Label Prediction Under Partial Monitoring

M. Hanawal, J. Wang, V. Saligrama, C. Szepesvári

I. Sensor Acquisition Problem

We consider the problem of efficient label prediction under partial monitoring. 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 select a subset of sensors in the cascade sequentially and stop at any depth.

Let $\{Y_t\}_{t>0}$ denote sequence of labels generated according to an unknown distribution $\mathcal{D}$. In round t, let $\hat{Y}_t^k$ denote the prediction of the k^{th} sensor. We denote K actions of the learner as $\mathcal{A} = \{1, \ldots, K\}$, where the action k indicates acquiring predictions from sensors $1, \ldots, k$ and classifying using the prediction $\hat{Y}_t^k$. The prediction error rate of the k^{th} sensor is denoted as $\gamma_k := \Pr\{Y_t \neq \hat{Y}_t^k\}$. We assume that $\gamma_{k-1} \geq \gamma_k$ 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 $k-1$.

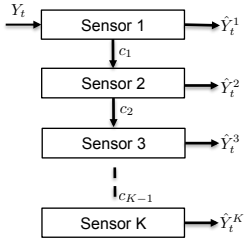

Figure 1. Cascade of sensors

Let $H_t(i)$ denote the feedback observed in round t from action i. When the learner selects action i at time t, the predictions of all sensors in the selected path are observed and the feedback is $H_t(i) = \{\hat{Y}_t^1, \ldots, \hat{Y}_t^i\}$. The loss incurred in each round is defined as follows. When the learner selects action i, the loss is the error in prediction of sensor i plus the sum of the costs incurred along the path $(c_1, \ldots, c_i)$. Let $L_t(i)$ denote the loss in round t for taking action i. Then,

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

We refer to the above setup as Sensor Acquisition Problem (SAP) and denote it as $\psi = (K, \mathcal{A}, (\gamma_i, c_{i-1})_{i\in[K]})$[1]. A policy $\pi = (\pi_1, \pi_2, \cdots)$ on ψ, where $\pi_t : H_{t-1} \to \mathcal{A}$, gives action selected in each round using the set of actions played and the corresponding feedback observed in the past. We compare the performance of policy with respect to the optimal policy (single best action in hindsight) and define expected regret of a policy π as follows

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

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 maximum expected regret, i.e.,

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

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

where Π denote the set of policies that maps past history to an action in $\mathcal{A}$ in each round.
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 $i^* = \arg\min_{i\in\mathcal{A}} \gamma_i + \sum_{j<i} c_j$. Then the optimal policy is to play action i^* in each round. If an action i is played in any round then it adds $\Delta_i := \gamma_i + \sum_{j<i} c_j - (\gamma_{i^*} + \sum_{j<i^*} c_i)$. to the expected regret. Let I_t denote the action slected in round t and $N_i(s)$ denote the number of times action i is selected till time s, i.e., $N_i(s) = \sum_{t=1}^{s} \mathbf{1}_{\{I_t=i\}}$. Then the expected regret can be expressed as

$$R_T^S(\pi) \;\; = \;\; \sum_{i\in\mathcal{A}} \mathbb{E}[N_i(T)]\Delta_i. \tag{5}$$

II. Background on Stochastic Multi-armed Bandits

A stochastic multi-armed bandit (MAB), denoted as $\phi := (K, (\nu_k)_{1\leq k\leq K})$, is a sequential learning problem where number of arms K is known and each arm $k \in [K]$ gives rewards drawn according an unknown distribution ν_k. A policy of the learner is any allocation strategy that selects an arm in each round based on the past history. Let $X_{i,n}$ denote the random reward from arm i in its nth play. For each am $i \in [K]$ $\{X_{i,t} : t > 0\}$ are independently and identically (i.i.d) distributed and they are independent across arms. We note that in the standard MAB setting, the learner only observes reward from the selected arm and no information from the other arms is revealed in that round. The performance of a policy is measured with respected to the optimal performance defined in terms of cumulative regret (or simply regret) as follows: Let π denote a policy that selects arm k_t in round t. The regret of policy π after T rounds is

$$R_T^B(\pi) = \max_{1\leq i\leq K} \mathbb{E}\left[\sum_{t=1}^{T} X_{i,t}\right] - \mathbb{E}\left[\sum_{i=1}^{T} X_{k_t,N_{k_t}(t)}\right], \tag{6}$$

where $N_k(t) = \sum_{s=1}^{t} \mathbf{1}\{k_s = k\}$ denotes the number of plays of arm k till round t. For each $k \in [K]$, let μ_k denote the mean value distribution ν_k and $i^* = \arg\min_{i\in[K]} \mu_i$ denote an arm with the smallest mean. The regret of a policy π can be expressed

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

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 [1] 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 consistent policy satisfies

$$\liminf_{n\to\infty} \frac{R_T^B(\pi)}{\log T} \geq \sum_{i\neq i^*} \frac{\mu_{i^*} - \mu_i}{D(\mu_{i^*}||\mu_i)}, \tag{7}$$

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

Auer et, al. [2] 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

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

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.

A. 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 $j \in \mathcal{K}_i$.The side-observation graph is known to the learner and remains fixed during the play.

A policy for ϕ_G maps the past history (including the side-observations) to an action in each round. With some abuse of notation, we denote the regret of any policy π_G that uses side-information over a period T as $R_T^B(\pi_G)$ and is given by (6). Note that, in each round, only reward from the arm played contribute to the regret and not that from the side-observations. In [3] the authors extended the lower bound in (7) to incorporate the effect of side-observations. Specifically, they established that any policy π_G on a graph where $i \in \mathcal{N}_i$ for all $i \in [K]$ satisfies [3]

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

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

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

Definition 1 (Domination number [3]): 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 denoted as $\xi(G)$.

The authors in [3] gave an UCB based strategy, named UCB-LP, that exploits the side-observations structure 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 $LP1$.

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

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 $LP2$. 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 $\mathcal{O}(\xi(G)\log T)$.

B. 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 $\Theta(\log T)$.

III. Regret Equivalence

In the SAP, the true labels are never revealed, hence the learner cannot learn the error rates of the sensors. Thus, SAP appears to be an hopeless case where nothing can be learned. In this section we establish that seemingly hopeless SAP can be efficiently solved provided the following dominance condition holds:

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

$$\hat{Y}_t^i = Y_t \rightarrow \hat{Y}_t^j \quad \forall j > i \geq 1 \tag{13}$$

In the following 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.

A. 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 [4]], 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 (13) holds. Then SAP ψ with $K = 2$ is regret-equivalent to a stochastic 1-armed bandit.

Proof: 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:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{fixed rewad arm} \\ \pi_t(H_{t-1} \cup \{2, r_{t-1}\}) \text{ if } a_{t-1} = \text{stochastic arm} \end{cases} \tag{14}$$

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:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1, c\}) \text{ if } a'_{t-1} = \text{action 1} \\ \pi'_t(H'_{t-1} \cup \{2, \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\}\}) \text{ if } a_{t-1} = \text{actions 2.} \end{cases} \tag{15}$$

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, $\Pr\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} = \gamma_1 - \gamma_2$.

$$\begin{aligned} \Pr\{\hat{Y}_t^1 \neq \hat{Y}_t^1\} &= \Pr\{\hat{Y}_t^1 = Y_t, \hat{Y}_t^2 \neq Y_t\} + \Pr\{\hat{Y}_t^2 = Y_t, \hat{Y}_t^1 \neq Y_t\} && (16) \\ &= \Pr\{\hat{Y}_t^2 = Y_t, \hat{Y}_t^1 \neq Y_t\} \text{ from assumption (13)} && (17) \\ &= \Pr\{\hat{Y}_t^1 \neq y_t\} \Pr\{\hat{Y}_t^2 = Y_t | \hat{Y}_t^1 \neq Y_t\} && (18) \\ &= \Pr\{\hat{Y}_t^1 \neq Y_t\} \left(1 - \Pr\{\hat{Y}_t^2 \neq Y_t | \hat{Y}_t^1 \neq Y_t\}\right) && (19) \\ &= \Pr\{\hat{Y}_t^1 \neq Y_t\} \left(1 - \frac{\Pr\{\hat{Y}_t^2 \neq Y_t, \hat{Y}_t^1 \neq Y_t\}}{\Pr\{\hat{Y}_t^1 \neq Y_t\}}\right) && (20) \\ &= \Pr\{\hat{Y}_t^1 \neq Y_t\} - \Pr\{\hat{Y}_t^2 \neq Y_t\} \text{ by contrapositve of (13)} && (21) \end{aligned}$$

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

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

Let $f(\pi)$ denote the policy for the 1-armed bandit obtained using the mapping (14). Now, for the 1-armed bandit, where the arm with stochastic rewards is optimal, we have

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

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 cases. 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,

$$\liminf_{T\to\infty} R_T^S(\pi)/\log T \geq \frac{|\gamma_1 - \gamma_2 - c|}{D(\gamma_1 - \gamma_2, c)} \tag{22}$$

where $D(\gamma_1 - \gamma_2, c)$ is the KL-divergence between $\gamma_1 - \gamma_2$ and c.

Proposition 2 (SAT 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 (15) is upper bounded as

$$R_T^S(g(\pi)) \leq \frac{6\log T}{|\gamma_1 - \gamma_2 - c|} + |\gamma_1 - \gamma_2 - c|(1 + \pi^2/3) \tag{23}$$

when $\pi' = UCB1$

B. 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 applying the dominance condition recursively to any pair of actions we can obtain information about the losses for all actions $i > 2$. Further, any information provided by action $i > 2$ is contained in that provided by all actions $j \geq i$. If action i 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 for all actions $j \leq i$. Hence, action i provides information about all actions $j \leq i$ 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).

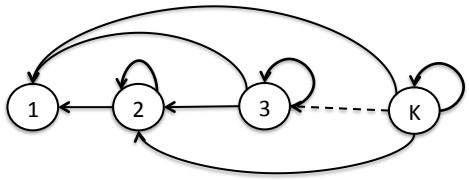

Figure 2. Side observation graph G^S

Theorem 2: Let the dominance condition (13) holds. Then SAP ψ with $K \geq 2$ is regret equivalent to a MAB with side-observation graph G^S.

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:

$$a_t = \begin{cases} \pi_t(H_{t-1} \cup \{1, \emptyset\}) \text{ if } a_{t-1} = \text{arm 1} \\ \pi_t(H_{t-1} \cup \{i, r_{t-1} \cup o_{t-1}\}) \text{ if } a_{t-1} = \text{arm i} \end{cases} \tag{24}$$

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:

$$a'_t = \begin{cases} \pi'_t(H'_{t-1} \cup \{1,0\}) \text{ if } a'_{t-1} = \text{action 1} \\ \pi'_t(H'_{t-1} \cup \{i, \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^2\} \cdots \mathbf{1}\{\hat{Y}_t^1 \neq \hat{Y}_t^i\}\}) \text{ if } a_{t-1} = \text{action i.} \end{cases} \tag{25}$$

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

$$R_T^S(\pi) = \sum_{i=1}^{K} \left[\left(\gamma_i + \sum_{j<i} c_j \right) - \left(\gamma_{i^*} + \sum_{j<i^*} c_j \right) \right] \mathbb{E}[N_i(T)] \tag{26}$$

$$\tag{27}$$

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

$$R_T^B(f_1(\pi)) = \sum_{i=1}^{K} \left[\left(\gamma_1 - \gamma_{i^*} - \sum_{j<i^*} c_j \right) - \left(\gamma_1 - \gamma_i - \sum_{j<i} c_j \right) \right] \mathbb{E}[N_i(T)] \tag{28}$$

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

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 $[0,1]$, they may not be directly applicable to our setting. However, this can be over come by setting the distributions of arm k, ν_k, to have mean $\gamma_1 - \gamma_i - \sum_{j<i} c_j + \sum_{k\leq K-1} c_k$. Note that we translated the mean of each arm by the same amount, which does not change the regret value. For $k=2$, this recovers the SAP with 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., $\mathbb{E}[N_i(T)] = o(T^a)$ for all $a > 0$ and $i \neq i^*$. Then,

$$\liminf_{T\to\infty} R_T^S(\pi)/\log T \geq \kappa \tag{29}$$

where

$$\begin{aligned} \kappa = & \min_{\{w_i\}} \sum_{i\in[K]} (\mu_{i^*} - \mu_i) w_i \\ & \text{subjected to} \sum_{ji} w_i \geq 1/D(\mu_i + \sum_{j<i} c_j || \mu_{i^*}) \text{ for all } i \in [K] \\ & w_i \geq 0 \text{ for all } i \in [K] \end{aligned} \tag{30}$$

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 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 (25) is upper bounded as

$$R_T^S(g(\pi)) \leq \mathcal{O}(\xi(G^S)\log T) \tag{31}$$

when $\pi' = UCB - LP$ [3].

IV. Extension to context based prediction

In this section we consider that learner gets contextual information. Let $x_t \in \mathcal{C} \subset \mathcal{R}^d$ denote the context associated with y_t, where $\mathcal{C}$ denotes a compact set.

Let $L_t(a|x)$ denote the loss incurred by selecting action a when the context is x. For any policy $\pi : \mathcal{C} \to \mathcal{A}$, let $\pi(x_t)$ denote the action selected when the context is x_t using the observed feedback from past actions. For any sequence fo samples $x_t, y_{t t>0}$, the regret of a policy π is defined as:

$$R_T(\pi) = \sum_{t=1}^{T} (L_t(\pi(x_t)) - L_t(a^*(x_t))) \tag{32}$$

where $a^*(x) = \arg\min_{a \in \mathcal{A}} \mathbb{E}[L_t(a|x)]$. The goal is to learn a policy that minimizes the expected regret, i.e.,

$$\pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}[R_T(\pi)]. \tag{33}$$

We first consider the case with two actions studied in Section **??**. The predictions of both the devices are context dependent denoted as $\gamma_1(x_t) = \Pr\{\hat{y}_t^1 \neq y_t | x_t\}$ and $\gamma_1(x_t) = \Pr\{\hat{y}_t^2 \neq y_t | x_t\}$. We assume that the difference of prediction errors satisfy $\gamma_1(x) - \gamma_2(x) = \theta' x$ for all $x \in \mathcal{C}$ for some $\theta \in \mathcal{R}^d$.

V. Relaxing the dominance assumption

We next consider weaker dominance condition. Suppose we interpret label-1 as 'threat', it is natural to assume that whenever device-2 does not label incoming instance as threat, then device-1 also does the same. Specifically, we assume that

$$\hat{y}_t^2 \neq 1 \implies \hat{y}_t^1 \neq 1, \tag{34}$$

which is equivalent to

$$\hat{y}_t^2 = 0 \implies \hat{y}_t^1 = 0. \tag{35}$$

As it turns out, this assumption does not lead to an identifiable system.

Cs: However, crowdsourcing model is also unidentifiable in a strict sens and the rank c assumption is what saves the day there.

VI. Calibrating the sensors

In this sections we assume that the predictions accuracy can be controlled. Let $\hat{\hat{y}}_t^i$ denote the measurement output by device-i in round t. The prediction of device-i is given by

$$\hat{y}_t^i = \text{sign}(\hat{\hat{y}}_t^i - \eta_i) \tag{36}$$

where $\eta_i \in \mathcal{R}$ is the calibration parameter set by the learner. Thus, setting η_i the learner can control the prediction accuracy of the devices. A policy involves selecting a device and setting its calibration parameter in each round. The goal is to learn a policy that minimizes the expected regret defined in **??**.

VII. Relationship to crowdsourcing

A basic problem formulation in crowdsourcing is the following:[2] One is given a $W \times T$ table Y of binary labels (for convenience, $Y \in \{\pm\}^{W \times T}$), the (w,t)th entry Y_{wt} in the table corresponding to the response of worker w for task t. It is assumed that the performance of each worker is stable across the tasks and that tasks are also randomly chosen from a fixed distribution. More precisely, $Y_{w,t} = (\xi_{w,t,-1}\mathbf{1}_{Y_t^*=-1} + \xi_{w,t,+1}\mathbf{1}_{Y_t^*=+1})Y_t^*$, where $Y_t^* \in \{\pm 1\}$ is the "true" unobserved label, $\xi_{w,t,y} \in \{\pm 1\}$ is the variable that indicates corruption of label $y \in \{\pm 1\}$ of worker w in task t. It is assumed that $(Y_t^*)_{1 \leq t \leq T}$, $(\xi_{w,t,y})_{w,t,y}$ are mutually independent of each other, while $Y_t^* \sim_D Y_{t'}^*$ and $\xi_{w,t,y} \sim_D \xi_{w,t',y}$ for any t, t', w, y. The joint distribution of the random variables in the observed table Y is thus uniquely determined by the

[2]See https://uwspace.uwaterloo.ca/bitstream/handle/10012/9841/Szepesvari_David.pdf?sequence=1&isAllowed=y and the references therein.

probabilities $\Pr\{Y_1^* = +1\}$ and $\Pr\{\xi_{w,1,y} = +1\}$, $1 \le w \le W$, $y \in \{\pm 1\}$. The model described dates back to the work of Dawid and Skene in 1979.[3]

One basic task in crowdsourcing is to infer the values of $(Y_t^*)_{1 \le t \le T}$ given the observed table Y. This is called the inference problem (this is often studied even in the lack of assumptions on the task generation process). Very often, this is studied under the so-called symmetric noise assumption when $\xi_{w,t,y}$ and $\xi_{w,t,-y}$ are identically distributed. In this case, the optimal way to aggregate the labels provided by the workers is to use a weighted majority vote, i.e., using $\mathrm{sgn}(\sum_{w=1}^{W} v_w Y_{w,t})$ to predict the label for task t, where $v_w^* = \ln(\frac{1+s_w^*}{1-s_w^*})$ and $s_w^* = \Pr\{\xi_{w,1,1} = 1\} - \Pr\{\xi_{w,1,1} = -1\}$ denotes the "skillfulness" of worker w. In the lack of the knowledge of worker skill levels, the skills are estimated. This is based on writing $Y_{w,t} = \xi_{w,t} Y_t^* = s_w^* Y_t^* + Z_{w,t} Y_t^*$, where $\mathbb{E}[Z_{w,t} | Y_t^*] = 0$ and we used that $\mathbb{E}[\xi_{w,t,1} | Y_t^*] = s_w^*$. Thus, Y can be viewed as a noisy observation of a rank-one matrix.

Note that in the lack of the symmetry assumption, the rank-one approximation becomes a rank-two approximation, a case, which, to the best of our knowledge, has not been studied theoretically in the literature so far.

REFERENCES

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Journal of Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Auer, Nicholó-Cesa-Bianchi, and P. Fischer, "Finite-time analysis of multiarmed bandit problem trade-offs," *Journal of Machine Learning*, vol. 3, pp. 235–256, 2002.

[3] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Stochastic bandits with side observation on networks," in *Proceeding of Sigmetrics*, 2014.

[4] D. P. Helmboat, P. Littlestone, and P. Long, "Apple tasting," *Journal of Information and Computation*, vol. 161, no. 2, pp. 85–139, 2000.

[3]Dawid, P., Skene, A. M., Dawid, A. P., and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics, pages 20–28.