Label Prediction Under Partial Monitoring

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

I. Problem Setup

We consider the problem of efficient label prediction under partial monitoring. Let $\{y_t\}_{t>0}$ denote sequence of binary labels generated according to an unknown but fixed distribution $\mathcal{D}$. The learner can use a 'cheap' sensor (device-1) or/and a 'costly' sensor (device-2) to predict the labels. In round t, let $\hat{y}_t^1$ and $\hat{y}_t^2$ denote the predictions of device-1 and device-2 respectively. We assume that device-1 has lower performance than device-2 in the sense that prediction error rate of device-1, denoted as $\gamma_1 := \Pr\{y_t \neq \hat{y}_t^1\}$, is larger than or equal to that of device-2, denoted as $\gamma_2 := \Pr\{y_t \neq \hat{y}_t^2\}$ ($\gamma_1 > \gamma_2$). In each round t, the learner can take the following actions:

- Action-1: use device-1.
- Action-2: use both the devices.

For ease of notion, we denote Action-1 as a_1 and Actions-2 as a_2 and write $\mathcal{A} = \{a_1, a_2\}$. Let H_t denote the feedback observed in round t by selecting an action. When action a_1 is selected, the learner observes $\hat{y}_t^1$, and when a_2 is selected, he observes both $\hat{y}_t^1$ and $\hat{y}_t^2$. That is,

$$H_t(a_t) = \begin{cases} \hat{y}_t^1 \ \text{ if } \ a_t = a_1, \\ \{\hat{y}_t^1, \hat{y}_t^2\} \ \text{ if } \ a_t = a_2. \end{cases} \tag{1}$$

The loss incurred in each round is defined as follows. When action a_1 is selected, the loss is 1 unit if prediction of device-1 (observed feedback) is incorrect, otherwise loss is zero. When actions a_2 is selected, a fixed loss of $c > 0$ is incurred in addition to the prediction loss of device-2, which is 1 unit if device-2's prediction is incorrect and 0 otherwise. Let $L_t(a_t)$ denote the loss in round t for taking action a_t. Then,

$$L_t(a_t) = \begin{cases} \mathbf{1}_{\{\hat{y}_t^1 \neq y_t\}} \ \text{ if } \ a_t = a_1, \\ \mathbf{1}_{\{\hat{y}_t^2 \neq y_t\}} + c \ \text{ if } \ a_t = a_2. \end{cases} \tag{2}$$

Names in alphabetical order

The learner uses a policy that selects an action in $\mathcal{A}$ in each round using the feedback observed in the past. The regret of a policy π that selects action $\pi_t \in \mathcal{A}$ in round t over a period T with respect to an action $a \in \{a_1, a_2\}$ is given as

$$R_T(\pi, a) = \sum_{t=1}^{T} (L_t(\pi_t) - L_t(a)). \tag{3}$$

The goal of the learner is to learn a policy that minimizes the maximum expected regret, i.e.,

$$\pi^* = \arg\min_{\pi \in \Pi} \max_{a \in \mathcal{A}} \mathbb{E}[R_T(\pi, a)], \tag{4}$$

where Π denote the set of policies that maps past history to an action in $\mathcal{A}$ in each round.

Optimal action in hindsight: Since for any t

$$\mathbb{E}[L_t(a)] = \begin{cases} \gamma_1 \ \text{ if } \ a = a_1, \\ \gamma_2 + c \ \text{ if } \ a = a_2, \end{cases} \tag{5}$$

for any $\pi \in \Pi$, $\mathbb{E}[R_T(\pi, a)]$ is maximized by an action a^* such that $a^* = a_1$ if $\gamma_1 \leq \gamma_2 + c$, and $a^* = a_2$ otherwise. We rewrite goal of the learner as

$$\pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}[R_T(\pi, a^*)]. \tag{6}$$

Let I_t denote the action taken in round t and $N_i(s)$ denote the number of times action i is selected till time s, i.e., $N_i(s) = \sum_{t=1}^{s} \mathbf{1}_{\{I_t = i\}}$. The expected regret can be expressed as

$$\mathbb{E}[R_T(\pi, a^*)] \ = \ \sum_{i=1}^{2} \mathbb{E}[N_i(T)] \Delta_i \tag{7}$$

where $\Delta_1 = \gamma_1 - \mathbb{E}[L(a^*)]$ and $\Delta_2 = \gamma_2 + c - \mathbb{E}[L(a^*)]$. Note that for all $i = 1, 2$, either $\Delta_i = |\gamma_1 - \gamma_2 - c|$ or $\Delta_i = 0$.

Assumptions: In the following, we assume that the labels and predictions are binary and that predictions of device-2 dominates that of device-1. More precisely, we assume that whenever device-1 makes no prediction error, device-2 is also guaranteed to make no prediction error, i.e., in every round t,

$$\hat{y}_t^1 = y_t \implies \hat{y}_t^2 = y_t. \tag{8}$$

Reduction to the apple tasting problem: The feedback after action a_1 reveals no information about the loss incurred in that round. However feedback after action a_2 reveals (partial) information about the loss of both actions. Suppose feedback is such that the predictions of devices

Algorithm 1 StAT

Input:
c cost of sensor
Initialization:
Play a_2 once, observe $X_{2,1}$
$N_2(1) \leftarrow 1$, $Y_1 \leftarrow X_{2,1}$ and $\hat{\mu}_1 \leftarrow \frac{Y_1}{N_2(1)}$
for $t = 2, 3 \cdots$ **do**
  if $\hat{\mu}_{t-1} + \sqrt{\frac{6 \log t}{N_2(t-1)}} > c$ **then**,
    play action a_2 and observe $X_{2,t}$,
    $N_2(t) \leftarrow N_2(t-1) + 1$, $Y_t \leftarrow Y_{t-1} + X_{2,t}$
  else
    play action a_1
  end if
  $\hat{\mu}_t \leftarrow \frac{Y_t}{N_2(t)}$
end for

disagree, i.e., $\hat{y}_t^1 \neq \hat{y}_t^2$ after action a_2. The dominance assumption 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 device-2 and loss is zero. Suppose the predictions of the devices agree, i.e., $\hat{y}_t^1 = \hat{y}_t^2$, then the dominance assumption 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 the prediction of both the devices agree, then the difference of loss of the actions is $L_t(a_2) - L_t(a_1) = c > 0$. And if the predictions of the devices disagree, then dominance assumption implies that $L_t(a_1) = 1$ and $L_t(a_2) = c$ or $L_t(a_2) - L_t(a_1) = c - 1 < 0$. Thus, each time learner selects action a_2, he gets to know whether or not he was better off by selecting the other action, and this is the only information he requires to distinguish the optimal arm. In the next section, we formalize this notion and give an algorithms that learns the optimal actions efficiently.

II. Algorithm and Regret bounds

Let $X_{2,t} = \mathbf{1}_{\{\hat{y}_t^1 \neq \hat{y}_t^2\}}$ denote whether the predictions of the devices agree or not in round t. Note that $X_{2,t}$ is observed only when the learner selects action a_2. Let $\hat{\mu}_s$ denote the empirical mean of the samples $\{X_{2,t}\}$ observed till time s, given by

$$\hat{\mu}_s = \frac{1}{N_2(s)} \sum_{t=1}^{s} X_{2,t} \mathbf{1}_{\{I_t = 2\}}. \tag{9}$$

The following algorithm named Stochastic Apple Tasting (StAT) is based on UCB strategy. In each round StAT checks if sum of the current estimate and the confidence term is larger than c. If the condition holds, it selects action a_2, otherwise it selects actions a_1. We show that StAT achieves logarithmic expected regret.

Theorem 1: Let the dominance assumption (8) holds. For any $c \in [0, 1]$, the expected regret of StAT is bounded as follows:

$$R_T(StAT, , a^*) \leq \frac{6 \log T}{|\mu - c|} + \frac{\pi^2}{6} + 1, \tag{10}$$

where $\mu = \gamma_1 - \gamma_2$.

A. Regret Analysis

The following lemma is immediate.

Lemma 1: Let the dominance assumption (8) holds. Then, $\Pr\{\hat{y}_t^1 \neq \hat{y}_t^2\} = \gamma_1 - \gamma_2$.

Proof:

$$\Pr\{\hat{y}_t^1 \neq \hat{y}_t^1\} = \Pr\{\hat{y}_t^1 = y_t, \hat{y}_t^2 \neq y_t\} + \Pr\{\hat{y}_t^2 = y_t, \hat{y}_t^1 \neq y_t\} \tag{11}$$

$$= \Pr\{\hat{y}_t^2 = y_t, \hat{y}_t^1 \neq y_t\} \text{ from assumption (8)} \tag{12}$$

$$= \Pr\{\hat{y}_t^1 \neq y_t\} \Pr\{\hat{y}_t^2 = y_t | \hat{y}_t^1 \neq y_t\} \tag{13}$$

$$= \Pr\{\hat{y}_t^1 \neq y_t\} \left(1 - \Pr\{\hat{y}_t^2 \neq y_t | \hat{y}_t^1 \neq y_t\}\right) \tag{14}$$

$$= \Pr\{\hat{y}_t^1 \neq y_t\} \left(1 - \frac{\Pr\{\hat{y}_t^2 \neq y_t, \hat{y}_t^1 \neq y_t\}}{\Pr\{\hat{y}_t^1 \neq y_t\}}\right) \tag{15}$$

$$= \Pr\{\hat{y}_t^1 \neq y_t\} - \Pr\{\hat{y}_t^2 \neq y_t\} \text{ by contrapositve of (8)} \tag{16}$$

Proof of Theorem 1:

First consider the case $\mu < c$, i.e., action a_1 is optimal. Suppose that in round t the condition

$$\hat{\mu}_t + \sqrt{\frac{6 \log t}{N_2(t-1)}} > c \tag{17}$$

holds and sub-optimal arm is played. This implies that one of the following must hold:

$$\hat{\mu}_t - \frac{\mu}{2} + \frac{1}{2}\sqrt{\frac{6 \log t}{N_2(t-1)}} > \frac{c}{2} \tag{18}$$

$$\frac{\mu}{2} + \frac{1}{2}\sqrt{\frac{6 \log t}{N_2(t-1)}} > \frac{c}{2}, \tag{19}$$

otherwise condition (17) is violated. Condition (19) bounds the number of plays of action a_2 as

$$N_2(t-1) \leq \frac{6\log t}{\Delta_2^2}. \tag{20}$$

Rest of the proof follows exactly as in the proof of UCB1 algorithm with minor modifications. We repeat it here for completeness. Conditions (18)- (20) implies that sub-optimal action a_2 is not selected in round t whenever $N(t-1) > u := \frac{6\log t}{\Delta_2^2} + 1$, or if a_2 is selected, (18) must hold. We thus have

$$\begin{aligned} N_2(T) &= \sum_{t=1}^{T} \mathbf{1}\{I_t = 2\} && (21)\\ &\leq \; u + \sum_{t=u+1}^{T} \mathbf{1}\{(18) \text{ holds} : N_2(t) > u\} && (22)\\ &\leq \; u + \sum_{t=u+1}^{T} \mathbf{1}\{(18) \text{ holds}\}. && (23) \end{aligned}$$

Now, it remains to to bound the probability of event (18). Note that event (18) satisfies

$$\left\{\hat{\mu}_t - \frac{\mu}{2} + \frac{1}{2}\sqrt{\frac{6\log t}{N_2(t-1)}} > \frac{c}{2}\right\} \subset \left\{\hat{\mu}_t - \mu + \frac{1}{2}\sqrt{\frac{6\log t}{N_2(t-1)}} > 0\right\} \tag{24}$$

Using union bound we get

$$\begin{aligned} \Pr\{(18 \text{ holds})\} &\leq \Pr\left\{\exists s \in \{1,2,\cdots,t\} \;\; \hat{\mu}_t - \frac{\mu}{2} + \frac{1}{2}\sqrt{\frac{6\log t}{s}} > \frac{c}{2}\right\} && (25)\\ &\leq \Pr\left\{\exists s \in \{1,2,\cdots,t\} \;\; \hat{\mu}_t - \mu > -\frac{1}{2}\sqrt{\frac{6\log t}{s}}\right\} && (26)\\ &\leq \sum_{s=1}^{t} \Pr\left\{\hat{\mu}_t - \mu > -\frac{1}{2}\sqrt{\frac{6\log t}{s}}\right\} && (27)\\ &\leq \sum_{s=1}^{t} \exp\{-\frac{12s\log t}{4s}\} = 1/t^2. && (28) \end{aligned}$$

Thus taking expectation in (23) we get

$$\mathbb{E}[N_2(T)] \leq \frac{6\log T}{\Delta_2^2} + \sum_{t=1}^{T} 1/t^2 \leq \frac{6\log T}{\Delta_2^2} + \pi^2/6 + 1. \tag{29}$$

Regret bound now follows by noting that $R_T(StAT, a^*) = \Delta_2 \mathbb{E}[N_2(T)]$.

Next consider the case $\mu > c$, i.e., action a_2 is optimal. Let the sub-optimal action a_1 is played in round t, i.e,

$$\hat{\mu}_t + \sqrt{\frac{6\log t}{N_2(t-1)}} \leq c, \tag{30}$$

We bound the probability of (30) as follows:

$$\begin{aligned}
\Pr\{(30) \text{ holds}\} &\leq \Pr\left\{\exists s \in \{1,2,\cdots,t\} \;\; \hat{\mu}_t + \sqrt{\frac{6\log t}{s}} \leq c\right\} && (31)\\
&\leq \sum_{s=1}^{t} \Pr\left\{\hat{\mu}_t + \sqrt{\frac{6\log t}{s}} \leq c\right\} && (32)\\
&= \sum_{s=1}^{t} \Pr\left\{\hat{\mu}_t - \mu \leq (c-\mu) - \sqrt{\frac{6\log t}{s}}\right\} && (33)\\
&\leq \sum_{s=1}^{t} \Pr\left\{\hat{\mu}_t - \mu \leq -\sqrt{\frac{6\log t}{s}}\right\} \;\; \text{as} \;\; c < \mu && (34)\\
&\leq t/t^{12} && (35)\\
& && (36)
\end{aligned}$$

Thus, the expected number of pulls of the sub-optimal arms is bounded as

$$\begin{aligned}
\mathbb{E}[N_1(T)] &= \sum_{t=1}^{T} \Pr\{I_t = a_2\} && (37)\\
&= \sum_{t=1}^{T} 1/t^{11} \leq \pi^2/6. && (38)
\end{aligned}$$

Corollary 1: Let the dominance assumption (8) holds. For any $c \in [0,1]$, the expected regret of StAT is bounded as follows:

$$R_T(StAT, a^*) \leq \sqrt{T(6\log T + \frac{\pi^2}{6} + 1)}. \tag{39}$$

Proof: From (7) and Cauchy-Schwartz inequality we have

$$\begin{aligned}
\mathbb{E}[R_T(StAT, a^*)] &= \sum_{i=1}^{2} \Delta_i \sqrt{\mathbb{E}[N_i(T)]}\sqrt{\mathbb{E}[N_i(T)]} && (40)\\
&\leq \sqrt{\sum_{i=1}^{2} \Delta_i^2 \mathbb{E}[N_i(T)] \sum_{j=1}^{2} \mathbb{E}[N_j(T)]} && (41)
\end{aligned}$$

Consider the case $\mu < c$. Substituting (29) and noting that $\sum_{j=1}^{2} \mathbb{E}[N_j(T)] = T$, we get the bound. For the case $\mu > c$, the expected regret is constant.

III. Extension to Multi-Stage and Multi-Action setting

We next consider the sensor selection problem with more than two sensors. In each round the learner can select a subset of sensors. We assume that the sensors form a cascade, i.e., the order in which the sensors can be selected is predetermined. The policy of the learner is to select a set of sensors, or when to stop use of sensors in the cascade, in each round.

As before, let $\{y_t\}_{t>0}$ denote sequence of binary labels generated according to an unknown but fixed distribution $\mathcal{D}$. The learner has access to a K sensors that can be used to predict the labels. In round t, let $\hat{y}_t^k$ denote the prediction of the k^{th} sensor.

In each round t, the learner can take one of K actions denoted as $\mathcal{A} = \{a_1, \ldots, a_K\}$, where the action a_k indicates acquiring sensors $1, \ldots, k$ and classifying using the prediction $\hat{y}_t^k$. We denote the prediction error rate of the k^{th} sensor as $\gamma_k := \Pr\{y_t \neq \hat{y}_t^k\}$.

Let H_t denote the feedback observed in round t by selecting an action. When the learner selects action a_k at time t, feedback is $H(a_k) = \{\hat{y}_t^1, \ldots, \hat{y}_t^k\}$, The loss incurred in each round is defined as follows. When the learner selects action a_k, the loss is the classification error incurred from the prediction $\hat{y}_t^k$ combined with the sum of the costs $c_1, \ldots, c_k$. Let $L_t(a_k)$ denote the loss in round t for taking action a_k. Then,

$$L_t(a_k) = \mathbf{1}_{\{\hat{y}_t^k \neq y_t\}} + \sum_{j=1}^{k} c_j. \tag{42}$$

The learner uses a policy that selects an action in $\mathcal{A}$ in each round using the feedback observed in the past. The regret of a policy π that selects action $\pi_t \in \mathcal{A}$ in round t over a period T with respect to an action $a \in \{a_1, \ldots, a_K\}$ is given as

$$R_T(\pi, a) = \sum_{t=1}^{T} (L_t(\pi_t) - L_t(a)). \tag{43}$$

The goal of the learner is to learn a policy that minimizes the maximum expected regret, i.e.,

$$\pi^* = \arg\min_{\pi \in \Pi} \max_{a \in \mathcal{A}} \mathbb{E}[R_T(\pi, a)], \tag{44}$$

where Π denote the set of policies that maps past history to an action in $\mathcal{A}$ in each round.

Optimal action in hindsight: Since for any t

$$\mathbb{E}[L_t(a_k)] = \Pr\{y_t \neq \hat{y}_t^k\} + \sum_{j=1}^{k} c_j. \tag{45}$$

We denote the optimal hindsight action, a^*, as the action that minimizes the expected loss:

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}[L_t(a)] \tag{46}$$

We can now rewrite the goal of the learner as

$$\pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}[R_T(\pi, a^*)]. \tag{47}$$

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 $x_t, y_{tt>0}$, the regret of a policy π is defined as:

$$R_T(\pi) = \sum_{t=1}^{T} (L_t(\pi(x_t)) - L_t(a^*(x_t))) \tag{48}$$

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.,

$$\pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}[R_T(\pi)]. \tag{49}$$

We first consider the case with two actions studied in Section I. 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

$$\hat{y}_t^2 \neq 1 \implies \hat{y}_t^1 \neq 1, \tag{50}$$

which is equivalent to

$$\hat{y}_t^2 = 0 \implies \hat{y}_t^1 = 0. \tag{51}$$

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

Cs: However, the crowdsourcing model is also unidentifiable in a strict sense; and the rank one 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

$$\hat{y}_t^i = \text{sign}(\hat{\hat{y}}_t^i - \eta_i) \tag{52}$$

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 (6).

VII. RELATIONSHIP TO CROWDSOURCING

A basic problem formulation in crowdsourcing is the following:[1] 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 $(Y_t^*)_{1\le t\le T}$, $(\xi_{w,t,y})_{w,t,y}$ 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 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.[2]

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

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

[2]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.

$\mathrm{sgn}(\sum_{w=1}^{W} v_w Y_{w,t})$ 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.