0.288	0.219	0.075
heart	0.305	0.169	0.051

Table 1: Error statistics

Before we develop this concept further we will motivate strong dominance based on experiments on a few real-world datasets. Table 5 lists the error probabilities of the classifiers (sensors) for the heart and diabetic datasets from UCI repository. For both the datasets, γ_1 denotes the test error of an SVM classifier (linear) trained with low cost features and γ_2 denotes test error of

SVM classifier trained using both low and high-cost features (cf. Section 8). The last column lists $\delta_{12} := \mathbb{P}\left(Y^1 = Y, Y^2 \neq Y\right)$, the probability that second sensor misclassifies an instance that is correctly classified by the first sensor. Strong dominance 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. Note that strong dominance is not merely a consequence of improved accuracy with availability of more features. It is related to better *recall rates* of high-cost features relative to low-cost features.

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

Theorem 1. *The set Θ_{SD} is learnable.*

We start with a proposition that will be useful beyond the proof of this result. In this proposition, $\gamma_i = \gamma_i(\theta)$ for $\theta = (P, c) \in \Theta_{\mathrm{SA}}$ and $(Y, Y^1, \dots, Y^K) \sim P$.

Proposition 2. *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. We construct a map as required by Proposition 1. Take an instance $\theta = (P, c) \in \Theta_{\mathrm{WD}}$ and let $P = P_S \otimes P_{Y|S}$ be its decomposition as defined above. Let $\gamma_i = \mathbb{P}\left(Y^i \neq Y\right)$, $(Y, Y^1, \dots, Y^K) \sim P$. For identifying an optimal action in θ, it clearly suffices to know the sign of $\gamma_i + C_i - (\gamma_j + C_j)$ for all pairs $i, j \in [K]^2$. Since $C_i - C_j$ is known, it remains to study $\gamma_i - \gamma_j$. Without loss of generality (WLOG) let $i < j$. Then,

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

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. □

Proof of Theorem 1. We construct a map as required by Proposition 1. Take an instance $\theta = (P, c) \in \Theta_{\mathrm{SD}}$ and let $P = P_S \otimes P_{Y|S}$ be its decomposition as before. Let $\gamma_i = \mathbb{P}\left(Y^i \neq Y\right)$, $(Y, Y^1, \dots, Y^K) \sim P$, $C_i = c_1 + \dots + c_i$. 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 2, $\gamma_i - \gamma_j = \mathbb{P}\left(Y^i \neq Y^j\right) - 2\mathbb{P}\left(Y^j \neq Y, Y^i = Y\right)$. 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. □

The proof motivates the definition of weak dominance, a concept that we develop next through a series of smaller propositions. In these propositions, as before $(Y, Y^1, \dots, Y^K) \sim P$ where $P \in M_1(\{0,1\}^{K+1})$, $\gamma_i = \mathbb{P}\left(Y^i \neq Y\right)$, $i \in [K]$, and $C_i = c_1 + \dots + c_i$. We start with a corollary of Proposition 2

Corollary 1. *Let $i < j$. Then $0 \leq \gamma_i - \gamma_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$.*

Proposition 3. *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. $\Rightarrow$: From the premise, it follows that $C_j - C_i \geq \gamma_i - \gamma_j$. Thus, by (3), $C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. $\Leftarrow$: 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 1. □

Proposition 4. *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. $\Rightarrow$: The condition $\gamma_i + C_i \leq \gamma_j + C_j$ implies that $\gamma_j - \gamma_i \geq C_i - C_j$. By Corollary 1 we get $\mathbb{P}\left(Y^i \neq Y^j\right) \geq C_i - C_j$. $\Leftarrow$: Let $C_i - C_j \leq \mathbb{P}\left(Y^i \neq Y^j\right)$. Then, by (4), $C_i - C_j \leq \gamma_j - \gamma_i$. $\square$

These results motivate the following definition:

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

$$\forall j > i \ : \ C_j - C_i \geq \mathbb{P}\left(Y^i \neq Y^j\right) \ . \tag{5}$$

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

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

We now relate weak dominance to the optimality condition described in Eq. (1). Weak dominance can be viewed as a more stringent condition for optimal actions. Namely, for an action to be optimal we also require that the marginal cost be larger than marginal *absolute* error:

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

The difference between marginal error in Eq. (1) and marginal absolute error is the presence of the absolute value. We will show later that weak-dominant set is a maximal learnable set, namely, the set cannot be expanded while ensuring learnability.

We propose the following action selector $a_{\mathrm{wd}} : M_1(\{0,1\}^K) \times \mathbb{R}_+^K \to [K]$:

Definition 3. *For $(P_S, c) \in M_1(\{0,1\}^K) \times \mathbb{R}_+^K$ let $a_{\mathrm{wd}}(P_S, c)$ 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.)

Proposition 5. *For any $\theta = (P, c) \in \Theta_{\mathrm{WD}}$ with $P = P_S \otimes P_{Y|S}$, $a_{\mathrm{wd}}(P_S, c)$ is well-defined.*

Proof. Let $\theta \in \Theta_{\mathrm{WD}}$, $i = a^*(\theta)$. Obviously, (7b) holds by the definition of Θ_{WD}. Thus, the only question is whether (7a) also holds. We prove this by contadiction: Thus, assume that (7a) does not hold, i.e., for some $j < i$, $C_i - C_j \geq \mathbb{P}\left(Y^i \neq Y^j\right)$. Then, by Corollary 1, $\mathbb{P}\left(Y^i \neq Y^j\right) \geq \gamma_j - \gamma_i$, hence $\gamma_j + C_j \leq \gamma_i + C_i$, which contradicts the definition of i, thus finishing the proof. $\square$

Proposition 6. *The map a_{wd} is sound over Θ_{WD}: In particular, for any $\theta = (P, c) \in \Theta_{\mathrm{WD}}$ with $P = P_S \otimes P_{Y|S}$, $a_{\mathrm{wd}}(P_S, c) = a^*(\theta)$.*

Proof. Take any $\theta \in \Theta_{\mathrm{WD}}$ and let $\theta = (P, c)$ with $P = P_S \otimes P_{Y|S}$, $i = a_{\mathrm{wd}}(P_S, c)$, $j = a^*(\theta)$. 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 1, $\mathbb{P}\left(Y^i \neq Y^j\right) \geq \gamma_i - \gamma_j$. 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 $i = j$. $\square$

Corollary 2. *The set Θ_{WD} is learnable.*

Proof. By Proposition 5, a_{wd} is well-defined over Θ_{WD}, while by Proposition 6, a_{wd} is sound over Θ_{WD}. By Proposition 1, Θ_{WD} is learnable, as witnessed by a_{wd}. $\square$

Cs: We should add definitions for these concepts.. namely, a_{wd} well-defined over Θ_{WD}, a_{wd} sound over Θ_{WD}, etc.

Proposition 7. *Let $\theta \in \Theta_{\mathrm{SA}}$, $\theta = (P, c)$, $P = P_S \otimes P_{Y|S}$ be such that a_{wd} is defined for P_S, c and $a_{\mathrm{wd}}(P_S, c) = a^*(\theta)$. Then $\theta \in \Theta_{\mathrm{WD}}$.*

Proof. Immediate from the definitions. $\square$

An immediate corollary of the previous proposition is as follows:

Corollary 3. *Let $\theta \in \Theta_{\mathrm{SA}}$, $\theta = (P, c)$, $P = P_S \otimes P_{Y|S}$. Assume that a_{wd} is defined for (P_S, c) and $\theta \notin \Theta_{\mathrm{WD}}$. Then $a_{\mathrm{wd}}(P_S, c) \neq a^*(\theta)$.*

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

Proposition 8. *Take any action selector map $a : M_1(\{0,1\}^K) \times \mathbb{R}_+^K \to [K]$ which is sound over Θ_{WD}. Then, for any (P_S, c) such that $\theta = (P_S \otimes P_{Y|S}, c) \in \Theta_{\mathrm{WD}}$ with some $P_{Y|S}$, $a(P_S, c) = a_{\mathrm{wd}}(P_S, c)$.*

Proof. Pick any $\theta = (P_S \otimes P_{Y|S}, c) \in \Theta_{\mathrm{WD}}$. If $A^*(\theta)$ is a singleton, then clearly $a(P_S, c) = a_{\mathrm{wd}}(P_S, c)$ 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 $\theta' = (P_S \otimes P'_{Y|S}, c)$ 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,

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

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 $(y^1, \ldots, y^{j-1}, y^{j+1}, \ldots, y^K) \in \{0,1\}^{K-1}$, 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. $\square$

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

Theorem 2. *Let $a : M_1(\{0,1\}^K) \times \mathbb{R}_+^K \to [K]$ be an action selector map such that a is sound over the instances of Θ_{WD}. Then there is no instance $\theta = (P_S \otimes P_{Y|S}, c) \in \Theta_{\mathrm{SA}} \setminus \Theta_{\mathrm{WD}}$ such that $(P_S, c) \in \mathrm{dom}(a_{\mathrm{wd}})$, the optimal action of θ is unique and $a(P_S, c) = a^*(\theta)$.*

Cs: It would be nice to remove this uniqueness assumption, but I don't see how this could be made to work.

Note that $\mathrm{dom}(a_{\mathrm{wd}}) \setminus \{(P_S, c) : \exists P_{Y|S} \text{ s.t. } (P_S \otimes P_{Y|S}, c) \in \Theta_{\mathrm{WD}}\} \neq \emptyset$, 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

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 3.

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 $\mathbb{P}\left(Y^1 \neq Y^2\right) = 1$).

Proof. Let a as in the theorem statement. By Proposition 8, a_{wd} is the unique sound action-selector map over Θ_{WD}. Thus, for any $\theta = (P_S \otimes P_{Y|S}, c) \in \Theta_{\mathrm{WD}}$, $a_{\mathrm{wd}}(P_S, c) = a(P_S, c)$. Hence, the result follows from Corollary 3. $\square$

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:

Theorem 3. 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.

Cs: The theorem statement should be refined or this text..

Cs: Add notation of $P_S(\theta)$ early on. Probably a good idea to add $P_S(\Theta)$ as a notation too for the "projection" of Θ to P_S. Also, we should probably remove c from the instance definition; in every case we are reasoning for a fixed c, hence it is superfluous to keep c in the instance definition.

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 SAP 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 $\mathbb{P}\left(Y^1 = Y^2\right)$) 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.

$\square$

6 Regret Equivalence

In this section we establish that SAP with strong dominance property is 'regret equivalent' to an instance of MAB with side-information and the corresponding algorithm for MAB can be suitably imported to solve SAP efficiently.

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

Table 3: Calculations for the proof of Theorem 3.

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.

Let $\mathcal{P}_{\rm SAP}$ be the set of SAPs 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 $(P, c) \in \mathcal{P}_{\rm SAP}$ and let $(Y, Y^1, \ldots, Y^K) \sim P$ be the unobserved state of environment in round s. 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).

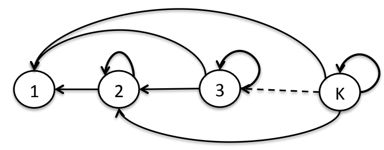

Figure 1: Neighborhood structure in bandit problem equivalent of SAP

The costs for different arms are defined to be independent of each other. Let $\mathcal{P}_{\rm side}$ denote the set of resulting bandit problems and let $f : \mathcal{P}_{\rm SAP} \to \mathcal{P}_{\rm side}$ be the map that transforms SAP instances to bandit instances by following the transformation that was just described.

Cs: Ok, so if they are independent of each other, then the joint distributions will not be same as if they were not independent of each other. Independence may lose information (e.g., may increase variance?). If we define them not to be independent of each other, we will need to be careful with the algorithms defined for bandits with side-observation: Do they use (in their proof) independence of rewards underlying different arms? I would think that they are not. The downside of not defining independent rewards is that the specification of bandits with side observations must allow this – complicating things a bit in the background. Another executive decision we should make is whether we like to see both costs and rewards.

Now let $\pi \in \Pi(\mathcal{P}_{\rm side})$ be a policy for $\mathcal{P}_{\rm side}$. Policy π can also be used on any (P, c) instance in $\mathcal{P}_{\rm SAP}$ 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 9. *If $\theta \in \Theta_{\rm SD}$, then the regret of π on $f(\theta) \in \mathcal{P}_{\rm side}$ 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. Also, 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

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

□

This implies that $\mathfrak{R}_T^*(\Theta_{\rm SD}) \le \mathfrak{R}_T^*(f(\Theta_{\rm SD}))$.

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 $\mathcal{P}_{\rm side}$) and if π' is a policy for $\mathcal{P}_{\rm SAP}$, then π' can be "used" on any instance $\theta \in \mathcal{P}_{\rm side}$ 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 $W_s = (B_s^1, \ldots, B_s^{A_s})$. Let the resulting policy be denoted by π. Then the following holds:

Proposition 10. *Let $\theta \in f(\Theta_{\rm SD})$. Then the regret of policy π on $\theta \in f(\Theta_{\rm SD})$ is the same as the regret of policy π' on instance $f^{-1}(\theta)$.*

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

Corollary 4. $\mathfrak{R}_T^*(\Theta_{\rm SD}) = \mathfrak{R}_T^*(f(\Theta_{\rm SD}))$.

Cs: So this could in theory be used for upper and lower bounds.. However, $\mathcal{P}_{\text{side}}$ is really special (because of the fixed costs) – hence it is unclear whether existing lower bounds, for example, would apply. The next step could be to describe policies for bandits with side observation starting from our paper with Yifan. We have two types of policies. One is asymptotically optimal, the other is minimax optimal. Can we have a single policy in our special problem that would be simultanously optimal in both cases? What happens when only weak dominance is satisfied?

7 Algorithm

The reduction of the previous section suggests that one can play in an SAP instance by utilizing an algorithm developed for stochastic bandits with side-observation. In this paper we make use of Algorithm 1 of Wu et al. (2015). While this algorithm was proposed for stochastic bandits with Gaussian side observations, as noted in the above paper, the algorithm is also suitable for problems where the payoff distributions are subgaussian. As Bernoulli random variables are $\sigma^2 = 1/4$-subgaussian (after centering), the algorithm is also applicable in our case.

Cs: Note the bug in the other paper.

Algorithm 1

1: Inputs: $\alpha > 0$ and $\beta : \mathbb{N} \to [0, \infty)$.
2: Play action K and observe the sensor outputs $Y^1, \dots, Y^K$.
3: Set $\hat{\gamma}(1) \leftarrow (0, \mathbb{I}_{\{Y^1 \neq Y^2\}}, \dots, \mathbb{I}_{\{Y^1 \neq Y^K\}})$.
4: Initialize the exploration count: $n_e \leftarrow 0$.
5: Initialize the allocation counts: $N_i(1) = \mathbb{I}_{\{i=K\}}$, $i \in [K]$.
6: **for** $t = 2, 3, \ldots$ **do**
7:     **if** $\frac{N(t-1)}{4\alpha \log t} \in C(\hat{\gamma}(t-1))$ **then**
8:         Set $I_t \leftarrow \operatorname{argmin}_{k \in [K]} c(k, \hat{\gamma}(t-1))$.
9:     **else**
10:         **if** $N_K(t-1) < \beta(n_e)/K$ **then**
11:             Set $I_t = K$.
12:         **else**
13:             Set I_t to some i for which $N_i(t-1) < u_i^*(\hat{\gamma}(t-1)) 4\alpha \log t$.
14:         **end if**
15:         Increment exploration count: $n_e \leftarrow n_e + 1$.
16:     **end if**
17:     Play I_t and observe the sensor outputs $Y^1, \dots, Y^{I_t}$.
18:     For $i \in [I_t]$, set $\hat{\gamma}_i(t) \leftarrow (1 - 1/t)\hat{\gamma}_i(t-1) + 1/t\, \mathbb{I}_{\{Y^1 \neq Y^i\}}$.
19: **end for**

For the convenience of the reader, we give the algorithm resulting from applying the reduction to Algorithm 1 of Wu et al. (2015) in an explicit form. For specifying the algorithm we need some extra notation. Recall that given a SAP instance $\theta = (P, c)$, we let $\gamma_k = \mathbb{P}\left(Y \neq Y^k\right)$ where $(Y, Y^1, \dots, Y^K) \sim P$ 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. 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: $C(\gamma) = \{u \in [0, \infty)^K \,:\, u_j + u_{j+1} + \cdots + u_K \geq \frac{2\sigma^2}{(\Delta_j^*(\theta))^2}, j \in [K]\}$ (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, $u_k^*(\gamma) = \min\{u_k \,:\, u \in C(\gamma)\}$ (here, $\langle x, y \rangle = \sum_i x_i y_i$ is the standard inner product of vectors x, y). 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 the algorithm shown as Algorithm 1 is as follows: The algorithm keeps track of an estimate $\hat{\gamma}(t)$ 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 current allocation count $N(t) \in \mathbb{N}^K$ is sufficiently informative (cf. line 7). If this holds, the action that is optimal under $\hat{\gamma}(t)$ is chosen (cf. line 8). 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 13). 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 extremely poorly estimated (line 10). 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 11 overall has a small impact on the regret, while it allows to stabilize the algorithm.

For $\theta = (P, c) \in \Theta_{\mathrm{SD}}$, let $\gamma(\theta)$ be the error probabilities for the various sensors. The following result follows from Theorem 6 of Wu et al. (2015):

Theorem 4. *Let $\epsilon > 0$, $\alpha > 2$ arbitrary and choose any non-decreasing $\beta(n)$ that satisfies $0 \leq \beta(n) \leq n/2$ and $\beta(m+n) \leq \beta(m) + \beta(n)$ for $m, n \in \mathbb{N}$. Then, for any $\theta = (P, c) \in \Theta_{\mathrm{SD}}$, letting $\gamma = \gamma(\theta)$ the expected regret of Algorithm 1 after T steps satisfies*

Cs: Actually, needs to be checked.. I also replaced $d_{\max}(\theta)$ with 1.

$$
\begin{aligned}
R_T(\theta, c) \leq \big(2K + 2 + 4K/(\alpha - 2)\big) + 4K \sum_{s=0}^{T} \exp\Big(-\frac{8\beta(s)\epsilon^2}{2K}\Big) \\
+ 2\beta\Big(4\alpha \log T \sum_{i \in [K]} u_i^*(\gamma, \epsilon) + K\Big) + 4\alpha \log T \sum_{i \in [K]} u_i^*(\gamma, \epsilon) d_i(\gamma)\,,
\end{aligned}
$$

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

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

Cs: I just copy&pasted this. We don't actually have a lower bound..

Corollary 5. *Suppose the conditions of Theorem 4 hold. Assume, furthermore, that $\beta(n)$ satisfies $\beta(n) = o(n)$ and $\sum_{s=0}^{\infty} \exp\Big(-\frac{\beta(s)\epsilon^2}{2K\sigma^2}\Big) < \infty$ for any $\epsilon > 0$, then for any θ such that $u^*(\theta)$ is unique,*

$$
\limsup_{T \to \infty} R_T(\theta, c)/\log T \leq 4\alpha \inf_{u \in C_\theta} \langle u, d(\gamma(\theta)) \rangle\,.
$$

Note that any $\beta(n) = an^b$ with $a \in (0, \frac{1}{2}]$, $b \in (0, 1)$ satisfies the requirements in Theorem 4 and Corollary 5.

dataset	γ_1	γ_2	p_{12}	δ_{12}
BSC	.2	.1	.261	.08
diabetic	0.288	0.219	0.219	0.075
heart	0.305	0.169	0.237	0.051

Figure 2: Error statistics

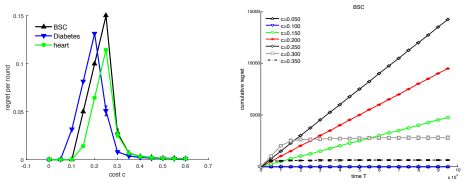

Figure 3: Left side figure plots regret per round against cost for all the datasets. The right side plots regret for different cost in BSC experiment

8 Experiments

In this section we apply bandit algorithms on SA-problem and evaluate its performance on synthetic and real datasets. For synthetic example, we consider data transmission over a binary symmetric channel, and for real world examples, we use diabetes (PIMA indiana) and heart disease (Clevland) from UCI dataset. In both datasets attributes/features are associated with costs, where features related to physical observations are cheap and that obtained from medical tests are costly. The experiments are setup as follows:

Synthetic: we consider data transmission over two binary symmetric channels (BSCs). Channel $i = 1, 2$ flips input bit with probability p_i and $p_1 \geq p_2$. Transmission over channel 1 is free and that over channel 2 costs $c_2 \in (0, 1]$ units per bit. Input bits are generated with uniform probability and we set $p_1 = .2$ and $p_2 = .1$.

Datatsets: we obtain a sensor acquisition setup from the datasets as follows: Two svm classifiers (linear, $C = .01$) are trained for each dataset, one using only cheap features, and the other using all features. These classifiers form sensors of a two stage SAP where classifier trained with cheap features is the first stage and that trained with all features forms the second stage. Cost of each stage is the sum of cost of features used to train that stage multiplied by a scaling factor λ (trade-off parameter for accuracy and costs). Specific details for each dataset is given below.

PIMA indians diabetes dataset consists of 768 instances and has 8 attributes. The labels identify if the instances are diabetic or not. 6 of the attributes (age, sex, triceps, etc.) obtained from physical observations are cheap, and 2 attributes (glucose and insulin) require expensive tests. First sensor of SAP is trained with 6 cheap attributes and costs \$6. Second sensor is trained from all 8 attributes that cost \$30. We set $c_1 = 6\lambda, c_2 = 30\lambda$ and $c = 24\lambda$.

Heart disease dataset consists of 297 instance (without missing values) and has 13 attributes. 5 class labels $(0, 1, 2, 3, 4)$ are mapped to binary values by taking value 0 as 'absence' of disease and values $(1, 2, 3, 4)$ as 'presence' of disease. First senor of SAP is trained with 7 attributes which cost \$1 each. Total cost of all attributes is \$568. We set $c_1 = 7\lambda, c_2 = 568\lambda$ and $c = 561\lambda$.

Various error probabilities for synthetic and datasets are listed in Table (8). The probabilities for the datasets are computed on 20% hold out data. To run the online algorithm, an instance is randomly selected from the dataset in each round and is input to the algorithm. We repeat the experiments 20 times and average is shown in (8) with 95% confidence bounds. The left Figure in 8 depicts regret per round vs. cost c for each setup. As seen, regret per round is positive over an interval where it is increasing and then drops to zero sharply. For all c in $[0.1\ 0.26]$, $[0.07\ 0.21]$, $[0.13,\ 0.237]$ for synthetic, diabetes and heart dataset, respectively, the regret per round is positive implying that regret is linear in these regions, and regret per round sharply falls to zero outside this region implying sublinear regret there. This is in agreement with the weak dominance property. For the BSC setup, regret is plotted on the right of Figure (8). As seen, regret is linear for all c in $[0.1\ 0.26]$ and is sublinear outside this region.

9 Conclusions

We need to conclude soon.

10 Appendix

Consider a K-armed stochastic bandit problem where reward distribution ν_i has mean $\gamma_1 - \gamma_i - \sum_{j<i} c_j$ for all $i > 1$ and arm 1 gives a fixed reward of value 0. The arms have side-observation structure defined by graph G_S. Given an arbitrary policy $\pi = (\pi_1, \pi_2, \cdots \pi_t)$ for the SAP, we obtain a policy for the bandit problem with side observation graph G_S from π as follows: Let H_{t-1} denote the history, consisting of all arms played and the corresponding rewards, available to policy π_{t-1} till time $t-2$. In round $t-1$, let a_{t-1} denote the arm selected by the bandit policy, r_{t-1} the corresponding reward and o_{t-1} the side-observation defined by graph G_S. Then, the next action a_t is obtained as follows:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{arm } 1 \\ \pi_t(H_{t-1} \cup \{i, r_{t-1} \cup o_{t-1}\}) \text{ if } a_{t-1} = \text{arm i} \end{cases} \tag{8}$$

Conversely, let $\pi' = \{\pi'_1, \pi'_2, \cdots\}$ denote an arbitrary policy for the K-armed bandit problem with side-observation graph. we obtain a policy the SAP as follows: Let H'_{t-1} denote the history, consisting of all actions played and feedback, available to policy π'_{t-1} till time $t-2$. Let a'_{t-1} denote the action selected by the SAP policy in round $t-1$ and observed feedback F_t. Then, the next action a'_t is obtained as follows:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1, 0\}) \text{ if } a'_{t-1} = \text{action } 1 \\ \pi'_t(H'_{t-1} \cup \{i, \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} \cdots \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^i\}\}) \text{ if } a_{t-1} = \text{action i.} \end{cases} \tag{9}$$

We next show that regret of a policy π on the SAP problem is same as that of the policy derived from it for the K-armed bandit problem with side information graph G_S, and regret of π' on the K-armed bandit with side-observation graph G_S is same as that of the policy derived from it for the SAP.

Given a policy π for the SAP problem let $f_1(\pi)$ denote the policy obtained by the mapping defined in (8). The regret of policy π that plays actions i, $N_i^\psi(T)$ times is given by

$$R_T^\psi(\pi) = \sum_{i=1}^{K} \left[\left(\gamma_i + \sum_{j<i} c_j \right) - \left(\gamma_{i^*} + \sum_{j<i^*} c_j \right) \right] \mathbb{E}[N_i^\psi(T)] \tag{10}$$

$$\tag{11}$$

Now, regret of regret policy $f_1(\pi)$ on the K-armed bandit problem with side-observation graph G_S

$$R_T^\phi(f_1(\pi)) = \sum_{i=1}^{K} \left[\left(\gamma_1 - \gamma_{i^*} - \sum_{j<i^*} c_j \right) - \left(\gamma_1 - \gamma_i - \sum_{j<i} c_j \right) \right] \mathbb{E}[N_i^\phi(T)], \tag{12}$$

where $N_i^\phi(T)$ is the number of times arm i is pulled by policy $f_1(\pi)$. Since the mapping is such that $N_i^\phi(T) = N_i^\psi(T)$, $R_T^\phi(f_1(\pi))$ is same as $R_T^\psi(\pi)$. Further, given a policy π' on ψ we can obtain a policy $f_2(\psi)$ for ψ as defined in (9) and we can argue similarly that they are regret equivalent. This concludes the proof.

11 Extension to context based prediction

In this section we consider that the prediction errors depend on the context X_t, and in each round the learner can decide which action to apply based on X_t. Let $\gamma_i(X_t) = \Pr\{\hat{Y}_t^1 \neq \hat{Y}_t^2 | X_t\}$ for all $i \in [K]$. We refer to this setting as Contextual Sensor Acquisition Problem (CSAP) and denote it as $\psi_c = (K, \mathcal{A}, \mathcal{C}, (\gamma_i, c_i)_{i \in [K]})$.

Given $x \in \mathcal{C}$, let $L_t(a|x)$ denote the loss from action $a \in \mathcal{A}$ in round t. A policy on ϕ^c maps past history and current contextual information to an action. Let Π^{ψ_c} denote set of policies on ψ_c and for any policy $\pi \in \Pi^{\psi_c}$, let $\pi(x_t)$ denote the action selected when the context is x_t. For any sequence $\{x_t, y_t\}_{t>0}$, the regret of a policy π is defined as:

$$R_T^{\phi_c}(\pi) = \sum_{t=1}^{T} \mathbb{E}\left[L_t(\pi(x_t)|x_t)\right] - \sum_{t=1}^{T} \min_{a \in \mathcal{A}} \mathbb{E}\left[L_t(a|x_t)\right]. \tag{13}$$

As earlier, the goal is to learn a policy that minimizes the expected regret, i.e., $\pi^* = \arg\min_{\pi \in \Pi^{\psi_c}} \mathbb{E}[R_T^{\psi_c}(\pi)]$.

In this section we focus on CSA-problem with two sensors and assume that sensor predictions errors are linear in the context. Specifically, we assume that there exists $\theta_1, \theta_2 \in \mathcal{R}^d$ such that $\gamma_1(x) = x'\theta_1$ and $\gamma_2(x) + c = x'\theta_2$ for all $x \in \mathcal{C}$, were x' denotes the transpose of x. By default all vectors are column vectors. In the following we establish that CSAP is regret equivalent to a stochastic liner bandits with varying decision sets. We first recall the stochastic linear bandit setup and relevant results.

11.1 Background on Stochastic Linear Bandits

In round t, the learner is given a decision set $D_t \subset \mathcal{R}^d$ from which he has to choose an action. For a choice $x_t \in D_t$, the learner receives a reward $r_t = x_t'\theta^* + \epsilon_t$, where $\theta^* \in \mathcal{R}^d$ is unknown and ϵ_t is random noise of zero mean. The learner's goal is to maximize the expected accumulated reward $\mathbb{E}\left[\sum_{t=1}^T r_t\right]$ over a period T. If the leaner knows θ^*, his optimal strategy is to select $x_t^* = \arg\max_{x \in D_t} x'\theta^*$ in round t. The performance of any policy π that selects action x_t at time t is measured with respect to the optimal policy and is given by the expected regret as follows

$$R_T^L(\pi) = \sum (x_t^*)'\theta^* - \sum x_t'\theta^*. \tag{14}$$

The above setting, where actions sets can change in every round, is introduced inAbbasi-Yadkori et al. (2011) and is a more general setting than that studied in Dani et al. (2008); Rusmevichientong & Tsitsiklis (2010) where decision set is fixed. Further, the above setting also specializes the contextual bandit studied in Li et al. (2010). The authors in Abbasi-Yadkori et al. (2011) developed an 'optimism in the face of uncertainty linear bandit algorithm' (OFUL) that achieves $\mathcal{O}(d\sqrt{T})$ regret with high probability when the random noise is R-sub-Gaussian for some finite R. The performance of OFUL is significantly better than $ConfidenceBall_2$ Dani et al. (2008), $UncertainityEllipsoid$ Rusmevichientong & Tsitsiklis (2010) and $LinUCB$ Li et al. (2010).

Theorem 5. *Consider a CSA-problem with $K = 2$ sensors. Let $\mathcal{C}$ be a bounded set and $\gamma_i(x) + c_i = x'\theta_i$ for $i = 1, 2$ for all $x \in \mathcal{C}$. Assume $x'\theta_1, x'\theta_2 \in [0\ 1]$ for all $x \in \mathcal{C}$. Then, equivalent to a stochastic linear bandit.*

11.2 Proof of Theorem 5

Let $\{x_t, y_t\}_{t \geq 0}$ be an arbitrary sequence of context-label pairs. Consider a stochastic linear bandit where $D_t = \{0, x_t\}$ is a decision set in round t. From the previous section, we know that given a context x, action 1 is optimal if $\gamma_1(x) - \gamma_2(x) - c < 0$, otherwise action 2 is optimal. Let $\theta := \theta_1 - \theta_2$, then it boils down to check if $x'\theta - c < 0$ for each context $x \in \mathcal{C}$.

For all t, define $\epsilon_t = \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} - x_t'\theta$. Note that $\epsilon_t \in [0\ 1]$ for all t, and since sensors do not have memory, they are conditionally independent given past contexts. Thus, $\{\epsilon_t\}_{t>0}$ are conditionally R-sub-Gaussian for some finite R.

Given a policy π on a linear bandit we obtain next to play for the CSAP as follows: For each round t define $a_t \in \mathcal{C}$ and $r_t \in \{0, 1\}$ such that $a_t = 0$ and $r_t = 0$ if action 1 is played in that round, otherwise set $a_t = x_t$ and $r_t = \mathbf{1}\{\hat{y}_t^1 \neq \hat{y}_t^1\}$. Let $\mathcal{H}_t = \{(a_1, r_1) \cdots (a_{t-1}, r_{t-1})\}$ denote the past actions and corresponding rewards observed till time $t-1$. In round t, after observing context x_t, we transfer $((a_{t-1}, r_{t-1}), D_t)$, where $D_t = \{0, x_t\}$. If π outputs $0 \in D_t$ as the optimal choice, we play action 1, otherwise we play action 2.

Conversely, suppose π' denote a policy for the CSAP problem we select action to play from decision set $D_t = \{0, x_t\}$ as follows. For each round t define $a_t' \in 1, 2$ and $r_t' \in \mathcal{R}$ such that $a_t' = 1$ and $r_t' = \emptyset$ if 0 is played otherwise set $a_t' = 2$ and $r_t' = x_t'\theta^* + \epsilon_t$ if x_t is played. Let $\mathcal{H}_t' = \{(a_1', r_1') \cdots (a_{t-1}', r_{t-1}')\}$ denote the past actions and corresponding rewards observed till time $t-1$. In round t, after observing set D_t, we transfer $((a_{t-1}', r_{t-1}'), x_t)$ to policy π'. If π outputs action 1 as the optimal choice, we play action 0, otherwise we play x_t.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, pp. 2312–2320, 2011.

Agrawal, Rajeev, Teneketzis, Demosthenis, and Anantharam, Venkatachalam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Transaction on Automatic Control*, 34:258–267, 1989.

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

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

Bartók, G., Foster, D., Pál, D., Rakhlin, A., and Szepesvári, Cs. Partial monitoring – classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39:967–997, 2014.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceeding of Conference on Learning Theory, COLT*, Helsinki, Finland, July 2008.

Draper, B., Bins, J., and Baek, K. Adore: Adaptive object recognition. In *International Conference on Vision Systems*, pp. 522–537, 1999.

Greiner, R., Grove, A., and Roth, D. Learning cost-sensitive active classifiers. *Artificial Intelligence*, 139:137–174, 2002.

Isukapalli, R. and Greiner, R. Efficient interpretation policies. In *International Joint Conference on Artificial Intelligence*, pp. 1381–1387, 2001.

Kapoor, A. and Greiner, R. Learning and classifying under hard budgets. In *ECML*, 2005.

Li, L., Wei, C., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceeding of International Word Wide Web conference, WWW*, NC, USA, April 2010.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *NIPS*, 2011.

Nan, F., Wang, J., and Saligrama, V. Feature-budgeted random forest. In *Proceeding of Conference on Neural Information Processing Systems, NIPS*, 2015.

Póczos, B., Abbasi-Yadkori, Y., Szepesvári, Cs., Greiner, R., and Sturtevant, N. Learning when to stop thinking and do something! In *ICML*, pp. 825–832, 2009.

Rusmevichientong, Paat and Tsitsiklis, John N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

Trapeznikov, K. and Saligrama, V. Supervised sequential classification under budget constraints. In *AISTATS*, pp. 235–242, 2013.

Trapeznikov, K., Saligrama, V., and Castanon, D. A. Multi-stage classifier design. *Machine Learning*, 39:1–24, 2014.

Wang, J., Trapeznikov, K., and Saligrama, V. Directed acyclic graph for resource constrained prediction. In *Proceeding of Conference on Neural Information Processing Systems, NIPS*, 2015.

Wu, Y., György, A., and Szepesvári, Cs. Online learning with gaussian payoffs and side observations. In *NIPS*, pp. 1360–1368, September 2015.

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