Learning without Feedback

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We propose a sensor acquisition problem (SAP) wherein sensors (and sensing tests) are organized into a cascaded architecture of increasingly informative and expensive tests and the goal is to choose a test with the optimal cost-accuracy tradeoff for a given instance. We consider the case where we obtain no feedback in terms of rewards for our chosen actions apart from test observations. Absence of feedback raises fundamentally new challenges since one cannot infer potentially optimal tests. We pose the problem in terms of competitive optimality with the goal of minimizing cumulative regret against optimally chosen actions in hindsight. In this context we introduce the notion of weak dominance and show that it is necessary and sufficient for realizing sub-linear regret. Weak dominance on a cascade supposes that a child node in the cascade has higher accuracy when its parent node makes correct predictions. When weak dominance holds we show that we can reduce SAP to a corresponding multi-armed bandit problem with side observations. Empirically we verify that weak dominance holds for many datasets.

1 Introduction

In many classification systems such as medical diagnosis and homeland security, sequential decisions are often warranted. For each instance, an initial diagnostic test is conducted and based on its results further tests maybe conducted. Tests have varying costs for acquisition, and these costs account for delay, throughput or monetary value [1]. Apart from these natural scenarios the problem also arises in the context of wireless communication systems, where a cascade of error-correcting decoders of increasing block lengths are designed to overcome channel noise.

Our goal is essentially a sensor acquisition problem (SAP), namely, to acquire the tests/sensors that achieves the optimal cost-accuracy tradeoff for that instance. We assume that the sensors/tests are organized into a diagnostic cascade architecture, where the ordering is based on costs/informativity of tests. Each stage in the cascade outputs a prediction of the underlying state of the instance (disease or disease-free, threat or no-threat etc.). We suppose that the classifiers (or predictors) corresponding to each node are part of the system and produce labeled outputs. This is often the case in diagnostic systems where a test ordering is a priori known and a report is produced by a human being or an automated mechanism corresponding to different sensor measurements. Thus our task in this paper is primarily to learn a decision rule to identify the collection of tests required for an instance.

Our problem can be framed as a version of a multi-armed bandit problem. Each arm of the bandit corresponds to a unique path from root to a node where the observation is a vector of outputs from

[1] As described in **?** security systems utilize a suite of sensors/tests such as X-rays, millimeter wave imagers (expensive & low-throughput), magnetometers, video, IR imagers and human search. Security systems must maintain a throughput constraint in order to keep pace with arriving traffic. In clinical diagnosis, doctors in the context of breast cancer diagnosis utilize tests such as genetic markers, imaging (CT, ultrasound, elastography) and biopsy. Imaging sensors are scored by humans. The different sensing modalities have diverse costs, in terms of health risks (radiation exposure) and monetary expense.

tests acquired along that path. Nevertheless, our problem is unconventional. Unlike a conventional bandit problem, where feedback (reward) is observed corresponding to each action, we do not get feedback of how well our action performed (either noisy or noiseless)[2].

Absence of reward information associated with chosen actions is fundamentally challenging since we cannot infer potential optimal actions. We pose the problem in terms of competitive optimality. In particular we consider a competitor who has the benefit of hindsight and can choose an optimal collection of tests for all the examples. Our goal is to choose an action for each instance so that the cummulative regret with respect to the competitor is sub-linear (and optimal).

We first provide negative results for the problem. We introduce the notion of weak dominance on tests. We show that weak dominance is fundamental, i.e., regardless of the algorithm, if this condition is not satisfied, we are left with a linear regret. On the other hand we develop UCB style algorithms that show that we can realize optimal regret (sub-linear regret) guarantees when the condition is satisfied. This leads to a sharp necessary and sufficient condition for learning under no feedback.

The weak dominance condition amounts to a stochastic ordering of the tests on the diagnostic cascade. Conceptually, the weak dominance condition says that the child node tends to be relatively more accurate when the parent is correct. Under weak dominance we show that the learner can partially infer losses of the stages. In particular, we reduce the SAP problem to a stochastic multi-armed bandit with side observations, where bandit arms are identified by the nodes of the cascade. The payoff of an arm is given by loss from the corresponding stage, and side observation structure is defined by the feedback graph induced by the cascade. Empirically we verify that weak dominance condition naturally holds for several datasets including breast-cancer and diabetes datasets. A stronger dominance condition is also shown to hold by design, namely, for error-correcting code cascades in the context of communication systems.

2 Sensor Acquisition Problem

The learner has access to $K \geq 2$ sensors that are ordered in terms of their prediction efficiency. Specifically, we consider that the sensors form a cascade (order in which the sensors are selected is predetermined) and in each round the learner can sequentially select a subset of sensors in the cascade and stop at any depth.

Let $\{Z_t, Y_t\}_{t>0}$ denote a sequence generated according to an unknown distribution. $Z_t \in \mathcal{C} \subset \mathcal{R}^d$, where $\mathcal{C}$ is a compact set, denotes a feature vector/context at time t and $Y_t \in \{0, 1\}$ its binary label. We denote output/prediction of the i^{th} sensor as $\hat{Y}_t^i$ when its input is Z_t. The set of actions available to the learner is $\mathcal{A} = \{1, \ldots, K\}$, where the action $k \in \mathcal{A}$ indicates acquiring predictions from sensors $1, \ldots, k$ and classifying using the prediction $\hat{Y}_t^k$.

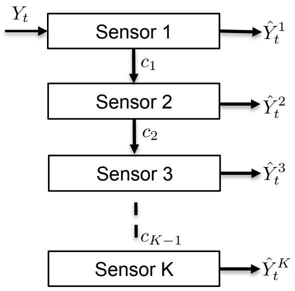

Figure 1: Cascade of sensors

The prediction error rate of the i^{th} sensor is denoted as $\gamma_i := \Pr\{Y_t \neq \hat{Y}_t^k\}$. In this section we assume that the error rate does not depend on the context and postpone the treatment with contextual information to Section 6. Further, the sensors are arranged such that the prediction error rate improves with depth in the cascade, i.e., $\gamma_{k-1} \geq \gamma_k$ for all $k > 2$. However, the learner incurs an extra cost of $c_k \geq 0$ to acquire output of sensor k after acquiring output of sensor $k-1$. The sensor cascade is depicted in the adjacent figure.

Let $H_t(k)$ denote the feedback observed in round t from action k. Since we observe predictions of all the first k senors by playing action k, we get $H_t(k) = \{\hat{Y}_t^1, \ldots, \hat{Y}_t^k\}$. The loss incurred in each round is defined in terms of the prediction error and the total cost involved. When the learner selects action k, loss is the prediction error of sensor k plus sum of the costs incurred

[2] This problem naturally arises in the surveillance and medical domains. We can perform a battery of tests on an individual in an airport but can never be sure whether or not he/she poses a threat.

along the path $(c_1, \ldots, c_k)$. Let $L_t : \mathcal{A} \rightarrow \mathcal{R}_+$ denote the loss function in round t. Then,

$$L_t(k) = \mathbf{1}_{\{\hat{Y}_t^k \neq Y_t\}} + \sum_{j=1}^{k} c_j. \tag{1}$$

We refer to the above setup as Sensor Acquisition Problem (SAP) and denote it as $\psi = (K, \mathcal{A}, (\gamma_i, c_{i-1})_{i\in[K]})^3$. A policy $\pi^\psi = (\pi_1^\psi, \pi_2^\psi, \cdots)$ on ψ, where $\pi_t^\psi : \mathcal{H}_{t-1} \rightarrow \mathcal{A}$, gives action selected in each round using history $\mathcal{H}_{t-1}$ that consists of all actions and corresponding feedback observed before t. Let Π^ψ denote set of policies on ψ. For any $\pi \in \Pi^\psi$, we compare its performance with respect to the optimal policy (single best action in hindsight) and define its expected regret as follows

$$R_T^\psi(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} L_t(a_t)\right] - \min_{k\in A} \mathbb{E}\left[\sum_{t=1}^{T} L_t(k)\right], \tag{2}$$

where a_t denotes the policy selected by π_t in round t. The goal of the learner is to learn a policy that minimizes the expected total loss, or, equivalently, to minimize the expected regret, i.e.,

$$\pi^* = \arg\min_{\pi\in\Pi^\psi} R_T^\psi(\pi). \tag{3}$$

Optimal action in hindsight: For any t, we have

$$\mathbb{E}[L_t(k)] = \Pr\{Y_t \neq \hat{Y}_t^k\} + \sum_{j=1}^{k} c_j = \gamma_k + \sum_{j=1}^{k} c_j. \tag{4}$$

Let $k^* = \arg\min_{k\in\mathcal{A}} \gamma_k + \sum_{i<k} c_i$. Then the optimal policy is to play action k^* in each round. If an action i is played in any round then it adds $\Delta_k := \gamma_k + \sum_{i<k} c_i - (\gamma_{k^*} + \sum_{i<k^*} c_i)$ to the expected regret. Let I_t denote the action selected in round t and $N_k^\psi(s)$ denote the number of times action k is selected till time s, i.e., $N_k^\psi(s) = \sum_{t=1}^{s} \mathbf{1}_{\{I_t=k\}}$. Then the expected regret can be expressed as

$$R_T^\psi(\pi) \quad = \quad \sum_{k\in\mathcal{A}} \mathbb{E}[N_k^\psi(T)]\Delta_k. \tag{5}$$

3 When is SAP Learnable?

In the SA-Problem feedback $H_t(\cdot)$ does not reveal any information about the true label Y_t in any round t. Hence the loss values are not known, and we are in a hopeless situation where linear regret is unavoidable. In this section we explore conditions that lead to policies that are Hannan consistent **?**, i.e, a policy $\pi \in \Pi^\psi$ such that $R_T^\psi(\pi)/T \rightarrow 0$.

Let us consider $K = 2$ sensors and start with a simple condition that if sensor 1 predicts the label 1 correctly, then sensor 2 also predicts it correctly[4], i.e.,

$$Y_t = 1 \text{ and } \hat{Y}_t^1 = 1 \implies \hat{Y}_t^2 = 1. \tag{6}$$

To fix ideas, we enumerate all the 8 possible tuples $(Y, \hat{Y}^1, \hat{Y}^2)$ as shown in Table 3, and write probability of the ith tuple $i = 1, 2, \cdots 8$ as p_{i-1}. From Table 3, we have $\gamma_1 = p_2 + p_3 + p_4 + p_5$ and $\gamma_2 = p_1 + p_3 + p_4 + p_6$, thus

$$\gamma_1 - \gamma_2 = p_2 + p_5 - p_1 - p_6. \tag{7}$$

[3] Note that $k \in \mathcal{A}$ implies that action k selects all sensors $1, 2, \cdots, k$, not just sensor k. We set $c_0 = 0$

[4] Suppose we interpret label 1 as 'threat', the condition implies that if sensor 1 detects threat correctly, the better sensor 2 also detects it.

Y	$\hat{Y}_t^1$	$\hat{Y}_t^2$	$\Pr(Y,\hat{Y}^1,\hat{Y}^2)$
0	0	0	p_0
0	0	1	p_1
0	1	0	p_2
0	1	1	p_3
1	0	0	p_4
1	0	1	p_5
1	1	0	p_6
1	1	1	p_7

$$\Pr(\hat{Y}^1,\hat{Y}^2) = \begin{cases} p_1 + p_5 \text{ if } (\hat{Y}^1,\hat{Y}^2) = (0,1) \\ p_2 + p_6 \text{ if } (\hat{Y}^1,\hat{Y}^2) = (1,0) \\ p_0 + p_4 \text{ if } (\hat{Y}^1,\hat{Y}^2) = (0,0) \\ p_3 + p_7 \text{ if } (\hat{Y}^1,\hat{Y}^2) = (1,1) \end{cases} \tag{8}$$

From (4), action 1 is optimal if $\gamma_1 - \gamma_2 \le c$, otherwise action 2 is optimal. If a policy learns the difference $\gamma_1 - \gamma_2$, it can play the optimal arm and it is Hannan consistent. Note that only sensor output $(\hat{Y}^1, \hat{Y}^2)$ are observed and not the true label Y. Hence only values of marginal probabilities $\Pr(\hat{Y}^1, \hat{Y}^2)$ as given in (8) can be used to learn the difference $\gamma_1 - \gamma_2$. The following example demonstrate that just knowing the values of marginals is not enough to decide which action is optimal.

Set $c = 0.35$ and consider the following two case: 1) $p_2 = 1/2, p_1 = 1/4 - 1/40, p_5 = 1/4 + 1/40$ and 2) $p_2 = 1/2, p_1 = 1/4 - 3/40, p_5 = 1/4 + 3/40$. From condition (6) we have $p_6 = 0$. Also, set $p_0 = p_4 = p_3 = p_7 = 0$ in both the cases. We get $\gamma_1 - \gamma_2 = 0.3$ in the first case, hence action 1 is optimal. Where as $\gamma_1 - \gamma_2 = 0.4$ in the second case, hence actions 2 is optimal. However, for both the cases the marginals $\Pr(\hat{Y}^1, \hat{Y}^1)$ are the same for all pairs $(\hat{Y}^1, \hat{Y}^1)$. Since we only observer the pairs $(\hat{Y}^1, \hat{Y}^1)$, one cannot hope to distinguish the cases and linear regret is unavoidable.

Next, assume that if sensor 0 predicts the label 0 correctly, then sensor 2 also predicts it correctly, i.e.,

$$Y_t = 0 \text{ and } \hat{Y}_t^1 = 0 \implies \hat{Y}_t^2 = 0. \tag{9}$$

We can argue similar to the previous example that under this conditions one cannot expect better than linear regret. Now assume that both (6) and (9) hold. Then, $p_2 = p_6 = 0$ and we get $\gamma_1 - \gamma_2 = p_5 - p_1$. Since $p_5 = \Pr(0,1)$ and $p_1 = \Pr(1,0)$, we can learn their values by observing $(0,1)$ and $(1,0)$ patterns and thus hope for a Hannan consistent policy. In the following we assume that (6) and (9) hold and refer to it as dominance condition. For the case of $K > 2$ sensors, it is given as follows:

Assumption 1 (Dominance Condition) *If sensor i predicts correctly, all the sensors in the subsequent stages of the cascade also predict correctly, i.e.,*

$$\hat{Y}_t^i = Y_t \rightarrow \hat{Y}_t^j \quad \forall j > i \geq 1 \tag{10}$$

In the following we establish that under the dominance condition efficient algorithms for a SAP problem can be derived from algorithms on a suitable stochastic multi-armed bandit problem. We first recall the stochastic multi-armed bandit setting and the relevant results.

4 Background on Stochastic Multi-armed Bandits

A stochastic multi-armed bandit (MAB), denoted as $\phi := (K, (\nu_k)_{1 \le k \le K})$, is a sequential learning problem where number of arms K is known and each arm $i \in [K]$ gives rewards drawn according to an unknown distribution ν_k. Let $X_{i,n}$ denote the random reward from arm i in its nth play. For each arm $i \in [K]$, $\{X_{i,t} : t > 0\}$ are independently and identically (i.i.d) distributed and for all $t > 0$, $\{X_{i,t}, i \in [K]\}$ are independent. We note that in the standard MAB setting the learner observes only reward from the selected arm in each round and no information from the other arms is revealed. A policy is any allocation strategy that maps the past history into an arm in each round, and let Π^ϕ denote a set of polices on ϕ. If the learner knows $\{\nu_k\}_{k \in [K]}$, then the optimal policy is to play the arm with highest mean. For any policy $\pi \in \Pi^\phi$, its performance is measured with respect to the optimal policy and is defined in terms of expected cumulative regret (or simply regret) as follows:

Let π selects arm i_t in round t. After T rounds, its regret is

$$R_T^{\phi}(\pi) = T\mu_{i^*} - \sum_{t=1}^{T} \mu_{i_t}, \tag{11}$$

where $\mu_i = \mathbb{E}[X_{i,n}]$ denotes mean of distribution ν_i for all $i \in [K]$ and $i^* = \arg\max_{i\in[K]} \mu_i$. Let $N_i^{\phi}(t) = \sum_{s=}^{t} \mathbf{1}\{i_s = i\}$ denote the number of pulls of arm i till time t. Then, the Regret of policy π can be expressed

$$R_T^{\phi}(\pi) = \sum_{i=1}^{K} (\mu_{i^*} - \mu_i)\mathbb{E}[N_i^{\phi}(T)].$$

The goal of the learner is to learn a policy that minimizes the regret.

MAB problems have been extensively studied in the literature. The seminal paper of Lai & Robbins **?** showed that for any consistent policy (that plays sub-optimal arms only sup-polynomially many times in the time horizon) the regret grows logarithmically in time horizon. Specifically, for a class of parametric reward distributions, they showed that regret of any consistent policy satisfies

$$\liminf_{n\to\infty} \frac{R_T^{\phi}(\pi)}{\log T} \geq \sum_{i\neq i^*} \frac{\mu_{i^*} - \mu_i}{D(\mu_{i^*}||\mu_i)}, \tag{12}$$

where $D(p||q)$ is the KL-divergence of $p, q \in [0\ 1]$. Further, the authors in **?** provided an upper confidence bound (UCB) based policy that asymptotically achieves the lower bound for a class of parmetric reward distributions.

Auer et, al. **?** proposed an anytime policy named UCB1, that is based on the UCB strategy and showed that it is optimal on any MAB with bounded rewards. Specifically, they showed that regret of UCB1 for any $T > K$ is upper bound as

$$R_T^{\phi}(\text{UCB1}) \leq \sum_{i\neq i^*} \frac{8\log n}{\mu_{i^*} - \mu_i} + (\pi^2/3 + 1)(\mu_{i^*} - \mu_i). \tag{13}$$

Thus demonstrating the optimality of UCB1. Since the work of Auer et. al. several works have proposed improvised UCB based policies like, KL-UCB **?**, MOSS **?**.

4.1 MAB With Side Information

In many applications playing an arm reveals information about the other arms which can be exploited to improve learning performance. Let $\mathcal{N}_i$ denote the set of arms such that playing arm i reveals rewards of all arms $j \in \mathcal{N}_i$. We refer to $\mathcal{N}_i$ as neighborhood of i and the graph induced by the neighborhood sets as side-information graph. Given a set of neighborhood $\{\mathcal{N}_i, i \in [K]\}$, let $\phi_G := (K, (\nu_k)_{1\leq k\leq K}, G)$ denote a MAB with side-information graph $G = (V, E)$, where $|V| = K$ and $(i, j) \in E$ if $j \in \mathcal{N}_i$. The side-observation graph is known to the learner and remains fixed during the play.

Let Π^{ϕ_G} denote the set of all policies on ϕ_G that map the past history (including the side-observations) to an action in each round. For any policy $\pi \in \Pi^{\phi_G}$, we denote the regret over a period T as $R_T^{\phi_G}(\pi)$ and is given by (11). Note that, in each round, only reward from the arm played contribute to the regret and not that from the side-observations. In **?** the authors extended the lower bound in (12) to incorporate the effect of side-observations. Specifically, they establish that any policy $\pi \in \Pi^{\phi_G}$ where side observation graph is such that $i \in \mathcal{N}_i$ for all $i \in [K]$ satisfies **?**

$$\liminf_{T\to\infty} R_T^{\phi_G}(\pi)/\log T \geq \eta(G) \tag{14}$$

where $\eta(G)$ is the optimal value of the following linear program

$$\begin{aligned} LP1:\ &\min_{\{w_i\}} \sum_{i\in[K]} (\mu_{i^*} - \mu_i)w_i \\ &\text{subjected to} \sum_{j\in\mathcal{N}_i} w_i \geq 1/D(\mu_i||\mu_{i^*}) \text{ for all } i \in [K] \\ &w_i \geq 0 \text{ for all } i \in [K] \end{aligned} \tag{15}$$

Definition 1 (Domination number ?) *Given a graph $G = (V, E)$, a subset $W \subset V$ is a dominant set if for each $v \in V$ there exists $u \in W$ such that $(u, v) \in E$. The size of the smallest dominant set is called weak domination number and is denoted as $\xi(G)$.*

The authors in **?** gave an UCB based strategy, named UCB-LP, that exploits the side-observations and explore arms at a rate in proportion to the size of their neighborhood. UCB-LP plays arms in proportions to the values $\{z_i^*, i \in [K]\}$ computed from the following linear programmer which is a relaxation of linear programme $LP1$.

$$\begin{aligned}
&LP2 : \min_{\{z_i\}} \sum_{i \in [K]} z_i \\
&\text{subjected to} \sum_{j \in \mathcal{N}_i} z_i \geq 1 \text{ for all } i \in [K] \\
&z_i \geq 0 \text{ for all } i \in [K]
\end{aligned} \tag{16}$$

The regret of UCB-LP is upper bounded by

$$\mathcal{O}\left(\sum_{i \in [K]} z_i^* \log T \right) + \mathcal{O}(K\delta), \tag{17}$$

where $\delta = \max_{i \in [K]} |\mathcal{K}_i|$ and $\{z_i^*\}$ are the optimal values of $LP2$. Since any dominating set is a feasible solution of $LP2$, we get $\sum_{i \in [K]} z_i^* \leq \xi(G)$, and the regret of UCB-LP is $\mathcal{O}(\xi(G) \log T)$.

4.2 Special case: 1-armed bandit

In the MAB problem when all the arms have a fixed reward except for one, we get 1-armed bandit. The learner knows the arms that give fixed reward the goal is to identify the quality of the arm that gives stochastic reward as fast as possible. A straightforward modification of UCB1 achieves optimal regret of $\Theta(\log T)$.

5 Regret Equivalence

In this section we establish that under the dominance condition SAP 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.

Definition 2 (Regret Equivalence) *Consider a SAP problem $\psi := (K, \mathcal{A}, (\gamma_i, c_{i-1})_{i \in [K]})$ and a bandit problem with $\phi_G := (N, (\nu_i)_{i \in [N]}, G)$ side-information graph G. We say that ψ is regret-equivalent to ϕ_G if given a policy π for problem ψ, one can come up with a policy π' that uses π, such that the regret of π' on any instance of ϕ_G is the same as the regret of π on some corresponding instance of ψ, and vice versa.*

In the following we first consider the SAP with 2 sensors and then the general case with more than 2 sensors. The 2 sensors case helps to draw comparison with the well studied apple tasting problem and understand role of the dominance condition.

5.1 SAP with two sensors

In the SAP with only two actions, the feedback from action $i = 1$ reveals no information about the loss incurred in that round. However feedback after action $i = 2$ reveals (partial) information about the loss of both actions. Suppose feedback is such that predictions of the sensors disagree, i.e., $\hat{Y}_t^1 \neq \hat{Y}_t^2$ after action 2. The dominance condition then implies that the only possible events are $\hat{Y}_t^1 \neq Y_t$ and $\hat{Y}_t^2 = Y_t$. I.e., the true label is that predicted by sensor-2, hence loss incurred is just c (prediction loss is zero). Suppose predictions of the sensors agree, i.e., $\hat{Y}_t^1 = \hat{Y}_t^2$, then the dominance condition implies that either predictions of both are correct or both are incorrect. Though the true loss is not known in this case, the learner can infer some useful knowledge: in round t, if prediction of both the sensors agree, then the difference in losses of the actions is $L_t(2) - L_t(1) = c > 0$.

And if predictions of the sensors disagree, then dominance assumption implies that $L_t(1) = 1$ and $L_t(2) = c$ or $L_t(2) - L_t(1) = c - 1 < 0$. Thus, each time learner plays action 2, he gets to know whether or not he was better off by selecting the other action. This setup sounds similar to the standard apple tasting problem **?**], but differs in terms of the information structure when action 2 is played.

Apple tasting problem: In the apple tasting problem, a learner gets a sequence of apples and some of them can be rotten. In each round, the learner can either accept or reject an apple. If an apple is accepted, the learner tastes it and incurs a penalty if it is rotten. If apple is rejectsed, he still incurs the penalty if it is rotten, but do not get to observe its quality. The goal of the learner is to taste more good apples. The SAP setting is a more general version than the apple tasting problem–in any round, actions 1 reveals no loss values. Action 2 reveals only partial information about the losses, but not the exact losses as in the apple tasting problem. However, we next show that the partial information is enough to achieve optimal performance.

Theorem 1 *Assume dominance condition (10) holds. Then SAP ψ with $K = 2$ is regret-equivalent to a stochastic 1-armed bandit.*

The following corollary follow immediately from the regret equivalence.

Proposition 1 (SAP regret lower bound) *Let π be any policy on SAT with 2 sensors such that it pulls the suboptimal arm only sub polynomial many times, i.e., $\mathbb{E}[N_i(T)] = o(T^a)$ for all $a > 0$ and $i \neq i^*$. Then,*

$$\liminf_{T\to\infty} R_T^{\psi}(\pi)/\log T \geq \frac{|\gamma_1 - \gamma_2 - c|}{D(\hat{\gamma}, \gamma^*)} \textit{ where } \gamma^* = \min\{\gamma_1, \gamma_2 + c\}, \hat{\gamma} = \max\{\gamma_1, \gamma_2 + c\} \quad (18)$$

and $D(\hat{\gamma}, \gamma^*)$ is the KL-divergence between $\hat{\gamma}$ and γ^*.

Proposition 2 (SAP regret upper bound) *Let π' denote a policy on a 1-armed stochastic bandit where one arm has mean $\gamma_1 - \gamma_2$ and the other gives fixed reward c. Then, the regret of a policy $g(\pi)$ for the SAT problem obtained according the mapping (27) is upper bounded as*

$$R_T^{\psi}(g(\pi)) \leq \frac{6 \log T}{|\gamma_1 - \gamma_2 - c|} + |\gamma_1 - \gamma_2 - c|(1 + \pi^2/3) \textit{ when } \pi' = \textit{UCB1}. \quad (19)$$

$$R_T^{\psi}(g(\pi)) \leq \frac{|\gamma_1 - \gamma_2 - c| \log T}{D(\hat{\gamma}, \gamma^*)} + \mathcal{O}(\sqrt{\log T}) \textit{ when } \pi' = \textit{KL-UCB}. \quad (20)$$

5.2 SAP with more than two actions

In the SAP with two sensors, only action 2 provides information about the losses. In the case with $K > 2$ sensors, by playing an action k, we can obtain information about the losses of all sensors $l < k$ by recursively applying the dominance condition between pair of sensors. Further, any information provided by action $k > 2$ is contained in that provided by all actions $k' \geq k$– if action k is played in round t, then we observe predictions $\{\hat{Y}_t^1, \hat{Y}_t^2, \cdots, \hat{Y}_t^i\}$ which includes the observed predictions of all actions $k' \leq i$. This side-observation can be represented by a directed graph $G^S = (V, E)$, where $|V| = K$ and $E = \{(i, j) : i1 < i \leq j \leq K\}$. Note that G^S has self loops for all nodes except for node 1. The nodes in G^S represents actions of the SAP and an edge $(i, j) \in E$ implies that actions i provides information about action j. The side-observation graph for the SAP is shown in Figure (2).

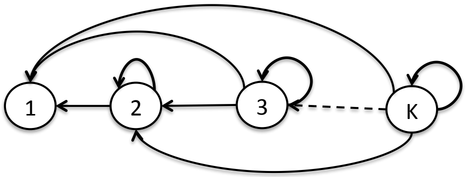

Figure 2: Side observation graph G^S

Theorem 2 *Let the dominance condition (10) holds. Then SAP ψ with $K \geq 2$ is regret equivalent to a MAB with side-observation graph G^S.*

Remark 1 *Note that the some of mean values $\{\gamma_1 - \gamma_i - \sum_{j \leq i} c_j\}$ need not be positive. Since the stochastic bandit algorithms assume that reward lie in the interval $[0, 1]$, we can ensure positive* *means by setting distribution ν_k, to have mean $\gamma_1 - \gamma_i - \sum_{j<i} c_j + \sum_{k \leq K-1} c_k$. Note that mean of each arm is shifted by the same amount, which does not change the regret value. This recovers the SAP with $K = 2$ actions and Theorem 1 holds.*

Proposition 3 (SAP regret lower bound) *Let π be any policy on SAT with 2 sensors such that it pulls the suboptimal arm only sub polynomial many times, i.e., $\mathbb{E}[N_i^{\psi}(T)] = o(T^a)$ for all $a > 0$ and $i \neq i^*$. Then,*

$$\liminf_{T \to \infty} R_T^{\psi}(\pi)/\log T \geq \kappa \textit{ where} \tag{21}$$

$$\begin{aligned} &\kappa = \min_{\{w_i\}} \sum_{i \in [K]} (\mu_{i^*} - \mu_i) w_i \\ &\textit{subjected to} \sum_{ji} w_i \geq 1/D(\mu_i + \sum_{j<i} c_j || \mu_{i^*}) \textit{ for all } i \in [K] \\ &w_i \geq 0 \textit{ for all } i \in [K] \end{aligned} \tag{22}$$

Proposition 4 (K-SAT regret upper bound) *Let π' denote a policy on a K-armed stochastic bandit where mean of arm $i > 1$ is $\gamma_1 - \gamma_i - \sum_{j<i} c_j$ and arm 1 has a fixed reward of value zero, and the side-observation graph is G^S. Then, the regret of a policy $g_1(\pi)$ for the SAT problem obtained from mapping (35) is upper bounded as*

$$R_T^{\psi}(g(\pi)) \leq \mathcal{O}(\xi(G^S) \log T + K^2) \tag{23}$$

when $\pi' = UCB - LP$ **?**.

6 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:

$$R_T^{\phi_c}(\pi) = \sum_{t=1}^{T} \mathbb{E}\left[L_t(\pi(x_t)|x_t)\right] - \sum_{t=1}^{T} \min_{a \in \mathcal{A}} \mathbb{E}\left[L_t(a|x_t)\right]. \tag{24}$$

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.

Note: c is a fixed cost and does not depend on context. We are assuming that error rate of sensor 2 offset by c is a linear quantity. Another possibility is, we can assume that there exists a $x_0 \in \mathcal{C}$ such that $c = x_0'\theta_2$ and we have oracle access to x_0. Then all the arguments hold.

7 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

$$R_T^L(\pi) = \sum (x_t^*)'\theta^* - \sum x_t'\theta^*. \tag{25}$$

The above setting, where actions sets can change in every round, is introduced in**?** and is a more general setting than that studied in **??** where decision set is fixed. Further, the above setting also specializes the contextual bandit studied in **?**. The authors in **?** 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 $ConfidenceBall_2$ **?**, $UncertainityEllipsoid$ **?** and $LinUCB$ **?**.

Theorem 3 *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.*

A Proof of Theorem 1

Consider a 1-armed stochastic bandit problem where arm with constant reward has value c and the arm that gives stochastic reward has mean value $\gamma_1 - \gamma_2$. Given an arbitrary policy $\pi = (\pi_1, \pi_2, \cdots \pi_t)$ for the SAP, we obtain a policy for the bandit problem 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 $t-2$. Let a_{t-1} denote the action selected by the bandit policy in round $t-1$ and r_{t-1} the observed reward. Then, the next action a_t is obtained as follows:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{fixed rewad arm} \\ \pi_t(H_{t-1} \cup \{2, r_{t-1}\}) \text{ if } a_{t-1} = \text{stochastic arm} \end{cases} \tag{26}$$

Conversely, let $\pi' = \{\pi'_1, \pi'_2, \cdots\}$ denote an arbitrary policy for the 1-armed bandit problem. we obtain a policy for 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 $t-1$. 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:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1, c\}) \text{ if } a'_{t-1} = \text{action 1} \\ \pi'_t(H'_{t-1} \cup \{2, \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\}\}) \text{ if } a_{t-1} = \text{actions 2.} \end{cases} \tag{27}$$

We next show that regret of π on the SAP is same as that of derived policy on the 1-armed bandit, and regret of π' on the 1-armed bandit is same as regret of the derived policy on SAP. We first argue that any policy on the SAP problem with 2 actions needs the information if whether the predictions of sensors match or not whenever action 2 is played. The following observation is straightforward.

Lemma 1 *Let dominance condition holds. Then,* $\Pr\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} = \gamma_1 - \gamma_2$.

$$\begin{aligned}
\Pr\{\hat{Y}_t^1 \neq \hat{Y}_t^1\} &= \Pr\{\hat{Y}_t^1 = Y_t, \hat{Y}_t^2 \neq Y_t\} + \Pr\{\hat{Y}_t^2 = Y_t, \hat{Y}_t^1 \neq Y_t\} && (28)\\
&= \Pr\{\hat{Y}_t^2 = Y_t, \hat{Y}_t^1 \neq Y_t\} \text{ from assumption (10)} && (29)\\
&= \Pr\{\hat{Y}_t^1 \neq y_t\} \Pr\{\hat{Y}_t^2 = Y_t | \hat{Y}_t^1 \neq Y_t\} && (30)\\
&= \Pr\{\hat{Y}_t^1 \neq Y_t\} \left(1 - \Pr\{\hat{Y}_t^2 \neq Y_t | \hat{Y}_t^1 \neq Y_t\}\right) && (31)\\
&= \Pr\{\hat{Y}_t^1 \neq Y_t\} \left(1 - \frac{\Pr\{\hat{Y}_t^2 \neq Y_t, \hat{Y}_t^1 \neq Y_t\}}{\Pr\{\hat{Y}_t^1 \neq Y_t\}}\right) && (32)\\
&= \Pr\{\hat{Y}_t^1 \neq Y_t\} - \Pr\{\hat{Y}_t^2 \neq Y_t\} \text{ by contrapositve of (10)} && (33)
\end{aligned}$$

From Lemma 1, mean of the observations $Z_t := \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\}$ from action 2 in the SAP is a sufficient statistics to identify the optimal arm. Thus, any SAP only needs to know Z_t in each round, and Z_t are i.i.d with mean $\gamma_1 - \gamma_2$. Our mapping of policies is such that any poilcy for SAP (1-armed bandits) and the derived policy on the 1-armed bandit (SAP) play the sub-optimal arm same number of times. For the sake of simplicity assume that action 1 is optimal for SAP ($\gamma_1 > \gamma_2 + c$) and let a policy π on SAP plays it $N_1(T)$ number if times. Then, we have

$$R_T^{\psi}(\pi) = \Delta_i \mathbb{E}[N_1^{\psi}(T)] = (\gamma_1 - \gamma_2 - c)\mathbb{E}[N_1(T)]$$

Let $f(\pi)$ denote the policy for the 1-armed bandit obtained using the mapping (26). Now, for the 1-armed bandit, where the arm with stochastic rewards is optimal, we have

$$R_T^{\phi}(f(\pi)) = (\mu_2 - \mu_1)\mathbb{E}[N_1(T)] = (\gamma_1 - \gamma_2 - c)\mathbb{E}[N_1^{\phi}(T)]$$

Thus the regret of π on the SAP problem and that of $f(\pi)$ on the 1-armed bandit are the same. We can argue similarly for the other case.

B Proof of Theorem 2

Consider a K-armed stochastic bandit problem where rewards 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 excluding that from the first arm. Then, the next action a_t is obtained as follows:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{arm 1} \\ \pi_t(H_{t-1} \cup \{i, r_{t-1} \cup o_{t-1}\}) \text{ if } a_{t-1} = \text{arm i} \end{cases} \tag{34}$$

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:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1, 0\}) \text{ if } a'_{t-1} = \text{action 1} \\ \pi'_t(H'_{t-1} \cup \{i, \mathbf{1}\{\hat{Y}^1_t \neq \hat{Y}^2_t\} \cdots \mathbf{1}\{\hat{Y}^1_t \neq \hat{Y}^i_t\}\}) \text{ if } a_{t-1} = \text{action i.} \end{cases} \tag{35}$$

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 information 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 (34). The regret of policy π that plays actions i, $N_i(T)$ times is given by

$$R_T^{\psi}(\pi) \;\;=\;\; \sum_{i=1}^{K} \left[\left(\gamma_i + \sum_{j<i} c_j \right) - \left(\gamma_{i^*} + \sum_{j<i^*} c_j \right) \right] \mathbb{E}[N_i^{\psi}(T)] \tag{36}$$

(37)

Now, regret of regret policy $f_1(\pi)$ on the K-armed bandit problem with side information graph G^S

$$R_T^{\phi_G}(f_1(\pi)) = \sum_{i=1}^{K} \left[\left(\gamma_1 - \gamma_{i^*} - \sum_{j<i^*} c_j \right) - \left(\gamma_1 - \gamma_i - \sum_{j<i} c_j \right) \right] \mathbb{E}[N_i^{\phi_G}(T)] \tag{38}$$

which is same as $R_T^{\phi}(\pi)$. This concludes the proofs.

C Proof of Theorem 3

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 $\epsilon_t = \mathbf{1}\{\hat{Y}^1_t \neq \hat{Y}^2_t\} - x'_t\theta$. 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 $r_t = \mathbf{1}\{\hat{y}^1_t \neq \hat{y}^1_t\}$. 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.