Unsupervised Sequential Sensor Acquisition

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 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 problems 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 result further tests maybe conducted in a fixed ordering, where ordering of the tests is often based on their cost. Given the outcomes of the test results at some stage, a classifier, which is part of the diagnostic architecture, produces a prediction of the unknown label of the instance to be classified. Apart from the above-mentioned natural scenarios, the problem also arises in human engineered systems, such as in the context of wireless communication systems, where a cascade of error-correcting decoders of increasing block lengths are designed to overcome channel noise, but using a larger block lengths incurs extra communication cost. In all these examples, tests have varying costs for acquisition, accounting for delay, throughput or monetary value.[1] Was the probability of error known for each classifier that uses an initial segment of the tests, a decision maker could optimally balance the cost of erroneous decisions and that of the sensor acquisitions'.

In the learning version of the problem, the misclassification probabilities are *a prirori* unknown and a learner must learn the optimal balance based on some feedback available to him. 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

[1] As described in Trapeznikov et al. (2014) security systems utilize a suite of sensors/tests such as X-rays, millimeter wave imagers (expensive & low-throughput), magnetometers, video, IR imagers 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. Sensors providing imagery are scored by humans. The different sensing modalities have diverse costs, in terms of health risks (radiation exposure) and monetary expense.

multiple rounds in a stochastic, stationary environment. Can a learner still achieve the optimal balance in this case? Can a learner achieve this in any of the above practical problems?

Our problem can be framed as a stochastic partial monitoring problem (Bartók et al., 2014), which itself is a generalization of multi-armed bandit problems, going back to Thompson (1933). 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 and could choose the best action. 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. To cast our problem as a partial monitoring problem, the key unknown information can be the misclassification error rates of the classifiers, an action is identified with the subset of sensors selected, the cost of an action is the sum of the misclassification cost of the classifiers that uses the selected sensor subset outputs and the cost of acquiring these sensor outputs, while the observed feedback is the vector of predicted labels by each of the classifiers that use the first, the first and second, etc., up to all sensor outputs from the sensors that were selected. Note that unlike in a conventional bandit problem, we do not get *direct* 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 may not be able to infer potential optimal actions. As usual, in sequential learning under uncertainty, 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 learn the action, while our learning algorithm's performance is evaluated using their cumulative regret with respect to the competitor.

We first prove an (unsurprising) result that states with no further assumptions, no learner can "learn", i.e., no learner can achieve sublinear regret. This negative result led us to 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. Thus, we identify weak dominance as the sharp necessary and sufficient condition for the learnability of our sensor acquisition problem.

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, a problem introduced by Mannor & Shamir (2011). In the reduction, the 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. The weak dominance condition occasionally can be shown to hold by design. For example, we do show that, in fact, a stronger dominance condition holds in the context of the communication systems and error-correcting code cascades, implying the weak dominance condition there. Empirically, with the help of labelled data, we verify that weak dominance condition naturally holds for several real-world problems, including diagnosing breast-cancer and diabetes.

Related Work.

Supervised, batch learning, the problem is well studied.

Related Work: Trapeznikov & Saligrama (2013)

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.

[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.

Greiner et al. (2002) consider 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. They consider the batch, supervised setting.

The literature of learning active classifiers is large (e.g., (Kapoor & Greiner, 2005; Draper et al., 1999; Isukapalli & Greiner, 2001)).

Online learning. Seldin et al. (2014): The decision maker can opt to pay for additional observations of the costs associated with other arms. Not unsupervised. Zolghadr et al. (2013): Online learning with costly features and labels. In each round, learner has to decide which features to observe, known that each features costs some money. The learner can also decide not to observe the label, but the learner always has the option to observe the label. Not unsupervised.

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)).

2 Unsupervised Sensor Acquisition Problem

A learner has access to $K \geq 2$ sensors that provide predictions of an unknown label. It is assumed that the sensors form a cascade (cf. Fig. 1), i.e., they are *ordered* in terms of their prediction efficiency, later sensors are more accurate in predicting the unknown label. However, acquiring the output of later sensor comes at a fixed cost. The dilemma of the learner is that while he knows the ordering of the sensors, the accuracies of the sensors are unknown. The learner's task is to minimize the total prediction cost, which includes both the cost of acquiring the sensor outputs and the cost incurred due to imperfect sensor output. The learner knows the costs, but does not know how efficient the sensors are and learns only the output of the sensors. Learning happens in a sequential setting, where in each round the learner can decide sequentially (within the round) which sensor outputs to observe, while respecting the ordering of the sensors. The output of the last sensor selected serves as the prediction for the round.

The formal specification of the learning problem is as follows: Learning happens sequentially. In round t ($t = 1, 2, \dots$), the environment generates $(Y_t, \hat{Y}_t^1, \dots, \hat{Y}_t^K) \in \{0,1\}^{K+1}$ from a distribution P unknown to the learner.

Here, Y_t is the unknown label of context/instance Z_t to be predicted in round t, while $\hat{Y}_t^k$ is the output of sensor k, a prediction of Y_t. We focus on the case where Z_t is not available to the learner. The case where they are observed is briefly discussed in the supplementary. At the cost of $c_1 + c_2 + \cdots + c_k$, the learner can choose to acquire the outputs of the first k sensors, where $k \in [K] := \{1, \dots, K\}$.

Here, $c_i \geq 0$ is the marginal cost of acquiring the output of sensor i. The costs $c := (c_1, \dots, c_K)$ are known to the learner. Having acquired the output of the first k sensors, the learner predicts the unknown label Y_t using the output of the last sensor acquired, i.e., using $\hat{Y}_t^k$, making the learner incur the loss

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

in round t. The feedback of learner for this round is then $H_t(k) = (\hat{Y}_t^1, \dots, \hat{Y}_t^k)$.

We refer to the above setup as Sensor Acquisition Problem (SAP). Based on the previous description, an instance of SAP is the tuple $\psi = (K, P, c)$, where $K \in \mathbb{N}$, $K \geq 2$, P is a distribution over $\{0,1\}^{K+1}$ and $c \in [0, \infty)^K$. A policy π on a K-sensor SAP problem is a sequence of maps, $(\pi_1, \pi_2, \cdots)$, where $\pi_t : \mathcal{H}_{t-1} \to [K]$ gives the action selected in round t given a history $h_{t-1} \in \mathcal{H}_{t-1}$ that consists of all actions and corresponding feedback observed before t. Let Π denote set of such policies. For any $\pi \in \Pi$, we compare its performance to that of the single best action in hindsight and define its

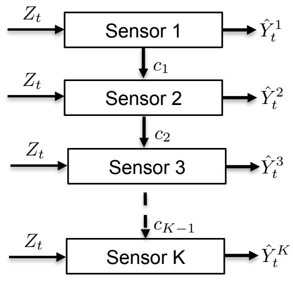

Figure 1: Cascade of sensors

expected regret as follows

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

where I_t denotes the action 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.,

$$\pi^* = \arg\min_{\pi \in \Pi} R_T^{\psi}(\pi). \tag{2}$$

Optimal action in hindsight: For any t, we have

$$\mathbb{E}[L_t(k)] = \Pr\{Y_t \neq \hat{Y}_t^k\} + \sum_{j=1}^{k} c_j = \gamma_k + \sum_{j=1}^{k} c_j\,, \tag{3}$$

where $\gamma_k = \Pr\{Y_t \neq \hat{Y}_t^k\}$ is the misclassification error rate of sensor k. Let $k^* = \arg\min_{k \in [K]} \gamma_k + \sum_{i \leq k} c_i$. Then the optimal policy is to play action k^* in each round. If an action i is played in any round then it adds $\Delta_k := \gamma_k + \sum_{i \leq k} c_i - (\gamma_{k^*} + \sum_{i \leq k^*} c_i)$ to the expected regret. Let $N_k(s)$ denote the number of times action k is selected till time s, i.e., $N_k(s) = \sum_{t=1}^{s} \mathbf{1}_{\{I_t = k\}}$. Then the expected regret can be expressed as

$$R_T^{\psi}(\pi) \quad = \quad \sum_{k \in [K]} \mathbb{E}[N_k(T)]\Delta_k\,. \tag{4}$$

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 Hannan (1957), i.e, a policy $\pi \in \Pi^{\psi}$ such that $R_T^{\psi}(\pi)/T \to 0$.

To fix ideas let us consider SA-Problem with 2 sensors. We enumerate all possible 8 tuples $(Y, \hat{Y}^1, \hat{Y}^2)$ as shown in Table 3, and write probability of 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

$$\gamma_1 - \gamma_2 = p_2 + p_5 - p_1 - p_6. \tag{5}$$

Y	$\hat{Y}^1$	$\hat{Y}^2$	$\Pr(Y, \hat{Y}^1, \hat{Y}^2)$
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

$$\Pr(\hat{Y}^1, \hat{Y}^2) = \begin{cases} p_1 + p_5 \text{ if } (\hat{Y}^1, \hat{Y}^2) = (0,1) \\ p_2 + p_6 \text{ if } (\hat{Y}^1, \hat{Y}^2) = (1,0) \\ p_0 + p_4 \text{ if } (\hat{Y}^1, \hat{Y}^2) = (0,0) \\ p_3 + p_7 \text{ if } (\hat{Y}^1, \hat{Y}^2) = (1,1) \end{cases} \tag{6}$$

Since we only observe feedbacks $(\hat{Y}_t^1, \hat{Y}_t^2)$ and not the true labels Y_t, only marginal probabilities $\Pr(\hat{Y}^1, \hat{Y}^2)$ as given in (6) can be estimated but not $\Pr(Y, \hat{Y}^1, \hat{Y}^2)$. Thus all the decision has to be based on the marginals only. To see when SAP has a Hannan consistent policy, let us consider the following conditions.

Condition 1 *If sensor* 1 *predicts label* 1 *correctly, then sensor* 2 *also predicts it correctly*[3]*, i.e.,*

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

Condition 2 *If sensor* 1 *predicts label* 0 *correctly, then sensor* 2 *also predicts it correctly, i.e.,*

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

The following example demonstrate marginals do not unambiguously decide optimal action under Condition 1. Set $c = 0.35$ and consider the following two cases: 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 (1) 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, whereas $\gamma_1 - \gamma_2 = 0.4$ in the second case. From 3, optimal action is 1 in the first case, whereas it is 2 in the second case. However, for both the cases the marginals $\Pr(\hat{Y}^1, \hat{Y}^2)$ are the same for all pairs $(\hat{Y}^1, \hat{Y}^2)$. Since we only observe $\Pr(\hat{Y}^1, \hat{Y}^2)$, the two cases cannot be distinguished and linear regret is unavoidable. We can argue similarly that Condition (2) is not sufficient for sub-linear regret.

Next, consider that both Condition (1) and (2) hold, i.e.,

Condition 3 *If sensor* 1 *is correct , then sensor* 2 *is also correct, i.e.,*

$$\hat{Y}_t^1 = Y_t \implies \hat{Y}_t^2 = Y_t.$$

Then, $p_1 = p_6 = 0$ and we get $\gamma_1 - \gamma_2 = p_2 + p_5$. Since $p_2 + p_5 = \Pr(\hat{Y}^1 \neq \hat{Y}^2)$, it can be estimated from observations $(\hat{Y}_t^1, \hat{Y}_t^2)$, and the optimal action can be found unambiguously. Thus Condition 3 gives a sufficient for existence of an Hannan consistent policy. In the following we refer to Condition (3) as strong dominance property. For the case of $K > 2$ sensors, its definition is as follows:

Definition 1 (Strong Dominance) *A SA-Problem is said to satisfy strong dominance property if sensor i predicts correctly, then all the sensors in the subsequent stages of the cascade also predict correctly, i.e.,*

$$\hat{Y}_t^i = Y_t \to \hat{Y}_t^j \quad \forall j > i \geq 1. \tag{7}$$

We will now establish necessary and sufficient conditions for SAP learnability For notional convenience rewrite $\gamma_1 - \gamma_2 = p_1 + p_2 + p_5 + p_6 - 2(p_1 + p_6) := p_{12} - 2\delta$, where $p_{12} := \Pr(Y^1 \neq Y^2)$ is the probability that sensors disagree and $\delta := \Pr(Y^2 \neq Y | Y^1 = Y)$ is the conditional probability that sensor 2 is incorrect given that sensor 1 is correct. We can estimate p_{12} from feedback $(\hat{Y}_t^1, \hat{Y}_t^2)$, but δ cannot be estimated.

Theorem 1 *For SA-Problem with $K = 2$, an Hannan consistent policy exists if and only if $c \notin [p_{12} - 2\delta, p_{12}]$.*

Proof: Under dominance condition $\delta = 0$, thus actions 1 is optimal if $p_{12} < c$, otherwise action 2 is optimal. Suppose dominance condition is violated, i.e., $\delta > 0$, but decisions are made assuming dominance condition holds (i.e., using estimates of p_{12} only), then the optimal action is correctly identified provided δ is such that $p_{12} - 2\delta < c \Rightarrow p_{12} < c$ or $p_{12} - 2\delta > c \Rightarrow p_{12} > c$. Now, notice that the latter implication is always true. So, whenever action 2 is optimal, violation of dominance condition does not miss the optimal action. However, the first implication holds if and only if $c \notin [p_{12} - 2\delta, p_{12}]$.

Clearly, when δ is small Hannan consistent policy exits for a large range of c. For the case of $K > 2$ sensors, its definition is as follows:

Definition 2 (Weak Dominance) *A SA-Problem is said to satisfy weak dominance property if $c_k \notin [p_{k-1,k} - 2\delta_{k-1,k}, p_{k-1,k}]$ for all $1 < k < K$, where $p_{k-1,k} = \Pr(Y^{k-1} \neq Y^k)$ and $\delta_{k-1,k} = \Pr(Y^k \neq Y | Y^{k-1} = Y)$.*

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

Many real world applications are designed to satisfy strong dominance property. For example, in wireless communication, increasing block length (more redundancy) improves tolerance against noise. Many practical datasets like, PIMA diabetes dataset and breast cancer dataset, conditional error probabilities are small. (i will add numerical values)

In the following we establish that if dominance property holds 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 Stochastic Multi-armed Bandits with Side Observations

A stochastic multi-armed bandit (MAB), denoted as $\phi := (K, (\nu_k)_{k\in[K]})$, 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_i. 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. However, 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 neighborhood of i such that playing arm i reveals rewards of all arms $j \in \mathcal{N}_i$. Given a set of neighborhood $\{\mathcal{N}_i, i \in [K]\}$, let $\phi_G := (K, (\nu_k)_{k\in[K]}, G)$ 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. To avoid cluttering, we henceforth drop subscript G in ϕ_G and it should be clear from context if side-observations exists or not.

Let Π^ϕ denote a set of polices on ϕ that maps the past history into an arm in each round.. If the learner knows $\{\nu_k\}_{k\in[K]}$, then the optimal policy is to play the arm with highest mean. Given a 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 (only reward from the arm played contribute to the regret and not that from the side-observations): Let π selects arm i_t in round t. After T rounds, its regret is

$$R_T^\phi(\pi) = T\mu_{i^*} - \sum_{t=1}^{T} \mu_{i_t}, \tag{8}$$

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 $N_i^\phi(t)$ 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 is to learn a policy that minimizes the regret.

Buccapatnam et al. (2014) establish that any policy $\pi \in \Pi^\phi$ where side observation graph is such that $i \in \mathcal{N}_i$ for all $i \in [K]$ satisfies

$$\liminf_{T\to\infty} R_T^\phi(\pi)/\log T \geq \eta(G) \tag{9}$$

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

$$\begin{aligned} \text{LP1}: \quad & \min_{\{w_i\}} \sum_{i\in[K]} (\mu_{i^*} - \mu_i) w_i \\ & \text{subjected to} \sum_{j\in\mathcal{N}_i} w_i \geq 1/D(\mu_i||\mu_{i^*}) \text{ and } w_i \geq 0 \text{ for all } i \in [K], \end{aligned} \tag{10}$$

$D(\mu_i||\mu_{i^*})$ here denotes the Kullback-Leibler divergence between ν_i and ν_{i^*}. When $\mathcal{N}_i = \{i\}$ for all $i \in [K]$, it reduces to the classical lower bound of $\sum_{i\neq i^*}(\mu_{i^*} - \mu_i)/D(\mu_i||\mu_{i^*})$ established in Lai & Robbins (1985). Further, Buccapatnam et al. (2014) also gave an UCB based strategy, named UCB-LP, that explores arms at a rate in proportion to the size of their neighborhood. Specifically, UCB-LP plays arms in proportions to the values $\{z_i^*, i \in [K]\}$ computed from the following linear optimization which is a relaxation of LP1.

$$\text{LP2} : \min_{\{z_i\}} \sum_{i \in [K]} z_i \quad \text{subjected to} \sum_{j \in \mathcal{N}_i} z_i \geq 1 \text{ and } z_i \geq 0 \text{ for all } i \in [K] \tag{11}$$

The regret of UCB-LP is upper bounded by

$$\mathcal{O}\left(\sum_{i \in [K]} z_i^* \log T\right) + \mathcal{O}(K\delta), \tag{12}$$

where $\delta = \max_{i \in [K]} |\mathcal{K}_i|$ and $\{z_i^*\}$ are the optimal values of $LP2$.

Definition 3 (Domination number Buccapatnam et al. (2014)) *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 $\xi(G)$.*

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 $\mathcal{O}(\xi(G) \log T)$.

5 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.

Definition 4 (Regret Equivalence) *Consider a SA-Poblem $\psi := (K, P, c)$ and a bandit problem $\phi := (N, (\nu_i)_{i \in [N]}, G)$ with side-information graph G. We say that ψ is regret-equivalent to ϕ if given a policy π for ψ, one can come up with a policy π' that uses π, such that the regret of π' on any instance of ϕ is the same as the regret of π on some corresponding instance of ψ, and vice versa.*

In the following we first consider SAP with 2 sensors and then the general case with more than 2 sensors. SAP with 2 sensors is useful to draw comparison with the well studied apple tasting problem Helmboat et al. (2000) and understand role of the dominance property.

5.1 SAP with two sensors

In SAP with two sensors, while action 1 reveals no information about the loss values, under dominance property, action 2 reveals (partial) information about the loss from both actions. To see this, let $I_t = 2$. If predictions of sensors disagree, i.e., $\hat{Y}_t^1 \neq \hat{Y}_t^2$, then dominance property implies that only sensor 2 is correct, i.e., $\hat{Y}_t^1 \neq Y_t$ and $\hat{Y}_t^2 = Y_t$. Hence $L_t(1) = 1$ and $L_t(2) = c$. On the other hand, if predictions agree, i.e., $\hat{Y}_t^1 = \hat{Y}_t^2$, then either predictions of both are correct or both are incorrect, and we can only infer that $L_t(2) - L_t(1) = c_1 + c_2 > 0$. Thus, each time learner plays action 2, loss from both actions is known only when sensor output disagree, otherwise.

Apple tasting Helmboat et al. (2000): In this problem, a learner gets a sequence of apples some of which can be rotten. In each round, the learner can either accept or reject an apple and irrespective of his action, a penalty is incurred if the apple is rotten in that round. If an apple is rejected, learner do not get to observe its quality, and if accepted, the learner tastes the apple and knows its quality. In the latter case loss incurred is known, and the learner can also know the loss he would have incurred if he opted to reject it. The goal of the learner is to taste more good apples. A SAP with dominance property is thus a general version than the apple tasting problem as unlike in apple tasting problem loss value are revealed only in few instances. We next show that SAP satisfying dominance property can be efficiently solved.

5.2 SAP with more than two actions

In SAP with two sensors, only action 2 provides partial information about the loss of both actions. In the case with $K > 2$ sensors, by playing an action k, partial information about the loss from actions $l < k$ can be inferred by recursively applying the dominance property to each pair of sensors. Further, any information provided by action $k > 2$ is contained in that provided by all actions $k' \geq k$ as $H_t(k) \subseteq H_t(k')$.

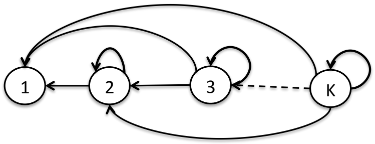

Figure 2: Side observation graph G_S

This information structure can be represented by a directed graph $G_S = (V, E)$, where $|V| = K$ and $E = \{(i, j) : 1 < i \leq j \leq K\}$. Note that G_S has self loops for all nodes except for node 1. The nodes in G_S represents actions of in 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).

We now have all the ingredients to relate SAP problem with MAB.

Theorem 2 *Let the dominance condition (7) holds. Then SAP is regret equivalent to a MAB with side-observation graph G_S.*

Then, from (9), we immediately obtain following regret lower bound for SAP.

Proposition 1 (SAP regret lower bound) *Let π be any policy on SAP 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,*

$$\liminf_{T \to \infty} R_T^{\psi}(\pi)/\log T \geq \kappa \textit{ where} \tag{13}$$

$$\begin{aligned}
\kappa = \min_{\{w_i\}} &\sum_{i \in [K]} (\mu_{i^*} - \mu_i) w_i \\
\textit{subjected to} &\sum_{ji} w_i \geq 1/D \left(\mu_i + \sum_{j<i} c_j \| \mu_{i^*} + \sum_{j<i^*} c_j \right) \textit{ for all } i \in [K] \qquad (14)\\
w_i \geq 0 &\textit{ for all } i \in [K].
\end{aligned}$$

Proposition 2 (K-SAT regret upper bound) *Given a SA-problem ψ, there exists a policy $\pi \in \Pi^{\psi}$ such that*

$$R_T^{\psi}(\pi) \leq \mathcal{O}(\log T) + \mathcal{O}(K^2). \tag{15}$$

As discussed in the proof of Theorem 2, using UCB-LP on side-observation graph G_S we can obtain a policy for SAP that maintains regret guarantee of UCB-LP which is given as $\mathcal{O}(\xi(G_S) \log T) + \mathcal{O}(K^2)$. Now the claim follows by noting that $\xi(G_S) = 1$.

6 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

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 3: Error statistics

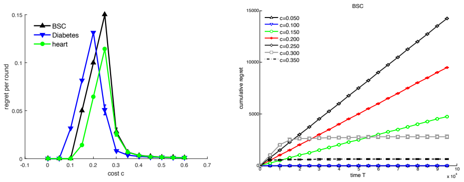

Figure 4: Left side figure plots regret per round against cost for all the datasets. The right side plots regret for different cost in BSC experiment

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 (6). 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 (6) with 95% confidence bounds. The left Figure in 6 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 (6). As seen, regret is linear for all c in $[0.1\ 0.26]$ and is sublinear outside this region.

7 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:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{arm } 1 \\ \pi_t(H_{t-1} \cup \{i, r_{t-1} \cup o_{t-1}\}) \text{ if } a_{t-1} = \text{arm i} \end{cases} \tag{16}$$

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:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1, 0\}) \text{ if } a'_{t-1} = \text{action } 1 \\ \pi'_t(H'_{t-1} \cup \{i, \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} \cdots \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^i\}\}) \text{ if } a_{t-1} = \text{action i.} \end{cases} \tag{17}$$

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 (16). The regret of policy π that plays actions i, $N_i^{\psi}(T)$ times is given by

$$\begin{aligned} R_T^{\psi}(\pi) \quad &= \quad \sum_{i=1}^{K} \left[\left(\gamma_i + \sum_{j<i} c_j \right) - \left(\gamma_{i^*} + \sum_{j<i^*} c_j \right) \right] \mathbb{E}[N_i^{\psi}(T)] \qquad (18) \\ & \qquad (19) \end{aligned}$$

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

$$R_T^{\phi}(f_1(\pi)) = \sum_{i=1}^{K} \left[\left(\gamma_1 - \gamma_{i^*} - \sum_{j<i^*} c_j \right) - \left(\gamma_1 - \gamma_i - \sum_{j<i} c_j \right) \right] \mathbb{E}[N_i^{\phi}(T)], \tag{20}$$

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 (17) and we can argue similarly that they are regret equivalent. This concludes the proof.

8 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:

$$R_T^{\phi_c}(\pi) = \sum_{t=1}^{T} \mathbb{E}\left[L_t(\pi(x_t)|x_t)\right] - \sum_{t=1}^{T} \min_{a \in \mathcal{A}} \mathbb{E}\left[L_t(a|x_t)\right]. \tag{21}$$

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.

8.1 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

$$R_T^L(\pi) = \sum (x_t^*)'\theta^* - \sum x_t'\theta^*. \tag{22}$$

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 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.*

8.2 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 $\epsilon_t = \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} - x_t'\theta$. 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.

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.

Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. Stochastic bandits with side observation on networks. In *Proceeding of Sigmetrics*, 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.

Hannan, J. Approximation to bayes risk in repeated plays. *Contributions to the Theory of Games*, 3: 97–139, 1957.

Helmboat, D. P., Littlestone, PN, and Long, P.M. Apple tasting. *Journal of Information and Computation*, 161(2):85–139, 2000.

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.

Lai, Tze Leung and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Journal of Advances in applied mathematics*, 6(1):4–22, 1985.

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.

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.

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