Active and Adaptive Sequential Learning with Per Time-step Excess Risk Guarantees

Yuheng Bu* Jiaxun Lu† Venugopal V. Veeravalli*
*University of Illinois at Urbana-Champaign † Tsinghua University
Email: bu3@illinois.edu, lujx14@mails.tsinghua.edu.cn, vvv@illinois.edu

***Abstract*—We consider solving a sequence of machine learning problems that vary in a bounded manner from one time-step to the next. To solve these problems in an accurate and data-efficient way, we propose an active and adaptive learning framework, in which we actively query the labels of the most informative samples from an unlabeled data pool, and adapt to the change by utilizing the information acquired in the previous steps. Our goal is to satisfy a pre-specified bound on the excess risk at each time-step. We first design the active querying algorithm by minimizing the excess risk using stochastic gradient descent in the maximum likelihood estimation setting. Then, we propose a sample size selection rule that minimizes the number of samples by adapting to the change in the learning problems, while satisfying the required bound on excess risk at each time-step. Based on the actively queried samples, we construct an estimator for the change in the learning problems, which we prove to be an asymptotically tight upper bound of its true value. We validate our algorithm and theory through experiments with real data.**

I. Introduction

Machine learning problems that vary in a bounded manner over time naturally arise in many applications. For example, in recommendation systems [1], the preferences of users might change with fashion trends. Since acquiring new training samples from users can be expensive in practice, a recommendation system needs to update the machine learning model and adapt to this change using as few new samples as possible.

In such problems, we are given a large set of unlabeled samples at each time t, and the learning tasks are solved by minimizing the expected value of an appropriate loss function on this unlabeled data pool. To capture the idea that the sequence of learning problems is changing in a bounded manner, we assume the following bound holds

$$\|\theta_t^* - \theta_{t-1}^*\|_2 \leq \rho, \quad \forall t \geq 2, \tag{1}$$

where θ_t^* is the true minimizer of the expected loss function at time t, and ρ is a finite upper bound on the change of minimizers, which needs to be estimated in practice.

To tackle this sequential learning problem, we design an *active* and *adaptive* algorithm to learn the approximate minimizers $\hat{\theta}_t$ of the loss function. At each time t, the algorithm actively queries the labels of K_t samples from the unlabeled data pool, with an appropriately designed active sampling distribution, which is adaptive to the change in the minimizers by utilizing the information acquired in the previous steps.

Our contributions in this paper can be summarized as follows. We propose an active and adaptive learning framework with theoretical guarantees to solve a sequence of learning problems in the maximum likelihood estimation (MLE) setting. The proposed algorithm ensures a bounded excess risk for each individual learning task when t is sufficiently large. We construct a new estimator of the change in the minimizers $\hat{\rho}_t$ with active learning samples and show that this estimate upper bounds the true parameter ρ almost surely. We apply our approaches in a recommendation system to track the changes in preferences of customers. Our experiments demonstrate that compared to the other baseline algorithms, the proposed algorithm achieves a better accuracy performance while being efficient in the use of training samples.

A. Related Work

The setting of our active and adaptive learning problem is similar to *online learning*, where a sequence of learning tasks arrive, and the goal is to minimize the regret over some large time horizon [2]. Thus, the theoretical guarantee of online learning is different from the per time-step excess risk guarantee provided in this paper.

Our work has relations with *active learning* [3], in which a learning algorithm is able to interactively query the labels of samples from an unlabeled data pool to achieve better performance. A standard approach to active learning is to select the unlabeled samples by optimizing specific statistics of these samples [4]. For example, with the goal of minimizing the expected excess risk in maximum likelihood estimation, the authors of [5], [6] propose a two-stage algorithm based on the Fisher information ratio to select the most informative samples, and show that it is optimal in terms of convergence rate. We apply similar algorithms in our problem, but the first stage of estimating the Fisher information using labeled samples to conduct active learning can be skipped by exploiting the bounded nature of the change, and utilizing information obtained in previous time-steps.

Our approach is closely related to prior work on adaptive sequential learning [7], [8], where the training samples are drawn passively and the adaptation is only in the selection of the number of training samples K_t at each time-step.

II. Problem Definition and Setting

Throughout this paper, we use lower-case letters to denote scalars and vectors, and use-upper case letters to denote random variables and matrices. We use I to denote an identity

978-1-7281-4300-2/19/$31.00 ©2019 IEEE

matrix of appropriate size. We use the superscript $(\cdot)^\top$ to denote the transpose of a vector or a matrix, and $\mathrm{Tr}(A)$ to denote the trace of a square matrix A. We use $\|x\|_A$ to denote $\sqrt{x^\top A x}$ for a vector x and a matrix A of appropriate dimensions.

We consider the active and adaptive sequential learning problem in the MLE setting. At each time t, we are given a pool $\mathcal{S}_t = \{x_{1,t}, \cdots, x_{N,t}\}$ of N_t unlabeled samples drawn from some instance space $\mathcal{X}$. We have the ability to interactively query the labels of K_t of these samples from a label space $\mathcal{Y}$. In addition, we are given a parameterized family of distribution models $\mathcal{M} = \{p(y|x,\theta_t), \theta_t \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$. We assume that there exists an unknown parameter $\theta_t^* \in \Theta$ such that the label y_t of $x_t \in \mathcal{S}_t$ is actually generated from the distribution $p(y_t|x_t,\theta_t^*)$.

For any $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\theta \in \Theta$, we let the loss function be the negative log-likelihood with parameter θ, i.e.,

$$\ell(y|x,\theta) \triangleq -\log p(y|x,\theta), \quad p(y|x,\theta) \in \mathcal{M}. \tag{2}$$

Then, the expected loss function over the uniform distribution on the data pool $\mathcal{S}_t$ can be written as

$$L_{U_t}(\theta) \triangleq \mathbb{E}_{X\sim U_t, Y\sim p(Y|X,\theta_t^*)}[\ell(Y|X,\theta)], \tag{3}$$

where we use U_t to denote the uniform distribution over the samples in $\mathcal{S}_t$. It can be seen that θ_t^* is one of the minimizers of $L_{U_t}(\theta)$. As in (1), we assume that θ_t^* changes at a bounded but unknown rate, $\|\theta_t^* - \theta_{t-1}^*\|_2 \le \rho$, for $t \ge 2$.

The quality of the algorithm outputs $\hat{\theta}_t$ are evaluated through an *excess risk criterion*, which means that the excess risk of $\hat{\theta}_t$ is bounded at each time-step t, i.e.,

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)] \le \varepsilon. \tag{4}$$

Thus, our goal is to actively and adaptively select the smallest number of K_t samples in $\mathcal{S}_t$ to query labels, and sequentially construct an estimate of $\hat{\theta}_t$ satisfying the above excess risk criterion for each time-step t. Note that it is allowed to query the label of the same sample multiple times.

Let Γ_t be an arbitrary sampling distribution on $\mathcal{S}_t$, and

$$\hat{\theta}_{\Gamma_t} \triangleq \arg\min_{\theta\in\Theta} \frac{1}{K_t}\sum_{k=1}^{K_t} \ell(Y_{k,t}|X_{k,t},\theta), \tag{5}$$

where $X_{k,t} \sim \Gamma_t$, $Y_{k,t} \sim p(Y|X_{k,t},\theta_t^*)$.

III. Algorithm Outline

We first provide an outline of the proposed active and adaptive sequential learning algorithm. Our algorithm consists of the following four steps, the technical details of which can be found in Section IV.

1) Construct active learning sampling distribution $\hat{\Gamma}_t^*$ based on the estimation acquired in the previous step $\hat{\theta}_{t-1}$, which queries the labels of the most informative samples (see Section IV-B).
2) Adaptively choose the minimal sample size K_t^* based on the estimated change in minimizers $\hat{\rho}_{t-1}$ to satisfy the excess risk criterion (see Section IV-C).

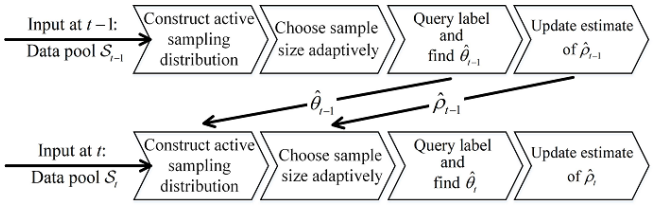


Fig. 1. Active and adaptive sequential learning framework.

3) Query the labels of K_t^* samples over the unlabeled data pool S_t using $\hat{\Gamma}_t^*$, and estimate $\hat{\theta}_t$ by solving (5).
4) Update the estimate of change rate $\hat{\rho}_t$ by using the actively labeled samples (see Section IV-D).

By executing this procedure iteratively, we can sequentially learn $\hat{\theta}_t$ over the considered time-steps. Fig. 1 illustrates our active and adaptive sequential learning framework.

IV. Analyses and Theoretical Guarantees

In this section, we present technical details and the theoretical analysis of our algorithm. We first introduce the assumptions needed. The proofs of the theorems and lemmas are presented in [9].

A. Assumptions

We require the following assumption on the Hessian matrix of $\ell(y|x,\theta)$ to design the active sampling distribution over the unlabeled data pool $\mathcal{S}_t$.

Assumption 1. *For any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\theta \in \Theta$, $H(x,\theta) \triangleq \frac{\partial^2 \ell(y|x,\theta)}{\partial \theta^2}$ is a function of only x and θ and independent on y.*

Assumption 1 holds for many practical models, such as generalized linear model, logistic regression and conditional random fields. Moreover, we denote The Fisher information matrix with sampling distribution Γ_t is given by: $I_{\Gamma_t}(\theta) \triangleq \mathbb{E}_{X\sim\Gamma_t}[H(X,\theta)]$.

The following regularity assumptions are required to establish the Local Asymptotic Normality of the MLE (see [10]).

Assumption 2 (Regularity conditions)**.**

1) ***Regularity conditions for MLE:***
   a) ***Strong Convexity****: For each t and $\theta \in \Theta$, $I_{U_t}(\theta) \succeq mI$ with $m > 0$, and hence $I_{U_t}(\theta)$ is positive definite.*
   b) ***Boundedness****: For all $\theta \in \Theta$, the largest eigenvalue of $I_{U_t}(\theta)$ is upper bounded by L_b.*
2) ***Concentration at θ_t^*****: For all t, and any $x_t \in \mathcal{S}_t$, $y_t \in \mathcal{Y}$,*

$$\begin{aligned} &\left\|\nabla \ell(y_t|x_t,\theta_t^*)\right\|_{I_{U_t}(\theta_t^*)^{-1}} \le L_1 \quad \textit{and} \\ &\left\|I_{U_t}(\theta_t^*)^{-1/2} H(x,\theta_t^*) I_{U_t}(\theta_t^*)^{-1/2}\right\| \le L_2 \end{aligned} \tag{6}$$

   holds with probability one.
3) ***Lipschitz continuity****: For all t, there exists a neighborhood B_t of θ_t^* and a constant L_3, such that for all $x_t \in \mathcal{S}_t$, $H(x_t,\theta)$ are L_3-Lipschitz in this neighborhood, namely,*

$$\begin{aligned} &\left\|I_{U_t}(\theta_t^*)^{-1/2}\big(H(x_t,\theta) - H(x_t,\theta')\big) I_{U_t}(\theta_t^*)^{-1/2}\right\| \\ &\qquad \le L_3 \|\theta - \theta'\|_{I_{U_t}(\theta_t^*)} \end{aligned} \tag{7}$$

holds for $\theta, \theta' \in B_t$.

In addition, we need the following assumption to prove that constructing the active sampling distribution using $\hat{\theta}_{t-1}$ instead of θ_t^* does not change the performance of the active learning algorithm in terms of the convergence rate.

Assumption 3 (Point-wise self-concordance)**.** *For all t, there exists a constant L_4, such that*

$$\begin{aligned} -L_4\|\theta_t - \theta_t^*\|_2 H(x, \theta_t^*) &\preceq H(x, \theta_t) - H(x, \theta_t^*) \\ &\preceq L_4\|\theta_t - \theta_t^*\|_2 H(x, \theta_t^*). \end{aligned} \tag{8}$$

This assumption is satisfied by many classes of models, e.g., the previously mentioned generalized linear model [5].

B. Active Sampling Distribution

The construction of Γ_t is motivated by the following lemma, which is a refinement of a similar result given in [11] and [5].

Lemma 1. *Suppose Assumptions 1 and 2 hold, and assume $\Theta_t \triangleq \{\theta_t | \|\theta_t - \theta_{t-1}^*\| \leq \rho\}$ is known. For any sampling distribution Γ_t on $\mathcal{S}_t$, suppose that $I_{\Gamma_t}(\theta_t^*) \succeq C I_{U_t}(\theta_t^*)$ holds for some constant $C < 1$. Then, for sufficiently large K_t, such that $\gamma_t \triangleq \mathcal{O}\big(\frac{1}{C^2}(L_1 L_3 + \sqrt{L_2})\sqrt{\frac{\log dK_t}{K_t}}\big) < 1$, the excess risk of $\hat{\theta}_{\Gamma_t}$ can be bounded as*

$$\begin{aligned} (1-\gamma_t)\frac{\tau_t^2}{K_t} - \frac{L_1^2}{CK_t^2} \leq& \mathbb{E}[L_{U_t}(\hat{\theta}_{\Gamma_t}) - L_{U_t}(\theta_t^*)] \\ &\leq (1+\gamma_t)\frac{\tau_t^2}{K_t} + \frac{2L_b\rho^2}{K_t^2} \end{aligned} \tag{9}$$

for all t, where $\tau_t^2 \triangleq \frac{1}{2}\mathrm{Tr}\big(I_{\Gamma_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big)$.

Lemma 1 shows that when ρ and θ_{t-1}^* are known, the convergence rate of the excess risk for $\hat{\theta}_{\Gamma_t}$ defined in (5) is $\mathrm{Tr}(I_{\Gamma_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*))/K_t$. Thus, the optimal sampling distribution Γ_t^* should be

$$\Gamma_t^* = \arg\min_{\Gamma_t} \mathrm{Tr}(I_{\Gamma_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)). \tag{10}$$

However, the true parameter θ_t^* in (10) is unknown, and hence we cannot solve Γ_t^* directly. Exploiting the bounded nature of the change in (1), we solve this problem by approximating θ_t^* with $\hat{\theta}_{t-1}$ and generate the estimate of Γ_t^* using

$$\hat{\Gamma}_t^* = \arg\min_{\Gamma_t} \mathrm{Tr}(I_{\Gamma_t}^{-1}(\hat{\theta}_{t-1}) I_{U_t}(\hat{\theta}_{t-1})). \tag{11}$$

Note that $\hat{\Gamma}_t^*$ may not have the full support of $\mathcal{S}_t$, which reduces the sampling diversity and further leads to biased estimates. Thus, we modify the active sampling distribution slightly by

$$\bar{\Gamma}_t = \alpha_t \hat{\Gamma}_t^* + (1-\alpha_t) U_t, \tag{12}$$

where $\alpha_t \in (0,1)$ is chosen via cross-validation.

Another issue is that Lemma 1 only characterizes the convergence rate for $\hat{\theta}_{\Gamma_t}$ without considering the error caused by optimization algorithm. In practice, we usually apply stochastic optimization algorithms, such as stochastic gradient descent (SGD) to find approximate minimizers in the original parameter space Θ. For the purpose of bounding the excess risk of the solution provided by SGD, we require the following condition on the optimization algorithm adopted to solve (5).

Condition 1. *Given an optimization algorithm that generates an approximate loss minimizer*

$$\hat{\theta}_t \triangleq \mathscr{A}\big(\hat{\theta}_{t-1},\ \{\nabla_\theta \ell(y_{k,t}|x_{k,t},\theta)\}_{k=1}^{K_t}\big)$$

using K_t stochastic gradients $\{\nabla_\theta \ell(y_{i,t}|x_{i,t},\theta)\}_{k=1}^{K_t}$ with initialization at $\hat{\theta}_{t-1}$, if $\mathbb{E}\|\hat{\theta}_{t-1} - \theta_t^*\|_2^2 \leq \Delta_t^2$, there exists a function $b(\tau_t^2, \Delta_t, K_t)$ such that

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t)] - L_{U_t}(\theta_t^*) \leq b(\tau_t^2, \Delta_t, K_t), \tag{13}$$

where $b(\tau_t^2, \Delta_t, K_t)$ increases monotonically with respect to τ_t^2, Δ_t and $1/K_t$.

The bound $b(\tau_t^2, \Delta_t, K_t)$ depends on the expectation of the difference between the initialization and the true minimizer Δ_t. As an example for this type of bound, for the Streaming Stochastic Variance Reduced Gradient (Streaming SVRG) algorithm in [11], it holds that $b(\tau_t^2, \Delta_t, K_t) = C_1\frac{\tau_t^2}{K_t} + C_2\big(\frac{\Delta_t}{K_t}\big)^2$ with constant C_1 and C_2. In addition, several examples of the bound $b(\tau_t^2, \Delta_t, K_t)$ with other variations of SGD algorithm are given in [8].

We have the following theorem characterizes the convergence rate of the proposed active learning algorithm.

Theorem 1. *Suppose Assumptions 1, 2 and 3 hold, and let $\beta_t \triangleq L_4(\rho + \frac{1}{\delta}\sqrt{\frac{2\varepsilon}{m}}) < 1$. Then, by using the active sampling distribution given in* (12)*, the excess risk of $\hat{\theta}_t$ which obtained by solving the optimization algorithm satisfying Condition 1 initialized at $\hat{\theta}_{t-1}$ is upper-bounded by*

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)] \leq b(\acute{\tau}_t^2, \Delta_t, K_t), \tag{14}$$

with probability 1-δ, where

$$\acute{\tau}_t^2 = \Big(\frac{1+\beta_t}{1-\beta_t}\Big)^2 \frac{\mathrm{Tr}\big(I_{\Gamma_t^*}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big)}{2\alpha_t}, \quad \Delta_t = \sqrt{\frac{2\varepsilon}{m}} + \rho, \tag{15}$$

$\delta \in (0,1)$ and Γ_t^* is the optimal sampling distribution minimizing $\mathrm{Tr}\big(I_{\Gamma_t^*}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big)$.

C. Sample Size Selection Rule

1) Case where ρ is known: We first consider the ideal case where ρ is known. If we can compute τ_t^2 and Δ_t, the sample size K_t can be simply determined by setting $b(\tau_t^2, \Delta_t, K_t)$ in Condition 1 to ε to satisfy the excess risk criterion.

However, θ_t^* in $\tau_t^2 = \frac{1}{2}\mathrm{Tr}\big(I_{\Gamma_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big)$ is unknown in practice. Thus, we use the fact that

$$\mathrm{Tr}\big(I_{\Gamma_t^*}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big) \leq \mathrm{Tr}\big(I_{U_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big) = d, \tag{16}$$

(recall d is the dimension of parameters) to get a conservative bound $b(d/2, \Delta_t, K_t)$ to choose K_t, which works for the uniform sampling distribution U_t.

To bound the difference between the initialization and the true minimizer Δ_t, we have the inequality $\mathbb{E}\|\hat{\theta}_{t-1} - \theta_t^*\|_2^2 \leq (\sqrt{2\varepsilon/m} + \rho)^2$ following from the triangle inequality, Jensen's

inequality and the strong convexity in Assumption 2. This inequality implies that $\Delta_t = \sqrt{2\varepsilon/m} + \rho$.

Therefore, if ρ is known, we can set

$$K_t^* = \min\Big\{K \geq 1 \Big| b\Big(d/2, \sqrt{\frac{2\varepsilon}{m}} + \rho, K\Big) \leq \varepsilon\Big\}, \quad (17)$$

for $t \geq 2$ to ensure that $\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)] \leq \varepsilon$.

2) Case where ρ is unknown: In this case, we can replace ρ with its estimate $\hat{\rho}_{t-1}$ to select the sample size. The following theorem characterizes the convergence guarantee using the sample size selection rule in Algorithm 1 and the estimator of $\hat{\rho}_t$ in Section IV-D.

Theorem 2. *If we choose*

$$K_t \geq K_t^* \triangleq \min\Big\{K \geq 1 \Big| b\Big(d/2, \sqrt{\frac{2\varepsilon}{m}} + \hat{\rho}_{t-1}, K\Big) \leq \varepsilon\Big\},$$

then $\limsup_{t\to\infty} \big(\mathbb{E}[L_{U_t}(\hat{\theta}_t)] - L_{U_t}(\theta_t^*)\big) \leq \varepsilon$ *almost surely.*

D. Estimating the Change in Minimizers

1) Estimating One-Step Change: As a consequence of strong convexity, the following lemma holds.

Lemma 2. *Suppose Assumption 2 holds, then*

$$\|\theta_{t-1}^* - \theta_t^*\|^2 \leq \frac{1}{m}\big[L_{U_t}(\theta_{t-1}^*) - L_{U_t}(\theta_t^*) + L_{U_{t-1}}(\theta_t^*) - L_{U_{t-1}}(\theta_{t-1}^*)\big]. \quad (18)$$

Motivated by Lemma 2, we can construct the following one-step estimation of ρ^2

$$\widetilde{\rho}_t^2 = \frac{1}{m}\big[\hat{L}_{U_t}(\hat{\theta}_{t-1}) - \hat{L}_{U_t}(\hat{\theta}_t) + \hat{L}_{U_{t-1}}(\hat{\theta}_t) - \hat{L}_{U_{t-1}}(\hat{\theta}_{t-1})\big],$$

where

$$\hat{L}_{U_t}(\hat{\theta}_{t-1}) \triangleq \frac{1}{K_t}\sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t}, \hat{\theta}_{t-1})}{N_t \bar{\Gamma}_t(X_{k,t})}. \quad (19)$$

Note that we are using the samples generated from the active learning distribution, i.e., $X_{k,t} \sim \bar{\Gamma}_t$ and $Y_{k,t} \sim p(Y|X_{k,t}, \theta_t^*)$. Thus, based on the idea of importance sampling, we normalize the estimate with the sampling distribution $\bar{\Gamma}_t$.

2) Combining One-Step Estimates: Then, we combine the one-step estimates to construct an overall estimate by using a class of window functions $h_W : \mathbb{R}^W \to \mathbb{R}$ that are non-decreasing in their arguments and satisfy $\mathbb{E}[h_W(\rho_j, \cdots, \rho_{j-W+1})] \geq \rho$. For example, $h_W(\rho_j, \cdots, \rho_{j-W+1}) = \frac{W+1}{W}\max\{\rho_j, \cdots, \rho_{j-W+1}\}$ satisfies the requirements. The combined estimate of ρ_t^2 is computed by applying the function h_W to a sliding window of one-step estimates of $\widetilde{\rho}^2$, i.e.,

$$\rho_t^2 = \frac{1}{t-1}\sum_{j=2}^{t} h_{\{\min[W,j-1]\}}(\widetilde{\rho}_j^2, \widetilde{\rho}_{j-1}^2, \cdots, \widetilde{\rho}_{\max[j-W+1,2]}^2).$$

The following theorem characterizes the performance of proposed estimator.

Theorem 3. *Suppose Assumptions 1 and 2 hold, and there exists a sequence* $\{r_t\}$ [1] *satisfying*

$$\sum_{t=1}^{\infty} \exp\Big\{-\frac{2m^2(t-1)r_t^2}{9L_b^2 \text{Diameter}^4(\Theta)}\Big\} < \infty.$$

Then, for all t large enough, $\hat{\rho}_t^2 \triangleq \hat{\rho}_t^2 + D_t + r_t \geq \rho^2$ almost surely with constant D_t specified in [9].

E. Algorithm

Our active and adaptive sequential learning algorithm is formally presented in Algorithm 1.

Algorithm 1 Active and Adaptive Sequential Learning

Input: Sample pool $\mathcal{S}_t = \{x_{1,t}, \cdots, x_{N,t}\}$, previous estimates $\hat{\theta}_{t-1}$, $\hat{\rho}_{t-1}$ and the pre-specified excess risk ε.

1: Solve the following semi-definite programming problem

$$\hat{\Gamma}_t^* = \underset{\Gamma_t \in \mathbb{R}^{N_t}}{\arg\min} \ \text{Tr}[I_{\Gamma_t}^{-1}(\hat{\theta}_{t-1}) I_{U_t}(\hat{\theta}_{t-1})]$$
$$\text{s.t.} \begin{cases} I_{\Gamma_t}(\hat{\theta}_{t-1}) = \sum_{i=1}^{N_t} \Gamma_{i,t} H(x_{i,t}, \hat{\theta}_{t-1}), \\ \sum_{i=1}^{N_t} \Gamma_{i,t} = 1, \ \Gamma_{i,t} \in [0,1]. \end{cases}$$

2: Choose K_t^* based on $\hat{\rho}_{t-1}$ such that it is the minimum number of samples required to meet the excess risk criterion.

3: Generate K_t^* samples using the distribution $\bar{\Gamma}_t = \alpha_t \hat{\Gamma}_t^* + (1-\alpha_t)U_t$ on unlabeled data pool $\mathcal{S}_t$, where $\alpha_t \in (0,1)$. Query their labels and get the labeled set $\mathcal{S}_t' = \{(x_{k,t}, y_{k,t})\}_{k=1}^{K_t^*}$.

4: Solve the MLE using labeled set S_t' with a SGD algorithm initialized at $\hat{\theta}_{t-1}$,

$$\hat{\theta}_t = \underset{\theta_t \in \Theta}{\arg\min} \sum_{(x_{k,t}, y_{k,t}) \in \mathcal{S}_t'} \ell(y_{k,t}|x_{k,t}, \theta_t).$$

5: Update the estimate of $\hat{\rho}_t$ using estimator defined in Theorem 3 for $\forall t \geq 2$.

Output: $\hat{\theta}_t$, $\hat{\rho}_t$.

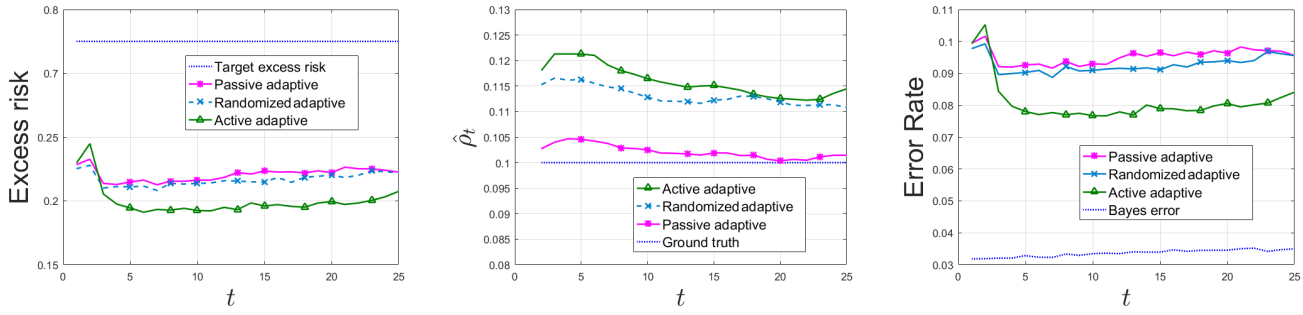
In Step 1, the active sampling distribution is constructed via solving a semi-definite programming (SDP) problem. Then, we use the minimum sample size K_t^* such that the excess risk criterion is satisfied, and actively draw samples from $\bar{\Gamma}_t$ to estimate $\hat{\theta}_t$ (Steps 2-4). As stated earlier, the distribution $\hat{\Gamma}_t^*$ is modified slightly to $\bar{\Gamma}_t$ in Step 3 to ensure it still has the full support of $\mathcal{S}_t$. Finally, based on the current and previous estimates $\hat{\theta}_t$ and $\hat{\theta}_{t-1}$, we update the estimate of the bounded change rate $\hat{\rho}_t$ using the result in in Theorem 3.

V. Experiments

In this section, we utilize a subset of Yelp 2017 dataset [2] to perform our experiments to validate our algorithm and the related theoretical results. We select the users that have at least 10 ratings from the original dataset and construct the dataset for this experiment. Our dataset contains ratings of $M = 473$

[1] Note that a choice of r_t that is greater than $1/\sqrt{t-1}$ in the order sense works here.

[2] https://www.yelp.com/dataset



(a) (b) (c)

Fig. 2. Experiments on user preference tracking: (a) Excess risk. (b) Estimated rate of change of minimizers. (c) Classification error.

users for $N = 858$ businesses. By converting the original 5-scale ratings to a binary label for all businesses with high ratings (4 and 5) as positive (1) and low ratings (3 and below) as negative (-1), we form the $N \times M$ binary rating matrix R, which is sparse and only 2.6% are observed. We complete the sparse matrix R to make recommendations by using the matrix factorization method [12]. The rating matrix R can be modeled by the following logistic regression model

$$p(R_{u,b}|\phi_b, \phi_u) = \frac{1}{1 + \exp^{-R_{u,b}\phi_u^\top \phi_b}}, \tag{20}$$

where ϕ_u and ϕ_b are the d-dimensional latent vectors representing the preferences of user u and properties of business b, respectively. Then, we train ϕ_u and ϕ_b with dimension $d = 10$ for each user and business in the dataset using maximum likelihood estimation by SGD. With the learned latent vectors, we can complete the matrix R and make recommendations to customers in a collaborative filtering fashion [1].

In practice, the preferences of users $\phi_{u,t}$ may vary with time t, and hence user features need to be retrained. Considering the fact that acquiring new ratings of users can be expensive, we apply our active and adaptive learning algorithm to reduce the number of new samples while maintaining a desired level of accuracy.

In the following experiment, we use a random subset of $\{\phi_b\}$ with size $N_t = 400$ as our unlabeled data pool, while the remaining serve as a test set to evaluate the algorithms. To model the bounded time-varying changes of user preferences $\phi_{u,t}$, we start from a randomly chosen user feature and update it by adding a random Gaussian drift with norm bounded by 0.1 at each time-step. Since we are unable to retrieve the actual answer from a real user, we generate the labels with the probabilistic model given by (20) with true parameter $\phi_{u,t}$. The target excess risk $\varepsilon = 0.75$ is set by cross-validation, which ensures the error rate (percentage of errors on the test set) is smaller than 10%.

We compare the proposed active and adaptive algorithm in Algorithm 1 with two other algorithms: the randomized adaptive algorithm, and the passive adaptive algorithm. The randomized adaptive algorithm is different from Algorithm 1 in that the active sampling distribution is constructed with a random point in Θ instead of the estimate in the previous time-step $\hat{\theta}_{t-1}$. The passive adaptive algorithm uses a uniform sampling distribution in place of the active sampling distribution. All the reported results are averaged over 1000 runs of Monte Carlo trials, and the number of time-steps considered is 25. We set $K_t = K_t^*$ for all the algorithms and use the estimator defined in Theorem 3 with window size $W = 3$ to estimate ρ.

Fig. 2(b) shows that $\hat{\rho}_t$ converges to a conservative estimate of ρ, and the corresponding sample size converges to $K_t^* = 26$ after two time-steps. Fig. 2(a) and Fig. 2(c) show that the proposed active and adaptive learning algorithm achieves a error rate of 8% with these samples and significantly outperforms the other algorithms.

ACKNOWLEDGMENT

This work was supported by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (IoBT), through the University of Illinois at Urbana-Champaign.

REFERENCES

[1] N. Rubens, M. Elahi, M. Sugiyama, and D. Kaplan, "Active learning in recommender systems," in *Recommender Systems Handbook*, pp. 809–846. Springer, 2015.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, Cambridge University Press, 2006.

[3] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 235–242.

[4] J. A Cornell, *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, John Wiley & Sons, 2011.

[5] K. Chaudhuri, S. M Kakade, P. Netrapalli, and S. Sanghavi, "Convergence rates of active learning for maximum likelihood estimation," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 1090–1098.

[6] J. Sourati, M. Akcakaya, T. K Leen, D. Erdogmus, and J. G Dy, "Asymptotic analysis of objectives based on fisher information in active learning," *J. Mach. Learn. Res.*, vol. 18, no. 34, pp. 1–41, 2017.

[7] C. Wilson and V. V Veeravalli, "Adaptive sequential optimization with applications to machine learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2642–2646.

[8] C. Wilson, V. V Veeravalli, and A. Nedich, "Adaptive sequential stochastic optimization," *IEEE Trans. Automat. Contr.*, 2018.

[9] Y. Bu, J. Lu, and V. V Veeravalli, "Active and adaptive sequential learning," *arXiv preprint arXiv:1805.11710*, 2018.

[10] A. W Van der Vaart, *Asymptotic Statistics*, Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2000.

[11] R. Frostig, R. Ge, S. M Kakade, and A. Sidford, "Competing with the empirical risk minimizer in a single pass," in *Proc. Annual Conference on Learning Theory (COLT)*, 2015, pp. 728–763.

[12] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.