Active and Adaptive Sequential learning

Yuheng Bu [*†] **Jiaxun Lu** [*‡] **Venugopal V. Veeravalli** [†]

Abstract

A framework is introduced for actively and adaptively solving a sequence of machine learning problems, which are changing in bounded manner from one time step to the next. An algorithm is developed that actively queries the labels of the most informative samples from an unlabeled data pool, and that adapts to the change by utilizing the information acquired in the previous steps. Our analysis shows that the proposed active learning algorithm based on stochastic gradient descent achieves a near-optimal excess risk performance for maximum likelihood estimation. Furthermore, an estimator of the change in the learning problems using the active learning samples is constructed, which provides an adaptive sample size selection rule that guarantees the excess risk is bounded for sufficiently large number of time steps. Experiments with synthetic and real data are presented to validate our algorithm and theoretical results.

1 Introduction

Machine learning problems that vary in a bounded manner over time naturally arise in many applications. For example, in personalized recommendation systems [9, 15], 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, and the learning tasks are solved by minimizing the expected value of an appropriate loss function on this unlabeled data pool at each time t. 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 loss functions 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 propose an *active* and *adaptive* algorithm to learn the approximate minimizers $\hat{\theta}_t$ of the loss function. At each time t, our algorithm actively queries the labels of K_t samples from the unlabeled data pool, with a well-designed active sampling distribution, which is adaptive to the change in the minimizers by utilizing the information acquired in the previous steps. In particular, we adaptively select K_t and construct $\hat{\theta}_t$ such that the excess risk [13] is bounded at each time t.

The challenges of this active and adaptive sequential learning problem arise in three aspects: 1) we need to determine which samples are more informative for solving the task at the current time step based on the information acquired in the previous time steps to conduct active learning; 2) to

[*]Equal contribution

[†]ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA. Email: {bu3, vvv}@illinois.edu

[‡]EE Department, Tsinghua University, Beijing, China. Email: lujx14@mails.tsinghua.edu.cn

Preprint. Work in progress.

achieve a desired bounded excess risk with as few new samples as possible, we need to understand the tradeoff between the solution accuracy and the adaptively determined sample complexity K_t; 3) the change in the minimizers ρ is unknown and we need to estimate it.

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, which 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 test our approaches on a synthetic regression problem, and further apply it to a recommendation system that tracks changes in preferences of customers. Our experiments demonstrate that our algorithm achieves a better performance compared to the other baseline algorithms in these scenarios.

1.1 Related Work

Our active and adaptive learning problem has relations with *multi-task learning* (MTL) and *transfer learning*. In multi-task learning, the goal is to learn several tasks simultaneously as in [2, 10, 21] by exploiting the similarities between the tasks. In transfer learning, prior knowledge from one source task is transferred to another target task either with or without additional training data [14]. Multi-task learning could be applied to solve our problem by running a MTL algorithm at each time, while remembering all prior tasks. However, this approach incurs a heavy memory and computational burden. Transfer learning lacks the sequential nature of our problem, and there is no active learning component in both works. For multi-task and transfer learning, there are theoretical guarantees on regret for some algorithms [1], while we provide an excess risk guarantee for each individual task.

In *concept drift* problem, stream of incoming data that changes over time is observed, and we try to predict some properties of each piece of data as it arrives. After prediction, a loss is revealed as the feedback in [17]. Some approaches for concept drift use iterative algorithms such as stochastic gradient descent, but without specific models on how the data changes, there is no theoretical guarantees for these algorithms.

Our work is of course related to active learning [8, 4], 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 [7]. For example, with the goal of minimizing the expected excess risk in maximum likelihood estimation, the authors of [6, 16] propose a two-stage algorithm based on Fisher information ratio to select the most informative samples, and show that it is optimal in terms of the 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 [20, 19], 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.

The rest of the paper is organized as follows. In Section 2, we describe the problem setting considered. In Section 3, we present our active and adaptive learning algorithm. In Section 4, we provide the theoretical analysis which motivates the proposed algorithm. In Section 5, we test our algorithm on synthetic and real data. Finally, in Section 6, we provide some concluding remarks.

2 Problem 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. All logarithms are the natural ones. We use I to denote an identity matrix of appropriate size. We use the superscript $(\cdot)^\top$ to denote the transpose of a vector or a matrix, and use $\mathrm{Tr}(A)$ to denote the trace of a square matrix A. We denote $\|x\|_A = \sqrt{x^\top A x}$ for a vector x and a matrix A of appropriate dimensions.

We consider the active and adaptive sequential learning problem in the maximum likelihood estimation (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 the minimizer of $L_{U_t}(\theta)$ is the true parameter θ_t^*. As mentioned in (1), we assume that θ_t^* is changing at a bounded but unknown rate, i.e., $\|\theta_t^* - \theta_{t-1}^*\|_2 \leq \rho$, for $t \geq 2$.

The quality of our approximate minimizers $\hat{\theta}_t$ are evaluated through a *mean tracking criterion*, which means that the excess risk of $\hat{\theta}_t$ is bounded at each time step t, i.e.,

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

Thus, our goal is to actively and adaptively select the smallest number of samples K_t in $\mathcal{S}_t$ to query labels, and sequentially construct an estimate of $\hat{\theta}_t$ satisfying the above mean tracking 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 $\mathcal{S}_t$. Then, the following MLE using Γ_t

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

can be viewed as an empirical risk minimizer (ERM) of (3), where $X_{k,t} \sim \Gamma_t$, $Y_{k,t} \sim p(Y|X_{k,t}, \theta_t^*)$.

To ensure that our algorithm works correctly, we require the following assumption on the Hessian matrix of $\ell(y|x, \theta)$, which determines the Fisher information matrix.

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 does not depend on y.*

Assumption 1 holds for many practical models, such as generalized linear model, logistic regression and conditional random fields [6]. Moreover, for $\theta \in \Theta$, we denote $I_{\Gamma_t}(\theta) \triangleq \mathbb{E}_{X \sim \Gamma_t}[H(X, \theta)]$ as the Fisher information matrix under sampling distribution Γ_t.

3 Algorithm

The main idea of our algorithm is to adaptively choose the number of samples K_t based on the estimated change in the minimizers $\hat{\rho}_{t-1}$ such that the mean tracking criterion in (4) is satisfied, then actively query the labels of these K_t samples with a well-designed sampling distribution Γ_t, and finally perform MLE in (5) using a stochastic gradient descent (SGD) algorithm over the labeled samples. By executing this algorithm iteratively, we can sequentially learn $\hat{\theta}_t$ over all the considered time steps. The algorithm is formally presented in Algorithm 1.

To ensure a good performance with limited querying samples, it is essential to construct Γ_t carefully. Motivated by Lemma 1 in Section 4.2, the convergence rate of the excess risk for ERM using K_t samples from Γ_t 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 the one that minimizes $\mathrm{Tr}(I_{\Gamma_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*))$, which relies on the unknown parameter θ_t^*. Based on the bounded nature of the change in (1), we solve this problem by approximating θ_t^* with $\hat{\theta}_{t-1}$ and generate the sampling distribution $\hat{\Gamma}_t^*$ by minimizing $\mathrm{Tr}(I_{\Gamma_t}^{-1}(\hat{\theta}_{t-1}) I_{U_t}(\hat{\theta}_{t-1}))$ (Step 1).

Then, as shown in Section 4.3, we use the minimum number of samples K_t^* such that the mean tracking criterion is satisfied, and actively draw samples from $\bar{\Gamma}_t$ to estimate $\hat{\theta}_t$ (Steps 2-4). Note that the distribution $\hat{\Gamma}_t^*$ is modified slightly to $\bar{\Gamma}_t$ in Step 3 to ensure it still has the full support of $\mathcal{S}_t$.

Algorithm 1 Active and Adaptive Sequential Learning

Input: Sample pool $\mathcal{S}_t = \{x_{1,t}, \cdots, x_{N,t}\}$, the previous estimation $\hat{\theta}_{t-1}$, $\hat{\rho}_{t-1}$ and the desired mean tracking accuracy ε.

1: Solve the following semidefinite programming problem (see Section 4.2)

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

2: Choose K_t^* based on $\hat{\rho}_{t-1}$ such that it is the minimum number of samples required to meet the mean tracking criterion (see Section 4.3).

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 $\hat{\theta}_{t-1}$,

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

5: Update the estimate of $\hat{\rho}_t$ using estimator defined in Section 4.4 for $\forall t \geq 2$.

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

Finally, based on the current and previous estimation $\hat{\theta}_t$ and $\hat{\theta}_{t-1}$, we update the estimate of the bounded change rate $\hat{\rho}_t$ by the estimator proposed in Section 4.4.

It is easy to see that the active nature of Algorithm 1 comes from the active sampling distribution, which is constructed by minimizing the Fisher information ratio as in Step 1. But the adaptivity of Algorithm 1 is more complex and results from the following three aspects: 1) The sampling distribution is adaptive to the bounded change through the replacement of θ_t^* with $\hat{\theta}_{t-1}$ in Step 1; 2) The sample size selection rule is adaptive through the selection of the minimum number of samples required in Step 2; 3) The SGD is adaptive through the initialization by $\hat{\theta}_{t-1}$ in Step 4.

4 Theoretical Performance Guarantees

In this section, we present the theoretical analysis of Algorithm 1. We first introduce the assumptions needed. Then, in Section 4.2, we provide the analysis of the active sampling distribution. In Section 4.3, we present theoretical guarantees on the sample size selection rules which meet the mean tracking criterion in (4). In Section 4.4, we describe the proposed estimator $\hat{\rho}_t$. The proofs of the theorems and all the supporting lemmas will be presented in the Appendices.

4.1 Assumptions

For the purpose of analysis, the following regularity assumption on the log-likelihood function ℓ is required to establish the standard Local Asymptotic Normality of the MLE [18].

Assumption 2 (Regularity conditions)**.**

1. ***Regularity conditions for MLE:***
   - (a) ***Compactness****: Θ is compact and θ_t^* is an interior point of Θ for each t.*
   - (b) ***Smoothness****: $\ell(y|x,\theta)$ is smooth in the following sense: the first, second and third derivatives of θ exist at all interior points of Θ.*
   - (c) ***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 and invertible.*
   - (d) ***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 $x_t \in \mathcal{S}_t$, $y_t \in \mathcal{Y}$,*

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

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,*

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

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

In addition, we need the following assumption to prove that replacing θ_t^* with $\hat{\theta}_{t-1}$ in Algorithm 1 does not change the performance of the active learning algorithm in terms of the convergence rate. This assumption is satisfied by many classes of models, including the generalized linear model [6].

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

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

4.2 Optimal Active Learning Sampling Distribution

In this subsection, we provide the intuition and analysis of Step 1 in Algorithm 1. The construction of the active sampling distribution Γ_t is motivated by the following lemma, which characterizes the convergence rate of the ERM solution $\hat{\theta}_{\Gamma_t}$ defined in (5) when ρ and θ_{t-1}^* are known.

Lemma 1. *Suppose Assumptions 1 and 2 hold, and let $\Theta_t \triangleq \{\theta_t | \|\theta_t - \theta_{t-1}^*\| \leq \rho\}$. 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*

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

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

In practice, the parameter space $\Theta_t = \{\theta_t | \|\theta_t - \theta_{t-1}^*\| \leq \rho\}$ is unknown and the ERM solution of (5) cannot be obtained directly due to the computational issue. To solve these problems, we can apply optimization algorithm such as SGD to find approximate minimizers in the original parameter space Θ with initialization at $\hat{\theta}_{t-1}$. Thus, we further build Algorithm 1 and our theoretical results with the SGD algorithm (which incidentally achieves the optimal convergence rate for ERM). We need the following assumptions on the optimization algorithm to solve (5):

Assumption 4. *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*

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t)] - L_{U_t}(\theta_t^*) \leq b(\tau_t^2, \Delta_t, K_t), \tag{10}$$

where $b(\tau_t^2, \Delta_t, K_t)$ monotonically increases 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 converge rate τ_t^2 and the expectation of the difference between the initialization and the true minimizer Δ_t, which correspond to the first and the second term in the upper bound of Lemma 1, respectively. As an example for this type of bound, for the Streaming Stochastic Variance Reduced Gradient (Streaming SVRG) algorithm in [11], it holds that

$$b(\tau_t^2, \Delta_t, K_t) = C_1 \frac{\tau_t^2}{K_t} + C_2 \big(\frac{\Delta_t}{K_t}\big)^2 \tag{11}$$

with constant C_1 and C_2. In addition, the paper [20] contains several examples of the bound $b(\tau_t^2, \Delta_t, K_t)$ with other variations of SGD algorithm.

Then, the following theorem characterizes the convergence rate of the active sampling distribution used in Algorithm 1 in the order sense.

Theorem 1. *Suppose Assumptions 1-4 hold, and let* $\beta_t \triangleq L_4(\rho + \frac{1}{\delta}\sqrt{\frac{2\varepsilon}{m}}) < 1$. *Then, the excess risk of* $\hat{\theta}_t$ *in Algorithm 1 is upper-bounded by*

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)] \leq b(\acute{\tau}_t^2, \Delta_t, K_t), \tag{12}$$

with probability 1-δ, *where*

$$\acute{\tau}_t^2 = \Big(\frac{1+\beta_t}{1-\beta_t}\Big)^2 \frac{\mathrm{Tr}\big(I_{\Gamma_t^*}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big)}{2\alpha_t}, \qquad \Delta_t = \sqrt{\frac{2\varepsilon}{m}} + \rho, \tag{13}$$

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

Remark 1. *A comparison between Theorem 1 and Lemma 1 shows that the convergence rate of Algorithm 1 that approximates* θ_t^* *with* $\hat{\theta}_{t-1}$ *in Step 1 is the same as the ERM solution with high probability, as long as the change in the minimizers* ρ *is small enough, i.e.,* $L_4(\rho + \frac{1}{\delta}\sqrt{2\varepsilon/m}) < 1$. *In certain cases such as linear regression model, the Hessian matrices are independent of* θ_t^*. *Thus, no approximation is needed in constructing the sampling distribution, and Algorithm 1 is rate optimal.*

4.3 Sample Size Selection Rule

In this subsection, we explain and analyze the sample selection rule of Step 2 in Algorithm 1. The idea starts with the bound $b(\tau_t^2, \Delta_t, K_t)$ from Assumption 4. If we can compute τ_t^2 and Δ_t, the sample size K_t can be determined by letting $b(\tau_t^2, \Delta_t, K_t) \leq \varepsilon$ to satisfy the mean tracking 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. Although we can approximate θ_t^* using $\hat{\theta}_{t-1}$ as we did in Step 1, this upper bound only holds with high probability as shown in Theorem 1, which means the mean tracking criterion will be satisfied with high probability. To avoid this issue, we use the fact that $\mathrm{Tr}\big(I_{\Gamma_t^*}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big) \leq \mathrm{Tr}\big(I_{U_t}^{-1}(\theta_t^*) I_{U_t}(\theta_t^*)\big) = d$ (recall d is the dimension of parameters) to form 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 $K_t^* = \min\Big\{K \geq 1 \Big| b\Big(d/2, \sqrt{\frac{2\varepsilon}{m}} + \rho, K\Big) \leq \varepsilon\Big\}$ for $t \geq 2$ to ensure that $\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)] \leq \varepsilon$. For $t = 1$, we could always use $\mathrm{diameter}(\Theta)$ to bound Δ_1 and select K_1. In general, if ρ is much smaller than $\mathrm{diameter}(\Theta)$, then we require significantly fewer samples K_t to meet the mean tracking criterion for $t \geq 2$.

For the case where the change of the minimizers ρ is unknown, we could replace ρ with an estimate $\hat{\rho}_{t-1}$ to select the sample size. The following theorem characterizes the convergence guarantee using the sample size selection rule of step 2 in Algorithm 1 and the estimator of $\hat{\rho}_t$ in Section 4.4.

Theorem 2. *If*

$$K_t \geq K_t^* \triangleq \min\Big\{K \geq 1 \Big| b\Big(d/2, \sqrt{\frac{2\varepsilon}{m}} + \hat{\rho}_{t-1}, K\Big) \leq \varepsilon\Big\}, \tag{14}$$

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

4.4 Estimating the Change in Minimizers

In this subsection, we construct an estimate $\hat{\rho}_t$ of the change in the minimizers ρ using the active learning samples for step 5 in Algorithm 1.

We first construct an estimate $\widetilde{\rho}_t$ for the one-step changes $\|\theta_{t-1}^* - \theta_t^*\|$. As a consequence of strong convexity, the following lemma holds.

Lemma 2. *Suppose Assumption 2 holds, then*

$$\|\theta_{t-1}^* - \theta_t^*\|^2 \leq \frac{1}{m}\big[L_{U_t}(\theta_{t-1}^*) - L_{U_t}(\theta_t^*) + L_{U_{t-1}}(\theta_t^*) - L_{U_{t-1}}(\theta_{t-1}^*)\big]. \tag{15}$$

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

$$\widetilde{\rho}_t^2 = \frac{1}{m}\big[\hat{L}_{U_t}(\hat{\theta}_{t-1}) - \hat{L}_{U_t}(\hat{\theta}_t) + \hat{L}_{U_{t-1}}(\hat{\theta}_t) - \hat{L}_{U_{t-1}}(\hat{\theta}_{t-1})\big], \tag{16}$$

where we use

$$\hat{L}_{U_t}(\hat{\theta}_{t-1}) \triangleq \frac{1}{K_t}\sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t}, \hat{\theta}_{t-1})}{N_t \bar{\Gamma}_t(X_{k,t})} \tag{17}$$

as the empirical estimation of $L_{U_t}(\theta^*_{t-1})$. 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 [5], we need to normalize the estimate with the sampling distribution $\bar{\Gamma}_t$.

Then, we combine the one-step estimates to construct an overall estimate. The simplest way to combine the one-step estimates would be to set $\acute{\rho}_t^2 = \max\{\widetilde{\rho}_2^2, \cdots, \widetilde{\rho}_t^2\}$. However, if we suppose that each estimate $\widetilde{\rho}$ is an independent Gaussian random variable, then this estimate goes to infinity as $t \to \infty$. To avoid this issue, we use a class of 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.,

$$\acute{\rho}_t^2 = \frac{1}{t-1}\sum_{j=2}^{t} h_{\{\min[W,j-1]\}}(\widetilde{\rho}_j^2, \widetilde{\rho}_{j-1}^2, \cdots, \widetilde{\rho}^2_{\max[j-W+1,2]}). \tag{18}$$

The following theorem characterizes the performance of proposed estimator in (18).

Theorem 3. *Suppose Assumptions 1 and 2 hold, and there exists a sequence* $\{r_t\}$ [4] *satisfying*

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

for all t *large enough, then* $\hat{\rho}_t^2 \triangleq \acute{\rho}_t^2 + D_t + r_t \geq \rho^2$ *almost surely with a constant* D_t.

5 Experiments

In this section, we present two experiments to validate our algorithm and the related theoretical results: one is to track a synthetic regression model and the other is to track the time-varying user preferences in a recommendation system. More experiments on binary classification are presented in the Appendices. We use three baseline algorithms for comparison: passive adaptive algorithm, active random algorithm and passive random algorithm. Compared with Algorithm 1, *Passive* means drawing new samples using a uniform distribution U_t in Step 3 and *Random* means replacing the estimate of $\hat{\theta}_{t-1}$ with a random point from Θ in Step 1 and 4. All reported results are averaged over 1000 runs of Monte Carlo trials. The sizes of the sample pools for all the test algorithms are the same with $N_t = 500$, and the number of considered time steps is 25. We construct the active sampling distribution with the exact solution of the SDP problem in Step 1. Note that approximation algorithms for SDP introduced in [16] can be applied to accelerate this process. We set $K_t = K_t^*$ for all the test algorithms and use the estimator defined in Section 4.4 with window size $W = 3$ to estimate ρ.

5.1 Synthetic Regression

The model of the synthetic regression problem is $y_t = \theta_t^T x_t + w_t$, where the input variable $x_t \sim \mathcal{N}(0, 0.1I)$ is a 5-dimensional Gaussian vector and the noise $w_t \sim \mathcal{N}(0, 0.5)$. We consider learning the parameter θ_t by minimizing the following negative log-likelihood function $\ell(y_{k,t}|x_{k,t}, \theta_t) = (y_{k,t} - \theta_t^T x_{k,t})^2$. In the simulations, the change of the true minimizers is $\rho = 10$, and the target excess risk is $\varepsilon = 1$. To highlight the time-varying nature of the problem, we implement the "all samples up front" method by using $\sum_{t=1}^{25} K_t^*$ samples at the first time step and keep this time-invariant regression model for the rest of considered time steps.

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

Fig. 1(a) shows that using K_t^* new samples, the passive adaptive algorithm meets the mean tracking criterion and our proposed active and adaptive learning algorithm outperforms all the other algorithms. The "all samples up front" algorithm outperforms the other algorithms initially, but it fails to track the time-varying underlying model after only a few time steps. Moreover, the excess risk of active random algorithm is almost the same as that of active adaptive algorithm, since the Hessian matrices in the regression task are independent of θ_t. In this case, no approximation is needed and the change rate ρ in the regression task can be arbitrarily large, as we mentioned in Remark 1. Fig 1(b) shows that $\hat{\rho}_t$ converges to a conservative estimate of ρ, which verifies Theorem 3. Moreover, the corresponding number of samples determined by Theorem 2 is depicted in Fig. 1(c), which shrinks adaptively as $\hat{\rho}_t$ converges.

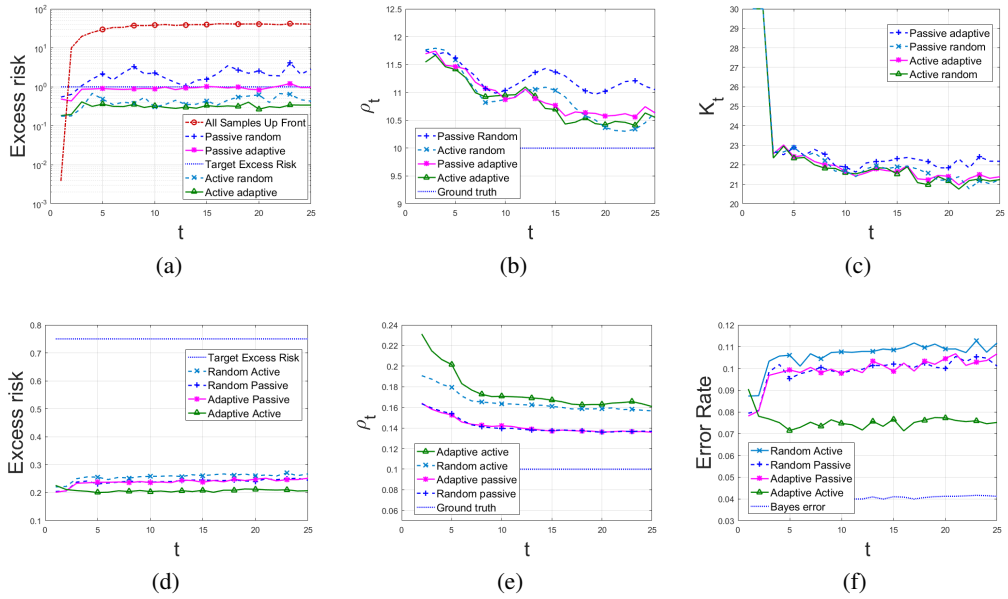


Figure 1: Experiments on synthetic regression: (a) Excess risk. (b) Estimated rate of change of minimizers. (c) Number of samples. Experiments on user preference tracking performance using Yelp data: (d) Excess risk. (e) Estimated rate of change of minimizers. (f) Classification error.

5.2 Tracking User Preferences in Recommendation System

We utilize a subset of Yelp 2017 dataset [5] to perform our experiments. We censor the original dataset such that each user has at least 10 ratings. After censoring procedure, our dataset contains ratings of $M = 473$ 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 very 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

$$p(R_{u,b}|\phi_b, \phi_u) = \frac{1}{1 + \exp^{-R_{u,b}\phi_u^\top \phi_b}}, \tag{19}$$

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 $d = 5$ 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 [9, 15].

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

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

active and adaptive learning algorithm to further reduce the number of new samples while maintaining the mean tracking accuracy.

In the following experiment, we use a random subset of $\{\phi_b\}$ with size N_t 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 Normal 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 (19) with true parameter $\phi_{u,t}$ instead. Note that one cannot ask a user the same question twice in a real recommendation system, and therefore we implement without replacement sampling by querying the labels of the samples having the largest K_t^* values in the active sampling distribution Γ_t.

Fig. 1(e) shows that $\hat{\rho}_t$ converges to a conservative estimate of ρ, and the corresponding sample size converges to $K_t^* = 13$ after two time steps. Fig. 1(d) and Fig. 1(f) show that our algorithm achieves a error rate of 6% with these samples and significantly outperforms the other algorithms. This is because the Hessian matrices of logistic regression are functions of θ_t, and hence the sampling distribution generated by the active and adaptive algorithm selects more informative samples.

6 Conclusions

In this paper, we propose an active and adaptive learning framework to solve a sequence of learning problems, which ensures a bounded excess risk for each individual learning task when the number of time steps is sufficiently large. We construct an estimator of the change in the minimizers $\hat{\rho}_t$ using active learning samples and show that this estimate upper bounds the true parameter ρ almost surely. We test our algorithm on a synthetic regression problem, and further apply it to a recommendation system that tracks changes in preferences of customers. Our experiments demonstrate that our algorithm achieves better performance compared to the other baseline algorithms.

References

[1] A. Agarwal, A. Rakhlin, and P. Bartlett. Matrix regularization techniques for online multitask learning. *Technical Report UCB/EECS-2008-138, EECS Department, University of California, Berkeley*, 2008.

[2] A. Agarwal, S. Gerber, and H. Daume. Learning multiple tasks using manifold regularization. In *Advances in Neural Information Processing Systems 23*, pages 46–54, 2010.

[3] R G Antonini and Yu V Kozachenko. A note on the asymptotic behavior of sequences of generalized subgaussian random vectors. *Random Operators and Stochastic Equations*, 13(1):39–52, 2005.

[4] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 65–72. ACM, 2006.

[5] O. Cappé, R. Douc, A. Guillin, J. Marin, and C. P Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

[6] K. Chaudhuri, S. M Kakade, P. Netrapalli, and S. Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems 28*, pages 1090–1098, 2015.

[7] J. A Cornell. *Experiments with mixtures: designs, models, and the analysis of mixture data*. John Wiley & Sons, 2011.

[8] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, pages 235–242, 2006.

[9] M. Elahi, F. Ricci, and N. Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50, 2016.

[10] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.

[11] R. Frostig, R. Ge, S. M Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, pages 728–763, 2015.

[12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[13] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[14] S. J Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

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

[16] J. Sourati, M. Akcakaya, T. K Leen, D. Erdogmus, and J. G Dy. Asymptotic analysis of objectives based on fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017.

[17] Zaid J Towfic, Jianshu Chen, and Ali H Sayed. On distributed online classification in the midst of concept drifts. *Neurocomputing*, 112:138–152, 2013.

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

[19] C. Wilson and V. V Veeravalli. Adaptive sequential optimization with applications to machine learning. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2642–2646, 2016.

[20] C. Wilson, V. V Veeravalli, and A. Nedich. Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control*, 2018.

[21] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. *arXiv preprint arXiv:1203.3536*, 2012.

A Proof of Lemma 1

To prove Lemma 1, we use the following result from [11]. In particular, the following lemma is a generalization of Theorem 5.1 in [11], and its proof follows from generalizing the derivation of that theorem and is omitted here.

Lemma 3. *Suppose $\psi_1(\theta), \cdots, \psi_K(\theta) : \mathbb{R}^d \to \mathbb{R}$ are random functions drawn i.i.d. from a distribution, where $\theta \in \Theta \subseteq \mathbb{R}^d$. Denote $P(\theta) = \mathbb{E}[\psi(\theta)]$ and let $Q(\theta) : \mathbb{R}^d \to \mathbb{R}$ be another function. Let*

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \sum_{k=1}^{K} \psi_k(\theta), \quad \text{and } \theta^* = \operatorname*{argmin}_{\theta \in \Theta} P(\theta).$$

Assume:

1. ***Regularity conditions**:*
   (a) *Compactness: Θ is compact, and θ^* is an interior point of Θ.*
   (b) *Smoothness: $\psi(\theta)$ is smooth in the following sense: the first, second and third derivatives exist at all interior points of Θ with probability one.*
   (c) *Convexity: $\psi(\theta)$ is convex with probability one, and $\nabla^2 P(\theta^*)$ is positive definite.*
   (d) *$\nabla P(\theta^*) = 0$ and $\nabla Q(\theta^*) = 0$.*
2. ***Concentration at** θ^*: Suppose*
   $$\Big\|\nabla\psi(\theta^*)\Big\|_{\nabla^2 P(\theta^*)^{-1}} \le L_1' \quad \text{and} \quad \Big\|\big(\nabla^2 P(\theta^*)\big)^{-1/2}\nabla^2\psi(\theta^*)\big(\nabla^2 P(\theta^*)\big)^{-1/2}\Big\|_2 \le L_2'$$
   hold with probability one.
3. ***Lipschitz continuity**: There exists a neighborhood B of θ^* and a constant L_3', such that $\nabla^2\psi(\theta)$ and $\nabla^2 Q(\theta)$ are L_3'-Lipschitz in this neighborhood, namely,*
   $$\Big\|\big(\nabla^2 P(\theta^*)\big)^{-1/2}\big(\nabla^2\psi(\theta) - \nabla^2\psi(\theta')\big)\big(\nabla^2 P(\theta^*)\big)^{-1/2}\Big\|_2 \le L_3'\|\theta - \theta'\|_{\nabla^2 P(\theta^*)},$$
   $$\Big\|\big(\nabla^2 Q(\theta^*)\big)^{-1/2}\big(\nabla^2 Q(\theta) - \nabla^2 Q(\theta')\big)\big(\nabla^2 Q(\theta^*)\big)^{-1/2}\Big\|_2 \le L_3'\|\theta - \theta'\|_{\nabla^2 P(\theta^*)},$$
   holds with probability one, for $\theta, \theta' \in B$,

Choose $p \ge 2$ and define

$$\gamma \triangleq c(L_1' L_3' + \sqrt{L_2'})\sqrt{\frac{p\log dK}{K}},$$

where c is an appropriately chosen constant. Let c' be another appropriately chosen constant. If K is large enough so that $\sqrt{\frac{p\log dK}{K}} \le c' \min\left\{\frac{1}{\sqrt{L_2'}}, \frac{1}{L_1' L_3'}, \frac{\text{diameter}(B)}{L_1'}\right\}$, then:

$$(1-\gamma)\frac{\tau^2}{K} - \frac{L_1'^2}{K^{p/2}} \le \mathbb{E}\big[Q(\hat{\theta}) - Q(\theta^*)\big] \le (1+\gamma)\frac{\tau^2}{K} + \frac{\max_{\theta\in\Theta}\left[Q(\theta) - Q(\theta^*)\right]}{K^p},$$

where

$$\tau^2 \triangleq \frac{1}{2K^2}\operatorname{Tr}\Bigg(\sum_{i,j}\mathbb{E}\big[\nabla\psi_i(\theta^*)\nabla\psi_j(\theta^*)^\top\big]\big(\nabla^2 P(\theta^*)\big)^{-1}\nabla^2 Q(\theta^*)\big(\nabla^2 P(\theta^*)\big)^{-1}\Bigg).$$

Then, we proceed to prove Lemma 1.

Proof of Lemma 1. We first use Lemma 3 to bound the excess risk, which is similar to the idea of Lemma 1 in [6]. We first define

$$\psi_k(\theta_t) = \ell(Y_{k,t}|X_{k,t}, \theta_t), \tag{20}$$

where $X_{k,t} \sim \Gamma_t$ and $Y_{k,t} \sim p(Y_{k,t}|X_{k,t}, \theta_t^*)$ for $1 \le k \le K_t$. Then,

$$P(\theta_t) = \mathbb{E}(\psi_k(\theta_t)) = L_{\Gamma_t}(\theta_t), \quad \text{and} \quad \nabla^2 P(\theta_t^*) = I_{\Gamma_t}(\theta_t^*). \tag{21}$$

Further, we choose

$$Q(\theta_t) = L_{U_t}(\theta_t), \quad \text{and} \quad \nabla^2 Q(\theta_t^*) = I_{U_t}(\theta_t^*). \tag{22}$$

As shown in Assumption 2, the assumptions of Lemma 3 are satisfied. Moreover, according to the condition that $I_{\Gamma_t}(\theta^*) \succeq C I_{U_t}(\theta^*)$ holds for some constant $C < 1$ in Lemma 1, we have

$$\begin{aligned}
&\Big\| I_{\Gamma_t}(\theta_t^*)^{-1/2}\big(H(x,\theta_t) - H(x,\theta_t')\big) I_{\Gamma_t}(\theta_t^*)^{-1/2} \Big\|_2 \\
&\qquad \le \frac{1}{C}\Big\| I_{U_t}(\theta_t^*)^{-1/2}\big(H(x,\theta_t) - H(x,\theta_t')\big) I_{U_t}(\theta_t^*)^{-1/2} \Big\|_2 \\
&\qquad \le \frac{L_3}{C}\|\theta - \theta'\|_{I_{U_t}(\theta_t^*)} \le \frac{L_3}{C^{3/2}}\|\theta - \theta'\|_{I_{\Gamma_t}(\theta_t^*)}
\end{aligned} \tag{23}$$

and

$$\begin{aligned}
&\Big\| I_{U_t}(\theta_t^*)^{-1/2}\big(H(x,\theta_t) - H(x,\theta_t')\big) I_{U_t}(\theta_t^*)^{-1/2} \Big\|_2 \\
&\qquad \le L_3\|\theta - \theta'\|_{I_{U_t}(\theta_t^*)} \le \frac{L_3}{\sqrt{C}}\|\theta - \theta'\|_{I_{\Gamma_t}(\theta_t^*)}.
\end{aligned} \tag{24}$$

Hence, $L_3' = \max\{L_3/C^{3/2}, L_3/\sqrt{C}\} = L_3/C^{3/2}$. Similarly, we have $L_1' = L_1/\sqrt{C}$ and $L_2' = L_2/C$. In summary, the Assumptions 2 and 3 in Lemma 3 are satisfied with constants

$$(L_1', L_2', L_3') = (L_1/\sqrt{C}, L_2/C, L_3/C^{3/2}). \tag{25}$$

Applying Lemma 3 with $p = 2$ and considering the fact that $\mathbb{E}_{x\sim\Gamma_t}\big[\nabla\ell(Y_{i,t}|X_{i,t},\theta_t^*)\nabla\ell(Y_{i,t}|X_{i,t},\theta_t^*)^\top\big] = I_{\Gamma_t}(\theta_t^*)$,

$$(1-\gamma_t)\frac{\tau_t^2}{K_t} - \frac{L_1^2}{CK_t^2} \le \mathbb{E}\big[L_{U_t}(\hat{\theta}_{\Gamma_t}) - L_{U_t}(\theta_t^*)\big] \le (1+\gamma_t)\frac{\tau_t^2}{K_t} + \frac{\max_{\theta\in\Theta_t}\left[L_{U_t}(\theta) - L_{U_t}(\theta_t^*)\right]}{K_t^2} \tag{26}$$

holds, where

$$\gamma_t = \mathcal{O}\bigg((L_1'L_3' + \sqrt{L_2'})\sqrt{\frac{\log dK_t}{K_t}}\bigg) = \mathcal{O}\bigg(\frac{1}{C^2}(L_1L_3 + \sqrt{L_2})\sqrt{\frac{\log dK_t}{K_t}}\bigg), \tag{27}$$

and $\tau_t^2 = \frac{1}{2}\mathrm{Tr}\Big(\big(I_{\Gamma_t}(\theta_t^*)\big)^{-1} I_{U_t}(\theta_t^*)\Big)$.

Note that if we assume the parameter set $\Theta_t \triangleq \{\theta_t | \|\theta_t - \theta_{t-1}^*\| \le \rho\}$ is known, then the second term in the right hand side of (26) can be further bounded as

$$\frac{\max_{\theta\in\Theta_t}\left[L_{U_t}(\theta) - L_{U_t}(\theta_t^*)\right]}{K_t^2} \le \frac{\max_{\theta\in\Theta_t}\left[L_b\|\theta - \theta_t^*\|^2\right]}{2K_t^2} \le \frac{L_b\mathrm{Diameter}(\Theta_t)^2}{2K_t^2} \le \frac{2L_b\rho^2}{K_t^2}, \tag{28}$$

where the inequalities follow from the boundedness condition in Assumption 2. Combining this result with the inequality in (26) completes the proof of Lemma 1. □

B Proof of Theorem 1

Proof of Theorem 1. The proof starts from the bound $b(\tau^2, \Delta_t, K_t)$ of the SGD algorithm in Assumption 4. To compute the convergence rate τ^2, we need to first study the approximation of θ_t^* using $\hat{\theta}_{t-1}$. The difference between $\hat{\theta}_{t-1}$ and θ_t^* can be bounded as

$$\big\|\hat{\theta}_{t-1} - \theta_t^*\big\|_2 \le \big\|\theta_{t-1}^* - \theta_t^*\big\|_2 + \big\|\hat{\theta}_{t-1} - \theta_{t-1}^*\big\|_2 \le \rho + \big\|\hat{\theta}_{t-1} - \theta_{t-1}^*\big\|_2. \tag{29}$$

To bound the second term, we use the strongly convexity assumption in Assumption 2,

$$\big\|\hat{\theta}_{t-1} - \theta_{t-1}^*\big\|_2^2 \le \frac{2}{m}(L_{U_{t-1}}\big(\hat{\theta}_{t-1}) - L_{U_{t-1}}(\theta_{t-1}^*)\big). \tag{30}$$

Suppose the excess risk bound $\mathbb{E}[L_{U_{t-1}}(\hat{\theta}_{t-1}) - L_{U_{t-1}}(\theta_{t-1}^*)] \le \varepsilon$ holds for $t-1$. Then, we have

$$\mathbb{E}(\big\|\hat{\theta}_{t-1} - \theta_{t-1}^*\big\|_2) \le \sqrt{\mathbb{E}(\big\|\hat{\theta}_{t-1} - \theta_{t-1}^*\big\|_2^2)} \le \sqrt{2\varepsilon/m}. \tag{31}$$

Then, $\big\|\hat{\theta}_{t-1} - \theta^*_{t-1}\big\|_2 \leq \frac{1}{\delta}\sqrt{\frac{2\varepsilon}{m}}$ holds with probability $1-\delta$ by Markov's inequality, for $\forall \delta \in (0,1)$. Thus,

$$\big\|\hat{\theta}_{t-1} - \theta^*_t\big\|_2 \leq \rho + \frac{1}{\delta}\sqrt{\frac{2\varepsilon}{m}} \tag{32}$$

holds with probability $1-\delta$. By the self-concordance condition in Assumption 3, we have that

$$(1-\beta_t)H(x_t, \theta^*_t) \preceq H(x_t, \hat{\theta}_{t-1}) \preceq (1+\beta_t)H(x_t, \theta^*_t), \quad x_t \in \mathcal{S}_t, \tag{33}$$

holds with probability $1-\delta$, where $\beta_t = L_4(\rho + \frac{1}{\delta}\sqrt{\frac{2\varepsilon}{m}})$. Then, for distribution Γ^*_t, $\hat{\Gamma}^*_t$ and U_t, we have

$$(1-\beta_t)I_{\Gamma^*_t}(\theta^*_t) \preceq I_{\Gamma^*_t}(\hat{\theta}_{t-1}) \preceq (1+\beta_t)I_{\Gamma^*_t}(\theta^*_t), \tag{34}$$

$$(1-\beta_t)I_{\hat{\Gamma}^*_t}(\theta^*_t) \preceq I_{\hat{\Gamma}^*_t}(\hat{\theta}_{t-1}) \preceq (1+\beta_t)I_{\hat{\Gamma}^*_t}(\theta^*_t), \tag{35}$$

$$(1-\beta_t)I_{U_t}(\theta^*_t) \preceq I_{U_t}(\hat{\theta}_{t-1}) \preceq (1+\beta_t)I_{U_t}(\theta^*_t). \tag{36}$$

Recall that $\bar{\Gamma}_t = \alpha_t\hat{\Gamma}^*_t + (1-\alpha_t)U_t$. Hence, $I_{\bar{\Gamma}_t}(\theta^*_t) \succeq \alpha_t I_{\hat{\Gamma}^*_t}(\theta^*_t)$ which implies that $I_{\bar{\Gamma}_t}(\theta^*_t)^{-1} \preceq \frac{1}{\alpha_t}I_{\hat{\Gamma}_t}(\theta^*_t)^{-1}$. Thus,

$$\tau_t^2 = \frac{1}{2}\mathrm{Tr}\big(I^{-1}_{\bar{\Gamma}_t}(\theta^*_t)I_{U_t}(\theta^*_t)\big) \leq \frac{1}{2\alpha_t}\mathrm{Tr}\big(I^{-1}_{\hat{\Gamma}^*_t}(\theta^*_t)I_{U_t}(\theta^*_t)\big). \tag{37}$$

From (35) and (36), (37) can be further upper bounded by

$$\begin{aligned}
\mathrm{Tr}\big(I^{-1}_{\hat{\Gamma}^*_t}(\theta^*_t)I_{U_t}(\theta^*_t)\big) &\leq \frac{1+\beta_t}{1-\beta_t}\mathrm{Tr}\big(I^{-1}_{\hat{\Gamma}^*_t}(\hat{\theta}_{t-1})I_{U_t}(\hat{\theta}_{t-1})\big) \\
&\overset{(a)}{\leq} \frac{1+\beta_t}{1-\beta_t}\mathrm{Tr}\big(I^{-1}_{\Gamma^*_t}(\hat{\theta}_{t-1})I_{U_t}(\hat{\theta}_{t-1})\big) \\
&\overset{(b)}{\leq} \left(\frac{1+\beta_t}{1-\beta_t}\right)^2 \mathrm{Tr}\big(I^{-1}_{\Gamma^*_t}(\theta^*_t)I_{U_t}(\theta^*_t)\big),
\end{aligned} \tag{38}$$

where (a) is because that $\hat{\Gamma}^*_t$ is the minimizer of $\mathrm{Tr}\big(I^{-1}_{\Gamma_t}(\hat{\theta}_{t-1})I_{U_t}(\hat{\theta}_{t-1})\big)$ and (b) follows from the results in (34) and (36).

To bound the difference between the initialization and the true minimizer, we use triangle inequality and Jensen's inequality to get

$$\sqrt{\mathbb{E}\|\hat{\theta}_{t-1} - \theta^*_t\|_2^2} \leq \sqrt{\mathbb{E}\|\hat{\theta}_{t-1} - \theta^*_{t-1}\|_2^2} + \|\theta^*_t - \theta^*_{t-1}\| \leq \sqrt{\mathbb{E}\|\hat{\theta}_{t-1} - \theta^*_{t-1}\|_2^2} + \rho. \tag{39}$$

From (31), we have

$$\mathbb{E}\|\hat{\theta}_{t-1} - \theta^*_{t-1}\|_2^2 \leq \frac{2\varepsilon}{m}, \tag{40}$$

which yields

$$\mathbb{E}\|\hat{\theta}_{t-1} - \theta^*_t\|_2^2 \leq \left(\sqrt{\frac{2\varepsilon}{m}} + \rho\right)^2 = \Delta_t^2. \tag{41}$$

Thus, combining the above result with the bound in (38), we can conclude that the following upper bound

$$\mathbb{E}[L_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta^*_t)] \leq b(\acute{\tau}_t^2, \Delta_t, K_t), \tag{42}$$

holds with probability 1-δ, where

$$\acute{\tau}_t^2 = \Big(\frac{1+\beta_t}{1-\beta_t}\Big)^2 \frac{\mathrm{Tr}\big(I^{-1}_{\Gamma^*_t}(\theta^*_t)I_{U_t}(\theta^*_t)\big)}{2\alpha_t}. \tag{43}$$

This completes the proof of Theorem 1.

□

C Proof of Lemma 2

Proof of Lemma 2. The following inequalities hold from the strong convexity assumption and the fact that $\nabla L_{U_t}(\theta_t^*) = \nabla L_{U_{t-1}}(\theta_{t-1}^*) = 0$:

$$L_{U_t}(\theta_{t-1}^*) \geq L_{U_t}(\theta_t^*) + \frac{1}{2}m\|\theta_t^* - \theta_{t-1}^*\|_2^2 \tag{44}$$

$$L_{U_{t-1}}(\theta_t^*) \geq L_{U_{t-1}}(\theta_{t-1}^*) + \frac{1}{2}m\|\theta_t^* - \theta_{t-1}^*\|_2^2. \tag{45}$$

Then, adding and rearranging these inequalities yields

$$\frac{1}{m}\Big[L_{U_t}(\theta_{t-1}^*) - L_{U_t}(\theta_t^*) + L_{U_{t-1}}(\theta_t^*) - L_{U_{t-1}}(\theta_{t-1}^*)\Big] \geq \|\theta_t^* - \theta_{t-1}^*\|_2^2. \tag{46}$$

□

Moreover, we have the following relation

$$\begin{aligned}
&\|\theta_t^* - \theta_{t-1}^*\|_2^2 \\
&\quad\leq \frac{1}{m}\Big[L_{U_t}(\theta_{t-1}^*) - L_{U_t}(\theta_t^*) + L_{U_{t-1}}(\theta_t^*) - L_{U_{t-1}}(\theta_{t-1}^*)\Big] \\
&\quad= \frac{1}{m}\Big[\mathbb{E}_{X\sim U_t}\big[D\big(p(Y|X,\theta_t^*)\|p(Y|X,\theta_{t-1}^*)\big)\big] + \mathbb{E}_{X\sim U_{t-1}}\big[D\big(p(Y|X,\theta_{t-1}^*)\|p(Y|X,\theta_t^*)\big)\big]\Big],
\end{aligned} \tag{47}$$

where

$$D(p\|q) \triangleq \int_{y\in\mathcal{Y}} p(y)\log\frac{p(y)}{q(y)}dy \tag{48}$$

is the KL divergence between distribution p and q.

Thus, an upper bound of ρ can be constructed by estimating the symmetric KL divergence between $p(y|x,\theta_t^*)$ and $p(y|x,\theta_{t-1}^*)$ using the data pool U_t and U_{t-1}, respectively.

D Proof of Theorem 3

To analyze the performance of the estimator of ρ, we need to introduce a few results for sub-Gaussian random variables including the following key technical lemma from [3]. This lemma controls the concentration of sums of random variables that are sub-Gaussian conditioned on a particular filtration.

Lemma 4. *Suppose we have a collection of random variables $\{V_i\}_{i=1}^n$ and a filtration $\{\mathscr{F}_i\}_{i=0}^n$ such that for each random variable V_i it holds that*

1. $\mathbb{E}[\exp\{s(V_i - \mathbb{E}[V_i|\mathscr{F}_{i-1}])\}|\mathscr{F}_{i-1}] \leq e^{\frac{1}{2}\sigma_i^2 s^2}$ *with σ_i^2 a constant.*
2. *V_i is $\mathscr{F}_i$-measurable.*

Then for every $\boldsymbol{a} \in \mathbb{R}^n$ it holds that

$$\mathbb{P}\left\{\sum_{i=1}^n a_i V_i > \sum_{i=1}^n a_i\mathbb{E}[V_i|\mathscr{F}_{i-1}] + t\right\} \leq \exp\left\{\frac{t^2}{2\nu}\right\}$$

with $\nu = \sum_{i=1}^n \sigma_i^2 a_i^2$. The other tail is similarly bounded.

If we can upper bound the conditional expectations $\mathbb{E}[V_i|\mathscr{F}_{i-1}] \leq \xi_i$ by some constants ξ_i, then we have

$$\mathbb{P}\left\{\sum_{i=1}^n a_i V_i > \sum_{i=1}^n a_i\xi_i + t\right\} \leq \exp\left\{\frac{t^2}{2\nu}\right\}. \tag{49}$$

For our analysis, we generally cannot compute $\mathbb{E}[V_i|\mathscr{F}_{i-1}]$ directly, but we can find the upper bound ξ_i. To compute σ_i^2 for use in Lemma 4, we employ the following conditional version of Hoeffding's Lemma.

Lemma 5. *(Conditional Hoeffding's Lemma): If a random variable V and a sigma algebra $\mathscr{F}$ satisfy $a \leq V \leq b$ and $E[V|\mathscr{F}] = 0$, then*

$$\mathbb{E}[e^{sV}|\mathscr{F}] \leq \exp\Big\{\frac{1}{8}(b-a)^2 s^2\Big\}.$$

Proof of Lemma Theorem 3. To simplify our proof, we look at a special case where $\|\theta_t^* - \theta_{t-1}^*\| = \rho$ holds. The proof for the case $\|\theta_t^* - \theta_{t-1}^*\| \leq \rho$ is similar, and more details about the window function h_W can be found in [20].

For the case $\|\theta_t^* - \theta_{t-1}^*\| = \rho$, we use the following estimator to combine the one-step estimator $\widetilde{\rho}_t$

$$\acute{\rho}_t^2 = \frac{1}{t-1}\sum_{i=2}^{t}\widetilde{\rho}_i^2 = \frac{1}{m(t-1)}\sum_{i=2}^{t}\big(\hat{L}_{U_i}(\hat{\theta}_{i-1}) - \hat{L}_{U_i}(\hat{\theta}_i) + \hat{L}_{U_{i-1}}(\hat{\theta}_i) - \hat{L}_{U_{i-1}}(\hat{\theta}_{i-1})\big). \quad (50)$$

We denote

$$\rho_t^2 \triangleq \frac{1}{m(t-1)}\sum_{i=2}^{t}\big(L_{U_i}(\theta_{i-1}^*) - L_{U_i}(\theta_i^*) + L_{U_{i-1}}(\theta_i^*) - L_{U_{i-1}}(\theta_{i-1}^*)\big) \geq \rho^2. \quad (51)$$

where the inequality follows from Lemma 2. We want to construct $\hat{\rho}_t$, such that $\hat{\rho}_t^2 \geq \rho_t^2 \geq \rho^2$ almost surely. Then, we have

$$\rho_t^2 - \acute{\rho}_t^2 = \frac{1}{m(t-1)}\Big(\sum_{i=2}^{t} L_{U_i}(\theta_{i-1}^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1}) + \sum_{i=2}^{t} L_{U_{i-1}}(\theta_i^*) - \hat{L}_{U_{i-1}}(\hat{\theta}_i) \quad (52)$$

$$+ \hat{L}_{U_1}(\hat{\theta}_1) - L_{U_1}(\theta_1^*) + 2\sum_{i=2}^{t-1}\hat{L}_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta_i^*) + \hat{L}_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)\Big). \quad (53)$$

Define

$$U_t \triangleq \frac{1}{(t-1)}\sum_{i=2}^{t}\frac{1}{m}\big(L_{U_i}(\theta_{i-1}^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1})\big), \quad (54)$$

$$V_t \triangleq \frac{1}{(t-1)}\sum_{i=2}^{t}\frac{1}{m}\big(L_{U_{i-1}}(\theta_i^*) - \hat{L}_{U_{i-1}}(\hat{\theta}_i)\big), \quad (55)$$

$$W_t \triangleq \frac{1}{m(t-1)}\Big(\hat{L}_{U_1}(\hat{\theta}_1) - L_{U_1}(\theta_1^*) + 2\sum_{i=2}^{t-1}\big(\hat{L}_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta_i^*)\big) + \hat{L}_{U_t}(\hat{\theta}_t) - L_{U_t}(\theta_t^*)\Big). \quad (56)$$

Then it holds that

$$\rho_t^2 - \acute{\rho}_t^2 = U_t + V_t + W_t. \quad (57)$$

Now, we look at bounding $\mathbb{E}\big[L_{U_i}(\theta_{i-1}^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1})\big]$, $\mathbb{E}\big[L_{U_{i-1}}(\theta_i^*) - \hat{L}_{U_{i-1}}(\hat{\theta}_i)\big]$ and $\mathbb{E}\big[\hat{L}_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta_i^*)\big]$ in U_t, V_t and W_t, respectively.

Note that, the samples at time step $i-1$ are independent with samples at time i, hence,

$$\begin{aligned}\mathbb{E}\big[\hat{L}_{U_i}(\hat{\theta}_{i-1})\big] &= \mathbb{E}\Big[\mathbb{E}_{X_{k,i}\sim\bar{\Gamma}_i, Y_{k,i}\sim p(Y|X_{k,i},\theta_i^*)}\big[\frac{1}{K_i}\sum_{i=1}^{K_i}\frac{\ell(Y_{k,i}|X_{k,i},\hat{\theta}_{i-1})}{N_i\bar{\Gamma}_i(X_{k,i})}\big]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}_{X_i\sim U_i, Y_i\sim p(Y|X_i,\theta_i^*)}\big[\ell(Y_t|X_t,\hat{\theta}_{i-1})\big]\Big] \\ &= \mathbb{E}\big[L_{U_i}(\hat{\theta}_{i-1})\big]. \end{aligned} \quad (58)$$

Thus,

$$\mathbb{E}\big[L_{U_i}(\theta_{i-1}^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1})\big] = \mathbb{E}\big[L_{U_i}(\theta_{i-1}^*) - L_{U_i}(\hat{\theta}_{i-1})\big], \quad (59)$$

$$\mathbb{E}\big[L_{U_{i-1}}(\theta_i^*) - \hat{L}_{U_{i-1}}(\hat{\theta}_i)\big] = \mathbb{E}\big[L_{U_{i-1}}(\theta_i^*) - L_{U_{i-1}}(\hat{\theta}_i)\big]. \quad (60)$$

We use Lemma 3 to construct bounds for these two terms. Let

$$Q(\theta) = \big(L_{U_i}(\theta^*_{i-1}) - L_{U_i}(\theta)\big)^2, \text{ and } \psi_k(\theta) = \ell(Y_k|X_k,\theta), \quad 1 \le k \le K_{i-1}, \tag{61}$$

where $X_k \sim \bar{\Gamma}_{i-1}$ and $Y_k \sim p(Y|X_k, \theta^*_{i-1})$. It can be verified that

$$\hat{\theta}_{i-1} = \operatorname*{argmin}_{\theta\in\Theta} \sum_{k}^{K_{i-1}} \psi_k(\theta), \quad \theta^* = \operatorname*{argmin}_{\theta\in\Theta} P(\theta) = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}[\psi(\theta)] = \theta^*_{i-1}, \tag{62}$$

and $\nabla Q(\theta^*_{i-1}) = 0$. All the conditions in Lemma 3 are satisfied. We have

$$\nabla^2 P(\theta^*) = I_{\bar{\Gamma}_{i-1}}(\theta^*_{i-1}), \quad \nabla^2 Q(\theta^*) = 2I_{U_i}(\theta^*_{i-1}). \tag{63}$$

Thus,

$$\begin{aligned}
&\big(\mathbb{E}\big[L_{U_i}(\theta^*_{i-1}) - L_{U_i}(\hat{\theta}_{i-1})\big]\big)^2 \\
&\quad\le \mathbb{E}\big[\big(L_{U_i}(\theta^*_{i-1}) - L_{U_i}(\hat{\theta}_{i-1})\big)^2\big] \\
&\quad\le (1+\gamma_{i-1})\frac{\mathrm{Tr}\big(I_{\bar{\Gamma}_{i-1}}(\theta^*_{i-1})^{-1} I_{U_i}(\theta^*_{i-1})\big)}{K_{i-1}} + \frac{\max_{\theta\in\Theta}\big[L_{U_i}(\theta) - L_{U_i}(\theta^*_{i-1})\big]^2}{K_{i-1}^2} \\
&\quad\triangleq A_i.
\end{aligned} \tag{64}$$

Similarly, we have

$$\begin{aligned}
&\big(\mathbb{E}[L_{U_{i-1}}(\theta^*_i) - L_{U_{i-1}}(\hat{\theta}_i)]\big)^2 \\
&\quad\le (1+\gamma_i)\frac{\mathrm{Tr}\big(I_{\bar{\Gamma}_i}(\theta^*_i)^{-1} I_{U_{i-1}}(\theta^*_i)\big)}{K_i} + \frac{\max_{\theta\in\Theta}\big[L_{U_{i-1}}(\theta) - L_{U_{i-1}}(\theta^*_i)\big]^2}{K_i^2} \\
&\quad\triangleq B_i.
\end{aligned} \tag{65}$$

For the term $\mathbb{E}\big[L_{U_i}(\theta^*_i) - \hat{L}_{U_i}(\hat{\theta}_i)\big]$ in W_t, suppose that the samples used to estimate $\hat{\theta}_i$ and the samples used to compute $\hat{L}_{U_i}$ are independent. This can be done by splitting the samples at each time step i. Note that this assumption is just required to proceed with the theoretical analysis; we will use all the samples to estimate $\hat{\theta}_i$ in practice.

Then, similar argument holds as in (58), and we have

$$\mathbb{E}\big[\hat{L}_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta^*_i)\big] = \mathbb{E}\big[L_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta^*_i)\big] \ge 0. \tag{66}$$

where the inequality follows from the fact that θ^*_t is the minimizer of $L_{U_t}(\theta)$. Applying the upper bound in Lemma 1, this term can be bounded as

$$\begin{aligned}
0 \le \mathbb{E}\big[L_{U_i}(\theta^*_i) - L_{U_i}(\hat{\theta}_i)\big] &\le (1+\gamma_i)\frac{\mathrm{Tr}\big(I^{-1}_{\bar{\Gamma}_i}(\theta^*_i) I_{U_i}(\theta^*_i)\big)}{2K_i} + \frac{\max_{\theta\in\Theta}\big[L_{U_i}(\theta) - L_{U_i}(\theta^*_i)\big]}{K_i^2} \\
&\triangleq C_i.
\end{aligned} \tag{67}$$

The resulting bounds on the expectation of U_t, V_t, and W_t denoted $\bar{U}_t$, $\bar{V}_t$, and $\bar{W}_t$ are as follows:

$$\bar{U}_t = \frac{1}{m(t-1)} \sum_{i=2}^{t} \sqrt{A_i}, \tag{68}$$

$$\bar{V}_t = \frac{1}{m(t-1)} \sum_{i=2}^{t} \sqrt{B_i}, \tag{69}$$

$$\bar{W}_t = \frac{1}{m(t-1)} (C_1 + 2\sum_{i=2}^{t-1} C_i + C_t). \tag{70}$$

Now, we find the upper bound ξ_i to upper bound the expectation as we mentioned in (49). Then it holds that

$$\begin{aligned}
&\mathbb{P}\Big\{\rho_t^2 - \acute{\rho}_t^2 > \bar{U}_t + \bar{V}_t + \bar{W}_t + r_t\Big\} \\
&\quad= \mathbb{P}\Big\{U_t + V_t + W_t > \bar{U}_t + \bar{V}_t + \bar{W}_t + r_t\Big\} \\
&\quad\le \mathbb{P}\Big\{U_t > \bar{U}_t + \frac{1}{3}r_t\Big\} + \mathbb{P}\Big\{V_t > \bar{V}_t + \frac{1}{3}r_t\Big\} + \mathbb{P}\Big\{W_t > \bar{W}_t + \frac{1}{3}r_t\Big\}.
\end{aligned} \tag{71}$$

To bound these probabilities with (49), we first bound the moment generating functions using Lemma 5,

$$\frac{1}{m}|\hat{L}_{U_i}(\hat{\theta}_i) - L_{U_i}(\theta_i^*)| \leq \frac{L_b}{2m}\max_{\theta\in\Theta}\|\theta-\theta_i^*\|^2 \leq \frac{L_b}{2m}\mathrm{Diameter}(\Theta)^2, \tag{72}$$

and

$$\begin{aligned}\frac{1}{m}|L_{U_i}(\theta_{i-1}^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1})| &\leq \frac{1}{m}|L_{U_i}(\theta_{i-1}^*) - L_{U_i}(\theta_i^*)| + \frac{1}{m}|L_{U_i}(\theta_i^*) - \hat{L}_{U_i}(\hat{\theta}_{i-1})| \\ &\leq \frac{L_b}{m}\mathrm{Diameter}(\Theta)^2. \end{aligned}\tag{73}$$

Then, we apply Lemma 4 and Lemma 5 with $\sigma_i^2 = \frac{L_b^2}{4m^2}\mathrm{Diameter}^4(\Theta)$ for the terms in U_t and V_t, and apply $\sigma_i^2 = \frac{L_b^2}{16m^2}\mathrm{Diameter}^4(\Theta)$ for the terms in W_t, respectively. We have

$$\nu_U = \nu_V = \frac{L_b^2}{4m^2}\mathrm{Diameter}(\Theta)^4\sum_{i=2}^{t}\frac{1}{(t-1)^2} = \frac{L_b^2}{4(t-1)m^2}\mathrm{Diameter}(\Theta)^4, \tag{74}$$

$$\nu_W \leq \frac{L_b^2}{16m^2}\mathrm{Diameter}(\Theta)^4\sum_{i=2}^{t}\big(\frac{2}{t-1}\big)^2 = \frac{L_b^2}{4(t-1)m^2}\mathrm{Diameter}(\Theta)^4. \tag{75}$$

Let $D_t \triangleq \bar{U}_t + \bar{V}_t + \bar{W}_t$. Then we obtain

$$\mathbb{P}\Big\{\rho_t^2 > \acute{\rho}_t^2 + D_t + r_t\Big\} \leq 3\exp\Big\{-\frac{2m^2(t-1)r_t^2}{9L_b^2\mathrm{Diameter}^4(\Theta)}\Big\}. \tag{76}$$

Then it follows the assumption in Theorem 3 that

$$\sum_{t=2}^{\infty}\mathbb{P}\Big\{\acute{\rho}_t^2 + D_t + r_t < \rho_t^2\Big\} \leq \sum_{t=2}^{\infty}3\exp\Big\{-\frac{2m^2(t-1)r_t^2}{9L_b^2\mathrm{Diameter}^4(\Theta)}\Big\} < \infty. \tag{77}$$

Therefore, by the Borel-Cantelli Lemma, for all t large enough it holds that

$$\hat{\rho}_t^2 = \acute{\rho}_t^2 + D_t + r_t \geq \rho_t^2 \tag{78}$$

almost surely. Finally, it holds that $\rho_t^2 \geq \rho^2$ from Lemma 2, which proves the result. □

E Proof of Theorem 2

To prove Theorem 2, we use the following result from Theorem 3 in [20].

Lemma 6. *If $\hat{\rho}_t \geq \rho$ almost surely for t sufficiently large, then with*

$$K_t \geq K_t^* \triangleq \min\Big\{K \geq 1\Big|b\Big(d/2, \big(\sqrt{\frac{2\varepsilon}{m}} + \hat{\rho}_{t-1}\big)^2, K\Big) \leq \varepsilon\Big\} \tag{79}$$

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

Proof of Theorem 2. From Theorem 3, we know that the proposed estimate $\hat{\rho}_t^2 \geq \rho^2$ almost surely, which implies $\hat{\rho}_t \geq \rho$ almost surely. Directly applying the above lemma completes the proof. □

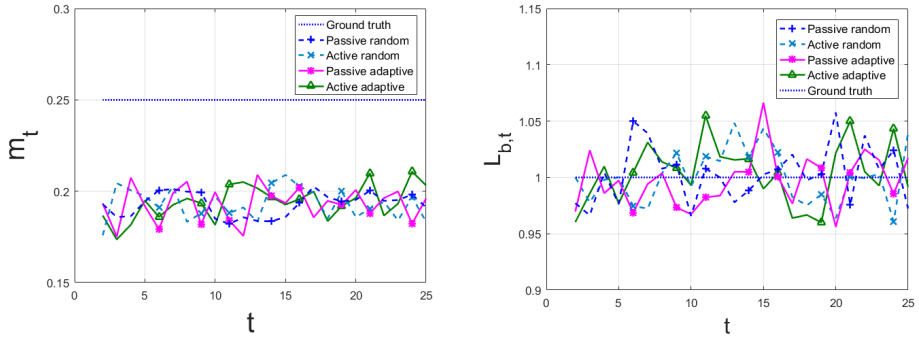
F Estimation of m and L_b

We construct the estimator of m and L_b with the samples drawn from distribution $\bar{\Gamma}_t$. By the assumption of strongly convexity, we have

$$L_{U_t}(\theta) \geq L_{U_t}(\theta') + \langle\nabla L_{U_t}(\theta'), \theta-\theta'\rangle + \frac{m}{2}\|\theta-\theta'\|^2,\ \forall\theta,\theta' \in \Theta, \tag{80}$$

which implies that

$$m \leq \frac{L_{U_t}(\theta) - L_{U_t}(\theta') - \langle\nabla L_{U_t}(\theta'), \theta-\theta'\rangle}{\frac{1}{2}\|\theta-\theta'\|^2} \tag{81}$$



(a) (b)

Figure 2: Estimated parameter on the regression task over synthetic data. (a) Estimated strongly convex parameter. (b) Estimated largest eigenvalue.

holds for any $\theta, \theta' \in \Theta$.

Since m is the smallest value satisfying (81) for any $\theta, \theta' \in \Theta$, we consider following estimator

$$\widetilde{m}_t \triangleq \min_{\theta,\theta' \in \Theta_t} \frac{2}{K_t} \sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t},\theta) - \ell(Y_{k,t}|X_{k,t},\theta') - \langle \nabla \ell(Y_{k,t}|X_{k,t},\theta'), \theta - \theta' \rangle}{N_t \bar{\Gamma}_t(X_{k,t}) \|\theta - \theta'\|^2}. \tag{82}$$

Following (82), we have

$$\begin{aligned}
\mathbb{E}(\widetilde{m}_t) &= \mathbb{E}_{X_{k,t} \sim \Gamma_t} \left\{ \min_{\theta,\theta' \in \Theta_t} \frac{2}{K_t} \sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t},\theta) - \ell(Y_{k,t}|X_{k,t},\theta') - \langle \nabla \ell(Y_{k,t}|X_{k,t},\theta'), \theta - \theta' \rangle}{N_t \Gamma_t(X_{k,t}) \|\theta - \theta'\|^2} \right\} \\
&\leq \min_{\theta,\theta' \in \Theta_t} \mathbb{E}_{X_{k,t} \sim \Gamma_t} \left\{ \frac{2}{K_t} \sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t},\theta) - \ell(Y_{k,t}|X_{k,t},\theta') - \langle \nabla \ell(Y_{k,t}|X_{k,t},\theta'), \theta - \theta' \rangle}{N_t \Gamma_t(X_{k,t}) \|\theta - \theta'\|^2} \right\} \\
&= \min_{\theta,\theta' \in \Theta_t} \mathbb{E}_{X_{k,t} \sim U_t} \left\{ \frac{2}{K_t} \sum_{k=1}^{K_t} \frac{\ell(Y_{k,t}|X_{k,t},\theta) - \ell(Y_{k,t}|X_{k,t},\theta') - \langle \nabla \ell(Y_{k,t}|X_{k,t},\theta'), \theta - \theta' \rangle}{\|\theta - \theta'\|^2} \right\} \\
&= \min_{\theta,\theta' \in \Theta_t} \frac{L_{U_t}(\theta) - L_{U_t}(\theta') - \langle \nabla L_{U_t}(\theta'), \theta - \theta' \rangle}{\frac{1}{2}\|\theta - \theta'\|^2} \\
&= m,
\end{aligned} \tag{83}$$

which implies that $\widetilde{m}_t$ is a conservative estimate of m. In practice, the strongly convex parameter m may also vary with time t. Thus, we use the following estimator to combine the one-step estimator $\widetilde{m}_t$,

$$\hat{m}_t = \min\{\widetilde{m}_{t-1}, \widetilde{m}_t\}, \tag{84}$$

for $t \geq 2$.

Moreover, following the boundedness assumption in Assumption 2, we have

$$\max_{\theta \in \Theta} \lambda_{\max} \left[I_{U_t}(\theta) \right] \leq L_b, \tag{85}$$

where $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a square matrix. In this case, we consider following estimator

$$\hat{L}_{b,t} \triangleq \max_{\theta_t \in \Theta_t} \lambda_{\max} \left[\frac{1}{K_t} \sum_{k=1}^{K_t} \frac{1}{N_t} \frac{1}{\Gamma_t(X_{k,t})} H(X_{k,t}, \theta_t) \right]. \tag{86}$$

Similarly, $\hat{L}_b$ is also a conservative estimate of L_{b}. That is,

$$\begin{aligned}
\mathbb{E}(\hat{L}_{b,t}) &= \mathbb{E}_{X_{k,t}\sim\Gamma_t}\left\{\max_{\theta_t\in\Theta_t}\ \lambda_{\max}\ \left(\frac{1}{K_t}\sum_{k=1}^{K_t}\frac{1}{N_t}\frac{1}{\Gamma_t(X_{k,t})}H(X_{k,t},\theta_t)\right)\right\} \\
&\geq \max_{\theta_t\in\Theta_t}\ \mathbb{E}_{X_{k,t}\sim\Gamma_t}\left\{\lambda_{\max}\ \left[\frac{1}{K_t}\sum_{k=1}^{K_t}\frac{1}{N_t}\frac{1}{\Gamma_t(X_{k,t})}H(X_{k,t},\theta_t)\right]\right\} \\
&\geq \max_{\theta_t\in\Theta_t}\ \lambda_{\max}\ [I_{U_t}(\theta_t)] \\
&= L_b.
\end{aligned} \tag{87}$$

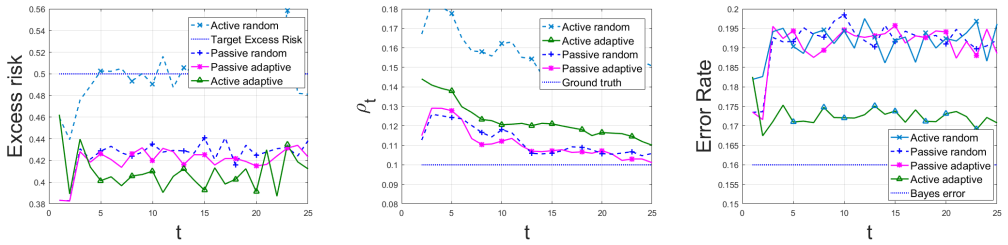
Fig. 2(a) and Fig. 2(b) demonstrate our estimation of $\hat{m}_t$ and $\hat{L}_{b,t}$ in the synthetic regression problem described in Section 5, respectively.

G Experiments on Synthetic Classification

We consider solving a sequence of binary classification problems by using logistic regression. At time t, the features of two classes are drawn from Gaussian distribution with different means $\mu_{1,t}$ and $-\mu_{1,t}$. More specifically, the features are 2-dimensional Gaussian vectors with $\|\mu_{1,t}\|_2 = 2$ and variance $0.25I$. The parameter θ_t is learned by minimizing the following log-likelihood function

$$\ell(y_{k,t}|x_{k,t},\theta_t) = \log(1+\exp^{-y_{k,t}\theta_t^\top x_{k,t}}). \tag{88}$$

To ensure the change of minimizers is bounded, we set that $\mu_{1,t}$ is drifting with a constant rate along a 2-dimensional sphere. We further set $\rho = 0.1$ and $\epsilon = 0.5$.



(a) (b) (c)

Figure 3: Experiments on synthetic classification: (a) Excess risk. (b) Estimated rate of change of minimizers. (c) Classification error.

Fig. 3 shows that active adaptive learning outperforms other baseline algorithms in the synthetic classification problem.