No-Regret Safety: Balancing Tests and Misclassification in Logistic Bandits

Tavor Z. Baharav[*,†]
Eric and Wendy Schmidt Center
Broad Institute
Cambridge, MA 02142
baharav@broadinstitute.org

Spyros Dragazis[*]
Department of Computer Science
Boston University
Boston, MA 02215
dragazis@bu.edu

Aldo Pacchiano[†]
Boston University & Broad Institute
Boston, MA 02215
pacchian@bu.edu

Abstract

We study the problem of sequentially testing individuals for a binary disease outcome whose true risk is governed by an unknown logistic model. At each round, a patient arrives with feature vector x_t, and the decision maker may either pay to administer a (noiseless) diagnostic test—revealing the true label—or skip testing and predict the patient's disease status based on their feature vector and prior history. Our goal is to minimize the total number of costly tests while guaranteeing that the fraction of misclassifications does not exceed a prespecified error tolerance α, with probability at least $1-\delta$. To address this, we develop a novel algorithm that (i) maintains a confidence ellipsoid for the unknown logistic parameter $\theta^\star$, (ii) interleaves label-collection and distribution-estimation to estimate both $\theta^\star$ and the context distribution, and (iii) computes a conservative, data-driven threshold τ_t on the logistic score $|x_t^\top \theta|$ over θ in the confidence set to decide when testing is necessary. We prove that, with probability at least $1-\delta$, our procedure does not exceed the target misclassification rate, and incurs only $\widetilde{O}(\sqrt{T})$ excess tests compared to the oracle baseline that knows both $\theta^\star$ and the patient feature distribution. This establishes the first no-regret guarantees for error-constrained logistic testing, with direct applications to cost-sensitive medical screening. Simulations corroborate our theoretical guarantees, showing that in practice our procedure efficiently estimates $\theta^\star$ while retaining safety guarantees, and does not require too many excess tests.

1 Introduction

Modern machine learning has recently provided solutions to real-world automated decision-making systems in various fields such as drug discovery [52, 11], recommendation systems [3, 56], online ad-allocation [49], and portfolio selection [43]. Bandit algorithms [36] and reinforcement learning [50] play a significant role in building interactive decision-making systems that collect feedback from users and improve their performance with each interaction. Two primary challenges exist in the aforementioned applications: the first is the learning challenge, estimating the problem parameters which are vital for decision-making; the second is the decision-making challenge, where effective performance is required concurrently with learning.

[*]Equal contribution.
[†]Equal senior contribution.

Preprint. Under review.

Although machine learning systems perform exceptionally well in practice, sometimes even surpassing human performance, when applied in human-centric scenarios safety constraints are paramount [27, 24]. Many mathematical formulations have been proposed to characterize what safety means in sequential decision making settings. The first one is based on satisfying cost constraints and is characterized by the requirement of playing actions that belong to a safe set as specified by a cost signal [40, 55, 21]. The second one, also known as conservative bandits, requires the learner to play actions that achieve a reward level comparable or superior to a fixed baseline [34]. In sequential decision making problems learning while satisfying a safety criterion typically makes reward acquisition more challenging. Thus the main challenge in these scenarios remains to understand how to optimally manage these tradeoffs.

Inspired by the COVID-19 pandemic, and more broadly medical triage application, we study an online learning problem with a different type of safety constraint. In our setting, patients sequentially arrive with an associated feature vector (fever, loss of smell, fatigue, blood oxygen saturation), and a latent unobserved disease state (whether they have COVID or not). The hospital has limited COVID tests due to resource constraints, and wants to minimize their usage. However, they want to ensure that they properly quarantine sick patients. Here, we posit a latent (unknown) logistic model between the patient's feature vector and their disease status; as more patients are observed, the hospital can learn that a low blood oxygen saturation and a high fever correspond to a high likelihood of COVID, and so the patient does not need to be tested but can immediately be classified as sick. Thus, the hospital must, as the data is being collected, learn a) the distribution of patients, b) the parameters of the logistic model, and c) the decision threshold of when to test.

Related problems have been studied in the active learning and selective sampling literature [47, 29, 39, 7, 20, 48]. These study settings where context information may be abundant but the labels are hard to come by [16]. More formally, in the active learning or online selective sampling literature at the start of every round the learner observes a context vector $X_t \in \mathbb{R}^d$ and has the option to query or not the label $Z_t \in \{0, 1\}$. The goal is to build a statistical learning algorithm that achieves similar performance (i.e., generalization error) to one that observes all the labels while minimizing the expected number of queries used. A connection between selective sampling and active learning can be found in [13].

By focusing on the classification task and changing the objective from minimizing the generalization error to minimizing the cumulative pseudo regret (with respect to the optimal labeling policy), various algorithms have been developed in the online selective sampling literature, such as [39, 46], by considering both stochastic and adversarial contexts. The objective in these works is to achieve sublinear regret while minimizing the expected number of queries made. A similar line of work is the one of online selective classification [22, 23, 25] where the learner has the right to abstain from classifying. The objective is to minimize the expected number of abstentions with the least amount of expected mistakes.

However, in real-world scenarios like the one in [6], it makes sense to ask that the training error remain under a safety threshold with high probability while minimizing the number of queries. For example in the streaming patient scenario we described above, where patients arrive one by one and the medical provider needs to classify them as sick or not. In this problem, due to the sensitive nature of making misclassification mistakes, the objective is to devise a selective testing procedure that can guarantee the total misclassification error remains below a safety threshold $\alpha \in [0, 1]$. Testing every patient clearly attains this safety threshold, but can be prohibitively expensive. Our question is thus:

Can we design an adaptive algorithm that minimizes the expected number of tests while maintaining a misclassification rate below a specified safety threshold?

In this work, we formalize this notion of (α, δ)-safety, where an algorithm attains a misclassification error rate of α with probability at least $1 - \delta$. We define a baseline testing policy, that is optimal when the α classification rate is only required to hold in expectation, which tests $p^\star \triangleq p^\star(\alpha)$ fraction of the time. We develop an adaptive algorithm to solve this problem, with (α, δ)-safety guarantees, which requires only a sublinear number of excess tests: $\mathcal{O}(\sqrt{\frac{dT}{p^\star(\alpha)} \log(T/\delta)})$.* We corroborate our theoretical results through comprehensive synthetic experiments.

*In Theorem 2 we show that the regret is upper bounded by $\tilde{\mathcal{O}}(\sqrt{dT/(p^\star \lambda_0)})$ where λ_0 is the minimum eigenvalue of the covariance matrix of the contexts observed under the optimal policy. When contexts are

2 Preliminaries

Notation We adopt the following notation throughout the paper. The inner product between two vectors $x, y \in \mathbb{R}^n$ will be denoted either as $x^\top y$ or as $\langle x, y \rangle$. We denote the ℓ_2 norm of a vector $x \in \mathbb{R}^d$ as $\|x\|_2 = \sqrt{\langle x, x \rangle}$ and $\|x\|_A = \sqrt{x^\top A x}$ for any positive semi-definite matrix $A \in \mathbb{R}^{d \times d}$ The minimum eigenvalue of a matrix $A \in \mathbb{R}^{d \times d}$ will be denoted as $\lambda_{\min}(A)$. The set $\{1, 2, \ldots, n\}$ is denoted as $[n]$. The logistic function is denoted as $\mu(z) = \frac{1}{1+\exp(z)}$ and $\mathbb{1}(E)$ denotes the indicator function of an event E. For two functions f, g we say that $f(x) \preccurlyeq g(x)$ when there exists an absolute constant $c > 0$ such that $f(x) \leq c g(x)$ for all x. We use upper case letters for random variables and lower case for scalars. For any measurable set A we denote the set of all distributions on A as $\Delta(A)$.

2.1 Problem Definition

We consider the following repeated interaction between a learner and the environment. At every round $t \in [T]$, the environment generates a context $X_t \in \mathbb{R}^d$ on the unit sphere. These contexts are identically distributed, and are drawn independently from an unknown distribution with density P. Every patient-context has an unseen random label $Y_t \in \{0, 1\}$ that represents their disease status. We assume that $Y_t \sim \text{Ber}(\mu(X_t^\top \theta^\star))$, independent from all other $X_{t'}$ and $Y_{t'}$. Here, $\theta^\star \in \Theta \subseteq \mathbb{R}^d$ is some fixed parameter vector unknown to the learner, such that $\|\theta^\star\|_2 \leq 1$.

At each round, the learner observes the patient's context X_t and must decide whether or not to test the patient, denoted by $Z_t \in \{0, 1\}$. Then, the learner must predict whether the patient is healthy or sick, denoted by $\hat{Y}_t \in \{0, 1\}$. If $Z_t = 1$, the patient is tested, and the learner observes the true label Y_t, and so can predict $\hat{Y}_t = Y_t$. The random variable Z_t can depend on information obtained prior to that decision, i.e. $\mathcal{H}_t = \{X_1, Z_1 Y_1, X_2, Z_2 Y_2, \ldots, X_t\}$ and possibly on internal randomization of the learner. Similarly, $\hat{Y}_t$ must be $\mathcal{F}_t = \sigma\{X_1, Z_1 Y_1, X_2, Z_2 Y_2, \cdots, X_t, Z_t Y_t\}$ measurable. The goal of the learner is to minimize the expected number of tests applied, while guaranteeing that the misclassification rate is less than a desired threshold α, with probability at least $1 - \delta$. We define this constraint as (α, δ)-safety:

Definition 1. *An algorithm outputting $\{\hat{Y}_t\}$ satisfies (α, δ)-safety if*

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\} \leq \alpha\right) \geq 1 - \delta$$

where the probability is computed with respect to the randomness in $\{X_t\}, \{Y_t\}$, and any randomness internal to the algorithm in constructing $\{\hat{Y}_t\}$.

Our safe learning objective can then be framed as below:

$$\min_{\{\hat{Y}_t\}, \{Z_t\}} \sum_{t=1}^{T} \mathbb{E} Z_t \quad \text{s.t.} \quad \{\hat{Y}_t\} \text{ satisfy } (\alpha, \delta) \text{ safety.} \tag{1}$$

2.2 Baseline policy

First, we characterize the optimal testing strategy satisfying the conditions of equation 1 in the case where the feature distribution P and optimal discriminator $\theta^\star$ are known a priori to the learner. Although many decision rules Z_t are possible, we will focus on threshold rules of the form:

$$Z_t = \mathbb{1}\{|X_t^\top \theta^\star| \leq \tau\} \qquad \hat{Y}_t = \begin{cases} 0 & \text{if } X_t^\top \theta^\star < -\tau, \\ Y_t & \text{if } |X_t^\top \theta^\star| \leq \tau, \\ 1 & \text{if } X_t^\top \theta^\star > \tau. \end{cases}$$

When P and $\theta^\star$ are known, a threshold decision rule is optimal when the constraint satisfaction is imposed only in expectation, as we show in the following proposition:

uniformly distributed on the unit sphere, $\lambda_0 = 1/\sqrt{d}$, recovering the linear dimension dependence of linear bandits.

Proposition 1. *Consider a variation of safe learning (Equation* (1)*) where the constraint holds in expectation, and both the sequence of contexts* $\{X_t\}_{t=1}^T$ *and the parameter vector* $\theta^\star$ *are known. The optimal policy for this problem is a threshold rule:*

$$\min_{\{\hat{Y}_t\}} \sum_{t=1}^{T} \mathbb{E} Z_t \quad \textit{s.t.} \quad \mathbb{E}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\}\right) \leq \alpha.$$

The proof of this proposition follows by relating this to the fractional knapsack problem, which we detail in Section 10. This motivates our use of a threshold policy as a baseline. We consider competing against the optimal threshold decision rule $\tau^\star$ that is a function of P, $\theta^\star$, and α.

To identify $\tau^\star$, we begin by analyzing the performance of a fixed threshold τ. To do this we define the function $p_{\text{err}}(\theta, P, \tau)$ as the probability of misclassification incurred by the the threshold τ, if θ was the underlying logistic parameter, and where the expectation is taken with respect to P:

$$p_{\text{err}}(\theta, P, \tau) = \int (1 + \exp(|x^\top \theta|))^{-1} \mathbb{1}\left\{|x^\top \theta| > \tau\right\} P(dx). \tag{2}$$

The term inside the integral $(1+\exp(|x^\top\theta|))^{-1} = \min\left\{\frac{1}{1+\exp(x^\top\theta)}, 1 - \frac{1}{1+\exp(x^\top\theta)}\right\}$ is the optimal misclassification error for a fixed x, θ pair. The term $\mathbb{1}\left\{|x^\top\theta| > \tau\right\}$ equals one only if we predict the label $\hat{y}$ without observing the real label y for context x, when using a threshold rule. Having defined the error probability for a given threshold τ, we can now easily define the optimal threshold. For any problem parameters $\theta \in \mathbb{R}^d, \alpha' \in [0,1]$, and distribution $\rho \in \Delta(\mathcal{X})$, we define the optimal decision threshold $\tau^\star$ as the minimum value of $\tau \in [0,1]$ that satisfies the α-fraction misclassification constraint:

$$\tau^\star(\theta, \rho, \alpha') \triangleq \min\{\tau : p_{\text{err}}(\theta, \rho, \tau) \leq \alpha'\}. \tag{3}$$

Evaluated at the true parameters $\theta^\star$, P, and α, the optimal threshold $\tau^\star$ alluded to in Proposition 1 implies that any any algorithm requires an expected number of tests $p^\star T$, where

$$\tau^\star \triangleq \tau^\star(\theta^\star, P, \alpha), \tag{4}$$

$$p^\star \triangleq \mathbb{P}\left(x : |x^\top\theta^\star| \leq \tau^\star\right). \tag{5}$$

Here, we have overloaded notation for $\tau^\star$ as both a function, and the evaluation of this function at the true problem parameters. Note that in practice, p_{err} must be estimated using $\hat{P}$, our observed samples from P, in addition to $\theta^\star$ being unknown.

Before introducing our "regret" objective, we examine the relationship between the safety parameter α, which serves as an input, and the baseline policy sampling probability $p^\star$. When the misclassification rate threshold α approaches zero, the system must minimize error rates, necessitating testing of all cases. This constraint leads to increased values of $\tau^\star$ and, consequently, higher values of $p^\star$. Conversely, in the degenerate scenarios where α approaches unity, policies become indifferent to misclassification errors and conduct vanishing testing, yielding values of $p^\star$ that approach zero.

This lets us define the "safe regret" of an algorithm as the number of excess tests it takes over this oracle baseline, while satisfying (α, δ)-safety. An algorithm could trivially sample at each time step and satisfy the misclassification criterion; the question is, for a given misclassification rate α, and error probability δ, can a learner achieve sublinear safe regret in T, as defined in Definition 2?

Definition 2. *For any policy* $\pi : \mathcal{X} \to \{0,1\}^2$ *that produces the sequence of actions and predictions* $\{Z_t\}_{t=1}^\infty, \{\hat{Y}\}_{t=1}^\infty$*, we define the safe regret of* π *as follows:*

$$\mathbb{E}\left[\sum_{t=1}^{T} Z_t - p^\star\right] \quad \textit{s.t.} \quad \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\} \leq \alpha\right) \geq 1 - \delta.$$

To analyze this quantity, we make the following natural assumptions.

Assumption 1. *The optimal baseline tests a nonzero fraction of the time, i.e.* $p^\star > 0$.

Other works such as, [39], [46], use the notation T_ε to describe the number of times the Bayes optimal classifier outputs a label with confidence less than a fixed parameter $\varepsilon > 0$. Our $p^\star$ is analogous to T_ε.

It serves as a measure to quantify the inherent difficulty of the problem instance (how many patients are close to the decision boundary).

We make two additional assumptions on the density of the contexts P, which are reasonable for patient data with continuous valued features.

Assumption 2. *The density P is upper and lower bounded by constants $[m, M]$, where $0 < m \leq P(x) \leq M < \infty$, for all x such that $\|x\|_2 \leq 1$.*

Assumption 3. *There exists a constant $\lambda_0 > 0$:*

$$\lambda_{\min}\left(\mathbb{E}\left[XX^\top \mid \left|X^\top\theta^\star\right| \leq \tau^\star\right]\right) \geq \lambda_0.$$

Adaptive sampling works such as [46, 29], and those tackling learning halfspaces, commonly assume the Tsybakov noise condition [51, 14]. The Tsybakov noise condition with parameters (α, A) states that $\mathbb{P}_{x\sim P}[\eta(x) \geq 1/2 - t] \leq At^{\frac{\alpha}{1-\alpha}}$ for any $0 < t \leq 1/2$, where $\eta(x) = \mathbb{P}(Y(x) = 1)$. This implies that, around the value of $1/2$ where the Bayes Optimal classifier is uncertain, the density of the contexts decays rapidly at a rate controlled by the parameters (α, A). In our setting, this assumption is not necessary or helpful, as near the uncertainty boundary the learner will simply test the patient. Another assumption in the literature is that the contexts are uniformly distributed over the surface of the unit sphere (Theorem 2 in [12]). Our assumption is much less stringent, and encompasses standard distributions such as smooth densities of the form $f(x) = g(\|x\|)$, or truncated Gaussian distributions.

Note that these assumptions are strictly for the *analysis* of our algorithm. We do not require knowledge of any of these parameters m, M, λ_0, or $p^\star$ as input to our algorithm. We are able to learn and adapt to them on the fly, they simply need to be strictly positive and finite.

Theorem 1 (Informal statement of Theorem 2)**.** *Under Assumptions 1-3, Algorithm 1 satisfies (α, δ)-safety, and has safe regret of order $\mathcal{O}(\sqrt{\frac{d}{p^\star\lambda_0}T\log(T/\delta)})$.*

2.3 Logistic Bandits tools

Our algorithm leverages confidence intervals for $\theta^\star$ from existing methods. [17] provides two methods (Appendix B.3): the first produces a confidence ellipsoid, while the second provides a tighter but non-convex confidence set. The advantage of the non-convex one is the lack of dependence on the quantity $\kappa \triangleq \sup_{(X,\theta)\in(\mathcal{X},\Theta)} \frac{1}{\dot{\mu}(\langle X,\theta\rangle)}$ that characterizes the non-linearity of the logistic function over the decision set $(\mathcal{X}, \Theta)$ and scales exponentially with the size of the decision set. In our setting, we utilize the first method to simplify the algorithm and its analysis. Moreover, we can bound the value of $\kappa = \frac{1}{\mu(1)(1-\mu(1))} \leq 6$ as $|\langle x, \theta^\star\rangle| \leq 1$ by Cauchy-Schwarz and boundedness assumptions for $\|x\|, \|\theta^\star\|$. Recently, tighter confidence intervals for the logistic bandit setting were proven by [37], but the results of [17] are sufficient for our needs.

Before stating our algorithm, we borrow some notation from [17]. We denote the labeled samples we use for estimating $\theta^\star$ that have been collected up to beginning of round t as $\mathcal{S}_\theta^t$, with $N_\theta^t = |\mathcal{S}_\theta^t|$. Since in this work we only collect labeled samples (X_t, Y_t) if we test at a given round, N_θ^t may not equal $t - 1$. We define the regularized log-likelihood as

$$\mathcal{L}_t(\theta) = \sum_{s\in\mathcal{S}_\theta^t}\left[y_s \log\mu(x_s^T\theta) + (1 - y_s)\log(1 - \mu(x_s^T\theta))\right] - \frac{1}{2}\|\theta\|_2^2,$$

and its maximum (regularized) likelihood estimator as $\hat{\theta}_t = \operatorname{argmax}_{\theta\in\mathbb{R}^d} \mathcal{L}_t(\theta)$. We also denote the design matrix as $V_t = \sum_{s\in\mathcal{S}_\theta^t} X_s X_s^\top + \kappa\mathbf{I}_d$, and for technical reasons we consider a projection θ_t^L of $\hat{\theta}_t$ onto the feasible set Θ defined as follows,

$$\theta_t^L \triangleq \operatorname*{argmin}_{\theta\in\Theta}\left\|g_t(\theta) - g_t(\hat{\theta}_t)\right\|_{V_t^{-1}} \text{ where } g_t(\theta) = \sum_{s\in\mathcal{S}_\theta^t}\mu(\langle x_s, \theta\rangle)x_s + \theta. \tag{6}$$

These allow us to define the confidence ellipsoid $\mathcal{C}_t$ for $\theta^\star$, which is implicitly a function of δ:

$$\mathcal{C}_t \triangleq \left\{\theta \in \Theta, \left\|\theta - \theta_t^L\right\|_{V_t} \leq B_t\right\}, \tag{7}$$

Where the confidence radius B_t is defined as:

$$B_t \triangleq 2\kappa \left(1 + \sqrt{\log\left(\frac{1}{\delta}\right) + 2d\log\left(1 + \frac{N_\theta^t}{\kappa d}\right)}\right). \tag{8}$$

We will track how quickly this decays with $\lambda_{\min}^t \triangleq \lambda_{\min}(V_t)$, which is computable from the observed data.

These confidence intervals [17] satisfy the following anytime, high probability guarantees:

Lemma 1. *[Lemma 12 of [17].] For any fixed choice of δ, let G_θ be the "good" event that the confidence intervals defined in Equation* (7) *are valid:*

$$\mathbb{P}(G_\theta) = \mathbb{P}\big(\forall t \geq 1, \theta^\star \in \mathcal{C}_t \mid N_\theta^t\big) \geq 1 - \delta.$$

Since the number of samples N_θ^t collected to estimate $\mathcal{C}_t$ is a random variable in our setting, we condition on its value in Lemma 1.

Before diving into our algorithm and its analysis, we discuss the role and behavior of key quantities that will arise. To begin, the numbers of samples collected N_θ^t to build our θ confidence intervals grows linearly in t, concretely $N_\theta^t \succeq p^\star t$.As a consequence, the bound B_t used in $\mathcal{C}_t$ (which satisfies $B_t \leq B_T$) grows extremely slowly in t, with $B_t \succeq \sqrt{d\log(1 + \frac{p^\star t}{d})}$. The other portion of the confidence interval involves $\|x\|_{V_t^{-1}}$, which we must upper bound. We show that $\|x\|_{V_t^{-1}} \leq \frac{\|x\|_2}{\lambda_{\min}(V_t^{-1})} \preceq \frac{1}{\sqrt{t\lambda_0}}$, from Assumption 3. Many prior works in Online Logistic Regression [8] or in Linear Bandits [1] utilize the elliptic potential lemma, which is unnecessary due to the stochastic contexts in our setting.

3 Algorithm design

After defining these logistic bandit preliminaries, we are now able to define and analyze our algorithm, SCOUT (Safe Contextual Online Understanding with Thresholds) in Algorithm 1. At every time step, SCOUT observes the label of a specific context if the inner product between this context and the estimated $\theta^\star$ is too close to an estimator of the true threshold $\tau^\star$. To iteratively refine the estimates of $\theta^\star$ and $\tau^\star$, SCOUT employs a classical sample-splitting trick to avoid dependencies, utilizing data from odd samples for estimation of the context distribution P (which is used to estimate $\tau^\star$), and data from even samples where a test was performed for $\theta^\star$ estimation.

The testing condition Z_t can be computed as follows: we defer the derivation and details to Section 6. Note that if we were to target α fraction misclassification, then half of the time we would exceed this, so instead we target $\alpha_t = \max(0, \alpha - \sqrt{\log(7t^2/\delta)/2t})$ (discussed inSection 12.3).

$$Z_t \triangleq \mathbb{1}\{|\langle X_t, \theta_t^L\rangle| \leq \tau_t\} \tag{9}$$

where $\tau_t \triangleq \hat{\tau}(\theta_t^L, \hat{P}_t, \alpha_t) + B_t/\sqrt{\lambda_{\min}^t} = \tau^\star(\theta_t^L, \hat{P}_t, \alpha_t - \zeta_t - 2B_t/\sqrt{\lambda_{\min}^t}) + 3B_t/\sqrt{\lambda_{\min}^t}$. This can be framed as $Z_t = \mathbb{1}\{c_t \leq 0\}$, where $c_t \triangleq |\langle X_t, \theta_t^L\rangle| - \tau_t$. This can be compared to the optimal rule $Z_t^\star = \mathbb{1}\{|\langle X_t, \theta^\star\rangle| \leq \tau^\star\} = \mathbb{1}\{c_t^\star \leq 0\}$ where $c_t^\star \triangleq |\langle X_t, \theta^\star\rangle| - \tau^\star$. The two sampling rules match except for the use of the estimated quantities $\theta_t^L, \hat{P}_t$, as opposed to the true unknown quantities, and the use of additional exploration bonuses and uncertainty penalties. ζ_t arises from confidence intervals on our estimates, where we only have samples $\hat{P}_t$ and not the true P:

$$\zeta_t \triangleq \sqrt{\frac{d\log(3/\varepsilon_Q) + \log\left(\frac{\pi^2 t^2}{3\delta}\right)}{4t}}. \tag{10}$$

The other term, $B_t \|X_t\|_{V_t^{-1}}$, arises from the fact that we only have the estimate θ_t^L and not $\theta^\star$. The final term ε_Q is a quantization parameter used in our analysis (discussed in Section 4.1.1), and should be thought of as some small quantity of order $1/T^2$.

Algorithm 1 SCOUT

Input: Number of rounds T, target error rate α, confidence level δ
Initialize: $\mathcal{S}_P^{(1)} = \emptyset$, $\mathcal{S}_\theta^{(1)} = \emptyset$. Maintain $N_P^t = |\mathcal{S}_P^t|$, $N_\theta^t = |\mathcal{S}_\theta^t|$
for $t = 1, 2, \ldots, T$ **do**
  Observe context X_t
  if $t \leq 2$ **then**
    Set $Z_t = 1$
  else
    Compute θ_t^L from (6) and τ_t from (9)
    Set $Z_t = \mathbb{1}\{|\langle \theta_t^L, X_t \rangle| \leq \tau_t\}$
  end if
  if $Z_t = 1$ **then**
    Observe Y_t
    Predict $\hat{Y}_t = Y_t$
  else
    Predict $\hat{Y}_t = \mathbb{1}\{\langle X_t, \theta_t^L \rangle > 0\}$
  end if
  if $Z_t = 1$ and t is even **then**
    Set $\mathcal{S}_\theta^{t+1} = \mathcal{S}_\theta^t \cup \{(X_t, Y_t)\}$
  end if
  if t is odd **then**
    Set $\mathcal{S}_P^{t+1} = \mathcal{S}_P^t \cup \{X_t\}$
  end if
end for

4 Stability of error estimates

To analyze our algorithm, we first study the stability properties of p_{err}. Concretely, the learner does not a priori know P or $\theta^\star$, and by extension $\tau^\star$. Thus, we must show that, as we gradually learn these quantities, our estimates of the error probabilities are not too far off.

4.1 Stability with respect to context sampling $\hat{P}_t$

Analyzing Equation (2), we note that we do not know the true distribution P, but only have access to samples from it. For any fixed θ and τ, (2) is a sum of i.i.d. $[0, 1/2]$ bounded random variables, enabling us to use standard concentration bounds.

Lemma 2. *Let $\hat{P}_t$ be the empirical distribution of constructed from $\lfloor t/2 \rfloor$ i.i.d. samples from P. Then, for any fixed θ and τ, with probability at least $1 - \delta$ over the randomness in $\hat{P}_t$:*

$$\left| p_{err}(\theta, \hat{P}_t, \tau) - p_{err}(\theta, P, \tau) \right| \leq \sqrt{\frac{\log\left(\frac{\pi^2 t^2}{3\delta}\right)}{4t}}$$

This requires the application of Hoeffding's inequality [54]. We defer additional proof details to Section 11.1.1.

We would like this bound to hold over all $\theta \in \Theta$ and $\tau \in [0, 1]$. However, this would preclude using a union bound over our estimators, due to the uncountable nature of these sets of these sets. Thus, we utilize an ϵ-net for both $\tau \in [0, 1]$ and $\theta \in \Theta$.

4.1.1 Quantization

We define quantized versions of τ and θ, so that we can safely union bound the failure probability of our estimators over the countable quantized set. We take an $\varepsilon_Q = T^{-2}$ covering of the unit interval for τ as $\mathcal{Q}_\tau \triangleq \mathcal{N}([0, 1], \varepsilon_Q)$, denoting the quantized τ value as $\tau_Q \in \mathcal{Q}_\tau$. We additionally take an ε_Q covering of the d dimensional unit sphere for θ as $\mathcal{Q}_\theta \triangleq \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon_Q)$, denoting the quantized θ value as $\theta_Q \in \mathcal{Q}_\theta$. Then, $|\mathcal{Q}_\tau| = \varepsilon_Q^{-1}$ and $|\mathcal{Q}_\theta| \leq (3/\varepsilon)^d$ [53].

To this end, we define the quantized optimized τ as $\tau_Q^\star$, which is close to the true $\tau^\star$:

$$\tau_Q^\star(\theta, \hat{P}, \alpha) \triangleq \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, \hat{P}, \tau_Q) \leq \alpha\}. \tag{11}$$

As p_{err} is monotonically decreasing in the threshold τ we have that

$$\tau^\star(\theta, \hat{P}, \alpha) \leq \tau_Q^\star(\theta, \hat{P}, \alpha) \leq \tau^\star(\theta, \hat{P}, \alpha) + \varepsilon_Q. \tag{12}$$

4.2 Stability of $\tau^\star$ with respect to θ

We now show that our estimate $p_{\text{err}}(\theta, \hat{P}, \tau)$ is close to $p_{\text{err}}(\theta^\star, \hat{P}, \tau)$ when θ is close to $\theta^\star$, for any distribution ρ and threshold τ.

Lemma 3. *For all $\theta, \theta' \in \Theta$, $\tau \geq \frac{\|\theta-\theta'\|_{V_t}}{\sqrt{\lambda_{\min}^t}}$, and density $\rho(x)$ on $\mathcal{X}$:*

$$p_{err}(\theta, \rho, \tau) \leq \; p_{err}\left(\theta', \rho, \tau - \frac{\|\theta - \theta'\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) + \frac{\|\theta - \theta'\|_{V_t}}{\sqrt{\lambda_{\min}^t}}.$$

We defer the proof to Section 11.2.1. This indicates that as our estimation of θ improves, so will our error probability estimates. To this end, we define the good event $G_{p_{\text{err}}}$ where our error probability estimates are uniformly bounded by ζ_t on our quantized set as:

$$G_{p_{\text{err}}} = \left\{\left|p_{\text{err}}(\theta_Q, \hat{P}_t, \tau_Q) - p_{\text{err}}(\theta_Q, P, \tau_Q)\right| \leq \zeta_t \; : \; \forall t \in [T], \forall \theta_Q \in \mathcal{Q}_\theta, \forall \tau_Q \in \mathcal{Q}_\tau\right\}. \tag{13}$$

The following lemma shows that this good event $G_{p_{\text{err}}}$ happens with overwhelming probability.

Lemma 4. *The good event $G_{p_{err}}$ satisfies $\mathbb{P}(G_{p_{err}}) \geq 1 - \delta$.*

We defer the proof of this Lemma to Section 11.2.2, which utilizes Lemma 2 and the union bound over the quantized sets $\mathcal{Q}_\theta$ and $\mathcal{Q}_\tau$. Conditioning on the good event $G_{p_{\text{err}}}$, we show that $\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha)$ is close to $\tau^\star$ when θ_Q is close to $\theta^\star$.

Lemma 5. *Conditioning on $G_{p_{err}}$, for any $\theta_Q \in \mathcal{Q}_\theta \cap \mathcal{C}_t, \theta \in \mathcal{C}_t$ such that $\frac{\|\theta_Q-\theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} \leq \tau^\star(\theta, P, \alpha - \zeta_t - \frac{\|\theta_Q-\theta^\star\|_{V_t}}{\sqrt{\lambda_{\min}^t}})$ it is true that:*

$$\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) \leq \tau^\star\left(\theta, P, \alpha - \zeta_t - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} + \varepsilon_Q,$$

$$\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) \geq \tau^\star\left(\theta, P, \alpha + \zeta_t + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}$$

We defer the proof of this Lemma to Section 11.3. This enables us to construct an estimator $\hat{\tau}(\theta_Q, \hat{P}_t, \alpha)$ by selecting $\theta = \theta^\star$, by noting that on the good events G_θ and $G_{p_{\text{err}}}$ that:

$$\|\theta_Q - \theta^\star\|_{V_t} \leq 2B_t. \tag{14}$$

Thus, for any $\theta_Q \in \mathcal{Q}_\theta \cap \mathcal{C}_t$ we can construct an estimator $\hat{\tau}$ as:

$$\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) \triangleq \tau_Q^\star\left(\theta_Q, \hat{P}_t, \alpha - \zeta_t - 2B_t/\sqrt{\lambda_{\min}^t}\right) + 2B_t/\sqrt{\lambda_{\min}^t}, \tag{15}$$

where conditioning on $G_{p_{\text{err}}}, G_\theta$, it holds that $\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) \geq \tau^\star(\theta^\star, P, \alpha)$ for all $\theta_Q \in \mathcal{Q}_\theta \cap \mathcal{C}_t$.

4.3 Smoothness of $\tau^\star$ with respect to α

The last property we will need for our analysis is that $\tau^\star$ does not vary too quickly with respect to α. We show that for small γ, $\tau^\star(\theta^\star, P, \alpha - \gamma)$ is not too much larger than $\tau^\star$. Note that while p_{err} is continuous with respect to τ when evaluated at the true distribution P, it is discontinuous when evaluated at $\hat{P}$ due to the resulting indicator functions. However, utilizing Assumption 2, the true distribution of contexts is upper and lower bounded by constants, and so p_{err}, which integrates the distribution, will change at an upper and lower bounded rate.

Lemma 6. *Under Assumptions 1 and 2,*

$$\tau^\star(\theta^\star, P, \alpha - \gamma) \leq \tau^\star(\theta^\star, P, \alpha) + K_{\max}\gamma + \varepsilon_Q, \tag{16}$$

for $\gamma \leq \gamma_{\max}$ (defined in Equation (18)*), where $K_{\max}$ is defined in Equation* (19)*.*

We defer the proof to Section 11.4.1, where the constants described are:

$$\tau_{\max} \triangleq (1 + \tau^\star)/2 \tag{17}$$

$$\gamma_{\max} \triangleq \operatorname*{argmax}_{\gamma < \alpha} \{\gamma : \tau^\star(\theta^\star, P, \alpha - \gamma) \leq \tau_{\max}\} \tag{18}$$

$$K_{\max} \triangleq \frac{(1+e)}{2m\pi \arccos(\tau_{\max})} \tag{19}$$

With these stability arguments in hand, we can now analyze the performance of SCOUT.

5 The good event

As is common practice in Multi-Armed Bandit analyses, we define a "good event" under which all concentration arguments hold, and condition on this event for the remainder of our analysis. To this end, we first define a collection of high-probability events under which our algorithm performs as anticipated.

We begin by showing that the confidence intervals $\mathcal{C}_t$ are valid, i.e., $\theta^\star \in \mathcal{C}_t$ for all t (utilizing the results from [17]), and then prove that we have collected sufficiently many samples that their radius decays at the desired rate. Although we cannot determine the exact distribution of the context, label pair samples to estimate $\theta^\star$, we can demonstrate that our policy is pessimistic and tests whenever the optimal policy would do so. By Assumption 1 the probability that the optimal policy tests at any given round is $p^\star$. Recall that $N_\theta^t = |\mathcal{S}_\Theta^t|$ denotes the number of samples (X_s, Y_s) collected to estimate $\theta^\star$ up to round t, and similarly $N_P^t = |\mathcal{S}_P^t|$ for the context estimation. The good event comprises the following constituent events.

Definition 3. *At round t the good event G_t holds that*

1. *G_θ: Defined in Lemma 1, the confidence sets $\mathcal{C}_t$ are valid in that $\theta^\star \in \mathcal{C}_t$ for all t.*
2. *$G_{p_{err}}$: The estimates of p_{err} on $\mathcal{Q}_\theta \times \mathcal{Q}_\tau$ are ζ_t accurate (Lemma 2).*
3. *$G_t^{(2)}$: the confidence sets $\mathcal{C}_t$ gets enough samples, that is $N_\theta^t \succeq p^\star t - \sqrt{t\log(t/\delta)} \succeq p^\star t/2$.*
4. *$G_t^{(3)}$: The minimum eigenvalue of the empirical covariance matrix formed by our testing policy grows linearly in t. Let $\lambda_{\min}^t \triangleq \lambda_{\min}\left(\sum_{s\in\mathcal{S}_\Theta^t} X_s X_s^\top\right)$. Then, without loss of generality, it holds for the same constant T_0 that for all $t \geq T_0$:*

$$\lambda_{\min}^t \geq \frac{p^\star t \lambda_0}{8}.$$

Let $G^{(i)} = \cap_{t=1}^T G_t^{(i)}$ for $i = 2, 3$. The good event G is the intersection of $G^{(i)}$ with $G_{p_{err}}$, i.e. $G = G_\theta \cap G^{(2)} \cap G^{(3)} \cap G_{p_{err}}$.

The first event, $\mathbb{P}(G_\theta \geq 1 - \delta$, follows from Lemma 1, i.e. the concentration inequality proven by [17]. The second one, $\mathbb{P}(G_{p_{\text{err}}}) \geq 1 - \delta$, is proved by Lemma 4. Then, we have to prove that $G^{(2)}$ holds with high probability. To do so, we utilize the fact that when the optimal policy tests, then when G_θ and $G_{p_{\text{err}}}$ hold our policy does the same, as proved in Lemma 8. Observe that on $G_t^{(2)}$, we have that $N_\theta^t \geq p^\star t/2$ for all $t \geq T_0$ for some constant T_0 (only a function of δ). For the last event, $\mathbb{P}(G^{(3)}) \geq 1 - 2\delta$ we use a covering argument to bound the minimum eigenvalue of the covariance matrix (see e.g. Section 4.4 of [53] for additional details). Then, we use $G^{(2)}$ as a lower bound of the number of samples collected to construct the empirical covariance matrix and the union bound. We see that G occurs with high probability in the following lemma.

Lemma 7.

$$\mathbb{P}(G) = \mathbb{P}(G_\theta \cap G^{(2)} \cap G^{(3)} \cap G_{p_{err}}) \geq 1 - 6\delta.$$

Detailed proofs are deferred to Section 13.

6 From Stability Analysis to Algorithmic Decisions and Safety

We design our testing rule based on two main principles. First, our testing rule must be "pessimistic", in that when the baseline policy tests, our policy does the same, even for the worst possible $\theta^\star$. Second, our testing rule must be computationally efficient. The second trivially follows from our stated algorithm, and we prove the first in the following lemma:

Lemma 8. *The testing rule Z_t defined in Algorithm 1 satisfies, conditioned on $G_{p_{err}}$ and G_θ, that $Z_t^\star = 1 \implies Z_t = 1$, i.e. $Z_t \geq Z_t^\star$ a.s.*

We defer the proof to Section 12.1. Another property of our testing rule is that it makes no additional errors beyond the baseline policy, on the good event $G_{p_{\text{err}}}$. As formalized in the following lemma, our algorithm makes predictions identical to those of the oracle policy when it does not test.

Lemma 9. *Let $\hat{Y}_t$ the prediction of our policy, where $Y_t^\star$ is the prediction of the oracle baseline policy. On the good event $G_{p_{err}}$, when $Z_t = 0$ (which implies that $Z_t^\star = 0$) then $\hat{Y}_t = \hat{Y}_t^\star$.*

We defer the proof to Section 12.2.

Having developed and motivated our testing rule we now formally prove that it satisfies our desired (α, δ)-safety guarantees, utilizing Lemma 8 and Lemma 9. In the first lemma, we proved that when the baseline policy tests, our policy tests too. In the second one, we proved that when the baseline policy predicts, our policy outputs the same prediction.

More formally, we define the Bernoulli random variable $\xi_t = \mathbb{1}\{\hat{Y}_t \neq Y_t\}$, that denotes whether the algorithm made a mistake at round t, and $\xi_t^\star = \mathbb{1}\{Y_t^\star \neq Y_t\}$ respectively for the baseline policy. When the algorithm tests (i.e. $Z_t = 1$) then we observe the label and it holds that $\xi_t = 0$. Conditioning on the good event G, $\xi_t \leq \xi_t^\star$ almost surely (formalized in Section 12.3). This implies a total error probability bound, stated in the following lemma.

Lemma 10. *On the good event G (Definition 3) the total error probability of Algorithm 1 is upper bounded by α with probability at least $1 - \delta$.*

We defer the proof to Section 12.3.

7 Regret Analysis

With safety in place, we now show that our algorithm achieves sublinear regret. To derive a regret bound, we begin by analyzing the regret at an arbitrary round $t > T_0$ in Lemma 14. We defer the proof to Section 12.4. Summing this lemma over rounds T yields the following overall Theorem.

Theorem 2. *Algorithm 1 satisfies (α, δ)-safety, and has safe regret (see Definition 2) at most*

$$T_0 + \frac{C}{\arccos\left(\tau^\star\right)} \sqrt{\frac{(d + \log(1/\delta))T \log(T)}{p^\star \lambda_0}},$$

for some absolute constant C.

The proof can be found at Section 12.5. Note that δ can even scale exponentially in T and the algorithm will still have sublinear regret. In the linear bandits literature, the dependency on the dimension is $\times(d)$. In our analysis, this extra $\mathcal{O}(\sqrt{d})$ is hidden inside the $1/\sqrt{\lambda_0}$ term where in the case that contexts are uniformly distributed over the unit sphere, $\lambda_0 = \Theta(1/\sqrt{d})$.

8 Numerical results

We corroborate our theoretical guarantees with numerical simulations, to show that our algorithm is able to efficiently compute the testing rule, and converge to the optimal error rate. We generate simulations varying the dimensionality and the target error rate α, showing the rapid convergence of our method when $p^\star$ is large. We see that in all instances our algorithm maintains the desired error rate, and has sublinear regret. Experiments were run on a 2023 Macbook Pro, and took under 5 minutes.

8.1 Modifications from written algorithm

For our numerical simulations, we implemented a version of SCOUT with a few minor modifications from Algorithm 1 to enable it to run faster in practice. These changes are common in practical applications of online learning algorithms to balance theoretical rigor with performance.

Batched Parameter Updates: as written, SCOUT updates the parameter estimate and the testing threshold at every time step t. In a setting with a large time horizon T, re-running the estimation procedures on ever-growing datasets at each step is computationally prohibitive, and wasteful as these will not change too much iteration to iteration. Instead, our implementation updates these estimates only periodically. Concretely, the estimates for θ and τ are cached and reused for a block of subsequent time steps. The frequency of these updates is decreased as the simulation progresses, reflecting the gradual convergence of the parameters.

Simplified Testing Condition: The testing rule in Equation (9) is given by $\langle X_t, \theta_t^L \rangle| \leq \tau_t$. This incorporates several uncertainty terms derived from our theoretical analysis. While crucial for the regret bounds, computing these quantities at every step is not necessary in practice, and the same performance can be obtained by simply collapsing these terms into a) the τ estimate, and b) a bound on $B_t \|X_t\|_{V_t^{-1}}$ (note that in practice this second term may not be known, as it will depend on λ_0, which SCOUT will adapt to). The testing decision becomes $Z_t = 1$ if $|\langle X_t, \theta_t^L \rangle|$ is less than the sum of these two terms.

Omission of the Projection Step: Our theoretical analysis utilizes two estimators. First, the regularized maximum likelihood estimator $\hat{\theta}_t = \operatorname{argmax}_{\theta \in \mathbb{R}^d} \mathcal{L}_t(\theta)$, where $\mathcal{L}_t(\theta)$ is the regularized log-likelihood. Second, for analysis purposes, a projection of this estimator, θ_t^L, is defined in Equation (6). This projection is in practice unneeded, and so we simply utilize $\hat{\theta}_t$ as our θ estimate.

In addition, we reduce the leading constants e.g. in the B_t bound. These adjustments allowed the algorithm to run efficiently while retaining the core principles of SCOUT. The empirical results, which demonstrate sublinear regret and adherence to the safety constraint, validate that these practical simplifications do not compromise the algorithm's performance in our simulated environments.

9 Discussion

In this work we introduced SCOUT, the first algorithm that provably balances **no-regret learning** with a **high-probability safety guarantee** on the empirical misclassification rate in logistic bandits. Our analysis shows that a simple, efficiently-computable testing rule suffices to achieve the order optimal $\widetilde{O}\big(\sqrt{dT/\lambda_0}\big)$ excess-test rate. The empirical results confirm that these bounds translate to practice on moderately large horizons.

In medical triage — our motivating use-case — SCOUT can be viewed as a "test-or-treat" policy that automatically calibrates how aggressively to screen as new evidence accrues. Because the policy is pessimistic by design, it never tests less than an oracle baseline that knows both the patient distribution and the ground-truth regression coefficients. This property is attractive in any high-stakes domain where misclassifications are costly (e.g. credit risk, fraud detection, or industrial quality control).

There are several straightforward theoretical extensions. First is anytime guarantees: replacing the fixed-horizon union bounds with stitched confidence sequences yields an anytime variant with identical regret up to log factors. Second is unequal Type-I / Type-II control. The threshold-selection step can be split to cap false positives and false negatives separately by using two one-sided versions of (2). Finally, here we utilized simple confidence bounds for our logistic bandits. Plugging the recent radius-free concentration results of [37] into Lemma 1 removes the κ factor in B_t.

There are several exciting directions of future work that are motivated by this work. First, we have the setting where the optimal baseline does not need to test, i.e. $p^\star = 0$. If the optimal policy never tests, can one detect *fast enough* that screening is unnecessary while still retaining the high-probability safety constraint? The second. is adversarial contexts, or any nonstationary context distribution. Can the ideas behind SCOUT be combined with online calibration tools to handle non-stationary or even adversarial X_t? Another consideration is to follow the line of work of conservative bandits [34] and, given a fixed baseline policy as input to our problem that satisfies the constraints, to compute a feasible policy for the problem that is competitive with the baseline policy.

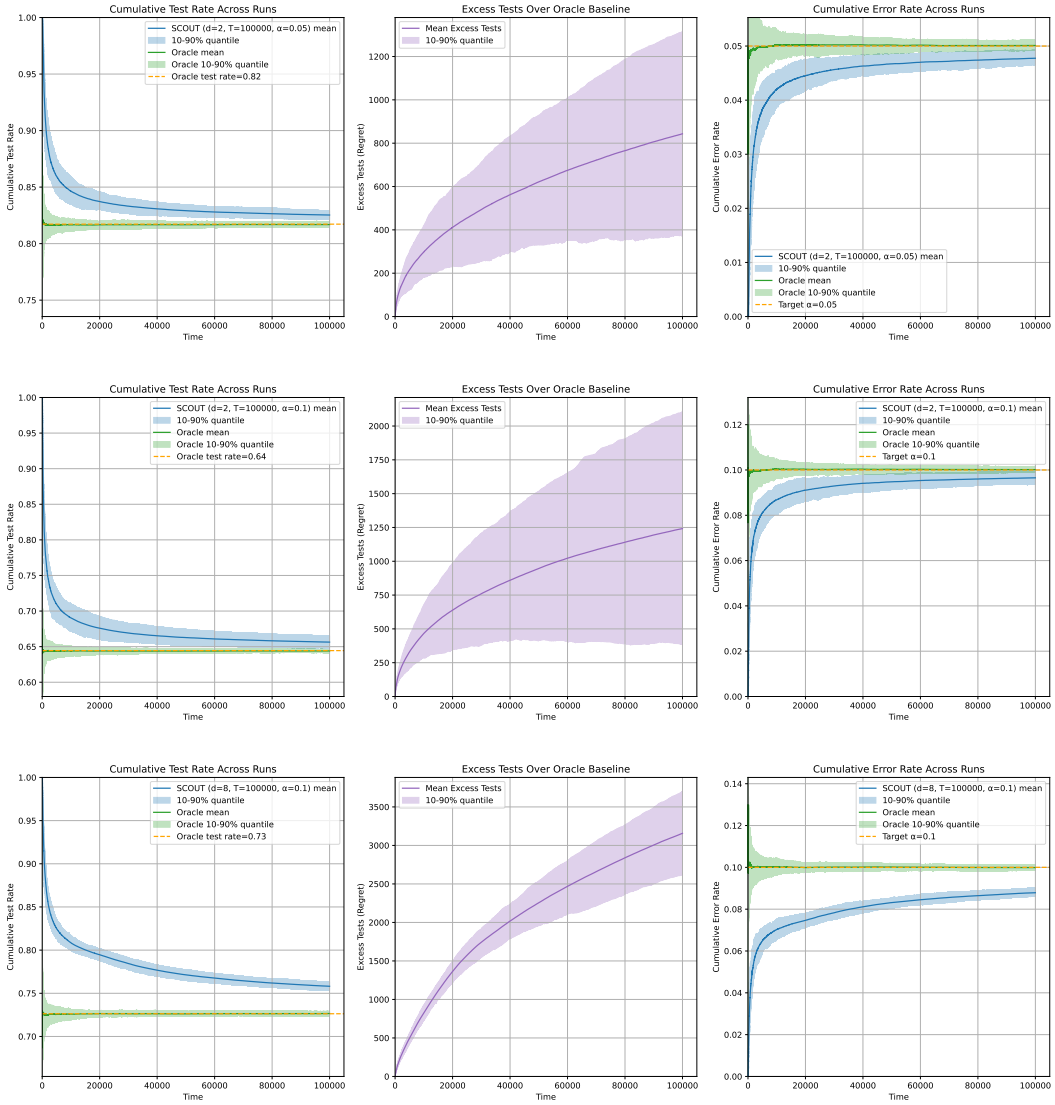

Figure 1: Simulation results. Plots in the first and second row correspond to$d = 2$, where the first row shows $\alpha = 0.05$, and the second $\alpha = 0.1$. Third row shows $d = 8, \alpha = 0.1$. In each row, the left plot shows the cumulative test rate as a function of round, where blue shows the performance of SCOUT (10-90%) quantiles shaded, with the oracle test rate shown in orange at $p^\star$ (empirical test rate for optimal threshold policy plotted in green). The middle plots show the excess number of tests, demonstrating the sublinear regret of SCOUT. The right plots show the misclassification rate of SCOUT, where we see that while the optimal baseline policy fluctuates around the desired threshold α, SCOUT starts far below, then learns to be more aggressive and increases its guessing, eventually approaching misclassification rate α. However, it never exceeds it (small δ chosen).

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

[3] M. M. Afsar, T. Crump, and B. Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

[4] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

[5] G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[6] H. Bastani, K. Drakopoulos, V. Gupta, J. Vlachogiannis, C. Hadjichristodoulou, P. Lagiou, G. Magiorkinis, D. Paraskevis, and S. Tsiodras. Interpretable operations research for high-stakes decisions: Designing the greek covid-19 testing system. *INFORMS Journal on Applied Analytics*, 52(5):398–411, 2022.

[7] N. Cesa-Bianchi, C. Gentile, L. Zaniboni, and M. Warmuth. Worst-case analysis of selective sampling for linear classification. *Journal of Machine Learning Research*, 7(7), 2006.

[8] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[9] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: a survey. *Internet mathematics*, 3(1):79–127, 2006.

[10] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 67–82. Springer, 2016.

[11] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan. Machine learning in drug discovery: a review. *Artificial intelligence review*, 55(3):1947–1999, 2022.

[12] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *International conference on computational learning theory*, pages 249–263. Springer, 2005.

[13] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1):2655–2697, 2012.

[14] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Efficiently learning halfspaces with tsybakov noise. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 88–101, 2021.

[15] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning halfspaces with massart noise under structured distributions. In *Conference on Learning Theory*, pages 1486–1513. PMLR, 2020.

[16] Y. Duan, Z. Zhao, L. Qi, L. Zhou, L. Wang, and Y. Shi. Towards semi-supervised learning with non-random missing labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16121–16131, 2023.

[17] L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

[18] L. Faury, M. Abeille, K.-S. Jun, and C. Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 546–580. PMLR, 2022.

[19] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

[20] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28:133–168, 1997.

[21] A. Gangrade, T. Chen, and V. Saligrama. Safe linear bandits over unknown polytopes. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1755–1795. PMLR, 2024.

[22] A. Gangrade, A. Kag, A. Cutkosky, and V. Saligrama. Online selective classification with limited feedback. *Advances in Neural Information Processing Systems*, 34:14529–14541, 2021.

[23] A. Gangrade, A. Kag, and V. Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2179–2187. PMLR, 2021.

[24] P. Giudici. Safe machine learning. *Statistics*, 58(3):473–477, 2024.

[25] S. Goel, S. Hanneke, S. Moran, and A. Shetty. Adversarial resilience in sequential prediction via abstention. *Advances in Neural Information Processing Systems*, 36:8027–8047, 2023.

[26] J. A. Grant and D. S. Leslie. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification. *arXiv preprint arXiv:2109.14412*, 2021.

[27] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[28] V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[29] S. Hanneke and L. Yang. Toward a general theory of online selective sampling: Trading off mistakes and queries. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2021.

[30] K. Harris, C. Podimata, and S. Z. Wu. Strategic apple tasting. *Advances in Neural Information Processing Systems*, 36:79918–79945, 2023.

[31] D. P. Helmbold, N. Littlestone, and P. M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.

[32] K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

[33] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[34] A. Kazerouni, M. Ghavamzadeh, Y. Abbasi Yadkori, and B. Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

[35] A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

[36] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[37] J. Lee, S.-Y. Yun, and K.-S. Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *Advances in Neural Information Processing Systems*, 37:124640–124685, 2025.

[38] N. A. Mehta, J. Komiyama, V. K. Potluru, A. Nguyen, and M. Grant-Hagen. Thresholded linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6968–7020. PMLR, 2023.

[39] F. Orabona, N. Cesa-Bianchi, et al. Better algorithms for selective sampling. In *Proceedings of the 28th international conference on machine learning: Bellevue, Washington, USA, june 28. july 2, 2011*, pages 433–440. Omnipress, 2011.

[40] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

[41] A. Pacchiano, S. Singh, E. Chou, A. Berg, and J. Foerster. Neural pseudo-label optimism for the bank loan problem. *Advances in Neural Information Processing Systems*, 34:6580–6593, 2021.

[42] M. Papini, A. Tirinzoni, M. Restelli, A. Lazaric, and M. Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pages 8371–8380. PMLR, 2021.

[43] M. Pinelis and D. Ruppert. Machine learning portfolio allocation. *The Journal of Finance and Data Science*, 8:35–54, 2022.

[44] V. Raman, U. Subedi, A. Raman, and A. Tewari. Revisiting the learnability of apple tasting. *arXiv preprint arXiv:2310.19064*, 2023.

[45] V. Raman, U. Subedi, A. Raman, and A. Tewari. Apple tasting: Combinatorial dimensions and minimax rates. *Proceedings of Machine Learning Research vol*, 247:1–23, 2024.

[46] A. Sekhari, K. Sridharan, W. Sun, and R. Wu. Selective sampling and imitation learning via online regression. *Advances in Neural Information Processing Systems*, 36:67213–67268, 2023.

[47] B. Settles. Active learning literature survey. 2009.

[48] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[49] A. Slivkins. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*, 2013.

[50] R. S. Sutton, A. G. Barto, et al. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134, 1999.

[51] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[52] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.

[53] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[54] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[55] J. Yao, E. Brunskill, W. Pan, S. Murphy, and F. Doshi-Velez. Power constrained bandits. In *Machine Learning for Healthcare Conference*, pages 209–259. PMLR, 2021.

[56] Z. Zhu and B. Van Roy. Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3636–3646, 2023.

10 Baseline policy

Here we provide some discussion and proofs regarding the optimal baseline we compare to.

10.1 Proof of Proposition 1

Proof. When the value of the parameter $\theta^\star$ and the collection of the contexts $\{X_t\}_{t=1}^T$ are known, we can equivalently write the problem as follows. Let $p_t = \mu(X_t^\top \theta^\star)$, the labels $Y_t \sim Ber(p_t)$ independently across t.

To compute the expected error, that is $\mathbb{E}(E_t) \triangleq \mathbb{E}(\mathbb{1}\{\hat{Y}_t \neq Y_t\})$, we only need to examine the case where we do not test. When we do test, we observe the true label and incur zero error. For $Z_t = 0$ then, the expected error is

1. If $\hat{Y}_t = 1$ then $\mathbb{E}(\mathbb{1}\{\hat{Y}_t \neq Y_t\} \mid \hat{Y}_t = 1) = 1 - p_t$.
2. Else if $\hat{Y}_t = 0$ then $\mathbb{E}(\mathbb{1}\{\hat{Y}_t \neq Y_t\} \mid \hat{Y}_t = 0) = p_t$.

The optimal policy then is to output the prediction with the smallest error. The expected error then is equal to

$$\mathbb{E}(\mathbb{1}\{\hat{Y}_t \neq Y_t\}) \triangleq \min\{1 - p_t, p_t\}.$$

We denote $\mathbf{P}(Z_t = 0) = \eta_t$. The optimal policy choice is reduced to the following optimization problem.

$$\min_{\{Z_t\}} \sum_{t=1}^T 1 - \eta_t \quad \text{s.t.} \quad \frac{1}{T}\sum_{t=1}^T \min\{1 - p_t, p_t\}\eta_t \leq \alpha, \quad 0 \leq \eta_t \leq 1. \tag{20}$$

Or equivalently can be written as.

$$\max_{\{Z_t\}} \sum_{t=1}^T \eta_t \quad \text{s.t.} \quad \frac{1}{T}\sum_{t=1}^T \min\{1 - p_t, p_t\}\eta_t \leq \alpha, \quad 0 \leq \eta_t \leq 1. \tag{21}$$

The solution of this Linear Program is the solution of the *Fractional Knapsack* problem with budget α. Solving optimally this problem, requires applying a greedy strategy that is to sort the coefficients $\min\{1 - p_t, p_t\}$ in an non-increasing order and assign $\eta = 1$ to the lowest "error" contexts until we do not violate the budget constraint α. This strategy is clearly a threshold strategy that depends on a.

□

10.2 Discussion of Assumption 3

Assumption 3 requires the covariance matrix of contexts selected by the optimal policy to be positive definite. We now demonstrate that under the distributional assumption in Assumption 2, this positive definiteness condition is indeed satisfied. While this result does not directly imply Assumption 3, it establishes that even a uniform testing policy would fulfill this eigenvalue requirement. That is summarized in the following lemma.

Lemma 11. *Let λ_0 the minimum eigenvalue of $\mathbb{E}_{\mathbf{x}\sim P}\mathbf{x}\mathbf{x}^\top$, then it is true that $\lambda_0 > 0$.*

Proof. We have assumed that all contexts lie on the unit sphere with $\|\mathbf{x}\|_2 = 1$. Then by Assumption 2, defining $\mathcal{B} \triangleq \mathcal{S}^{d-1}$:

Lemma 12. *Let $\Sigma = \mathbb{E}_{\mathbf{x}\sim P}\mathbf{x}\mathbf{x}^\top$ and $\Sigma_{tr} = \int_{\mathcal{B}} \mathbf{x}\mathbf{x}^\top m d\mathbf{x}$. For any arbitrary $\mathbf{v} \in \mathbb{R}^d$ it holds that*

$$\mathbf{v}^\top \Sigma \mathbf{v} \geq \mathbf{v}^\top \Sigma_{tr} \mathbf{v}.$$

Proof. We can write $\mathbf{v}^\top \Sigma \mathbf{v}$ as follows

$$\mathbf{v}^\top \Sigma \mathbf{v} = \mathbb{E}_{\mathbf{x}\sim P} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} \tag{22}$$
$$= \mathbb{E}_{\mathbf{x}\sim P} (\mathbf{x}^\top \mathbf{v})^2, \tag{23}$$

and analogously $\mathbf{v}^\top \Sigma_{tr} \mathbf{v}$ as

$$\mathbf{v}^\top \Sigma_{tr} \mathbf{v} = \int_{\mathcal{B}} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} m d\mathbf{x} \tag{24}$$
$$= m \int_{\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 d\mathbf{x}. \tag{25}$$

By using our assumption that $p(\mathbf{x}) \geq m > 0$ we derive that for all $\mathbf{x} \in \mathcal{B}$

$$(\mathbf{x}^\top \mathbf{v})^2 p(\mathbf{x}) \geq (\mathbf{x}^\top \mathbf{v})^2 m. \tag{26}$$

Integrating over the entire domain yields that:

$$\implies \int_{x\in\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 p(\mathbf{x}) d\mathbf{x} \geq \int_{x\in\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 m d\mathbf{x} \tag{27}$$
$$\mathbf{v}^\top \Sigma \mathbf{v} \geq \mathbf{v}^\top \Sigma_{tr} \mathbf{v} \tag{28}$$

□

The previous lemma applies for any arbitrary vector $\mathbf{v}$, so $\Sigma \succeq \Sigma_{tr}$. Let $(\lambda_0, \mathbf{v}_0)$ the eigen-pair of the corresponding minimum eigenvalue of Σ. Let us apply the previous result for $\mathbf{v}_0$. Then:

$$\lambda_0 \|\mathbf{v}_0\|_2^2 \geq m \int_{\mathcal{B}} (\mathbf{x}^\top \mathbf{v}_0)^2 d\mathbf{x} \tag{29}$$

Let $V_d(r)$ the volume of the *d-dimensional* ball with radius r. The density of the uniform distribution of a *d-dimensional* ball with radius r is $1/V_d(r)$ in the interior of the ball and zero outside. By multiplying and dividing on the right hand side of the previous inequality with $V_d(1)$ we derive that

$$\lambda_0 \|\mathbf{v}_0\|_2^2 \geq m V_d(1) \int_{\mathcal{B}} \frac{(\mathbf{x}^\top \mathbf{v}_0)^2}{V_d(1)} d\mathbf{x} \tag{30}$$
$$= m V_d(1) \int_{\|\mathbf{x}\|_2^2 \leq 1} \frac{(\mathbf{x}^\top \mathbf{v}_0)^2}{V_d(1)} d\mathbf{x} \tag{31}$$

The quantity $\int_{\|\mathbf{x}\|_2^2 \leq 1} \frac{(\mathbf{x}^\top \mathbf{v}_0)^2}{V_d(1)} d\mathbf{x}$ is equal to $\mathbb{E}[\langle \mathbf{x}, \mathbf{v}_0 \rangle^2]$ when $\mathbf{x}$ is uniformly distributed over the unit *d-dimensional* ball. This quantity can equivalently be written as

$$\mathbb{E}[\langle \mathbf{x}, \mathbf{v}_0 \rangle^2] = \mathbb{E}[\mathbf{v}_0^\top \mathbf{x}\mathbf{x}^\top \mathbf{v}_0]$$
$$= \mathbf{v}_0^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \mathbf{v}_0$$

The quantity $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is the covariance matrix of the uniform over the unit *d-dimensional* ball. This matrix can be written as $a\mathbf{I}_d$ due to spherical symmetry.

This is because, by a change of variables, we can obtain that $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = -\mathbb{E}[\mathbf{x}_i \mathbf{x}_j]$ for $i \neq j$, implying that $\mathbb{E}[\mathbf{x}_i \mathbf{x}_j] = 0$.

To compute the diagonal entries:

$$
\begin{aligned}
\mathbb{E}[x_i^2] &= \frac{1}{d}\mathbb{E}[\mathbf{x}^2] \\
&= \frac{1}{d}\int_{\|\mathbf{x}\|_2^2 \le 1} \frac{\mathbf{x}^2}{V_d(1)} d\mathbf{x} \\
&= \frac{1}{dV_d(1)}\int_{\mathcal{S}^{d-1}}\int_{0\le r\le 1} r^2 r^{d-1} dr d\sigma(\omega) \\
&= \frac{S_d(1)}{V_d(1)}\frac{1}{d(d+2)},
\end{aligned}
$$

where $S_d(1)$ is the surface of the unit sphere and $d\sigma$ any surface measure.

By combining them all we derive

$$
\lambda_0 \|\mathbf{v}_0\|_2^2 \ge \frac{mV_d(1)S_d(1)}{d(d+2)V_d(1)} \|\mathbf{v}_0\|_2^2 \tag{32}
$$

$$
\lambda_0 \ge \frac{mS_d(1)}{d(d+2)} > 0. \tag{33}
$$

□

11 Stability analysis of $p_{\text{err}}(\theta, \rho, \tau)$

11.1 Stability of $\tau^\star$ with respect to $\hat{P}_t$

11.1.1 Proof of Lemma 2

Proof. First, we collect a context as a sample at every odd round, so at round t it holds that $|\mathcal{S}_P^t| = \lceil t/2 \rceil$. Indexing these samples as x_i, we can write the empirical error $p_{\text{err}}(\theta, \hat{P}_T, \tau)$ as follows:

$$
\begin{aligned}
p_{\text{err}}(\theta, \hat{P}_T, \tau) - p_{\text{err}}(\theta, P, \tau) &= \int (1+\exp(|x^\top\theta|))^{-1} \mathbb{1}\left\{|x^\top\theta| > \tau\right\} \hat{P}_t(dx) - p_{\text{err}}(\theta, P, \tau) \\
&= \frac{1}{\lceil t/2 \rceil}\sum_{i=1}^{\lceil t/2 \rceil} \left((1+\exp(|x_i^\top\theta|))^{-1}\mathbb{1}\left\{|x_i^\top\theta| > \tau\right\} - p_{\text{err}}(\theta, P, \tau)\right)
\end{aligned} \tag{34}
$$

As $0 \le (1+\exp(z))^{-1} \le \frac{1}{2}$, the summands are i.i.d. [0,1/2] random variables with mean 0, so we can apply Hoeffding's inequality [54]:

$$
\mathbb{P}\left(\left|\frac{1}{\lceil t/2 \rceil}\sum_{i=1}^{\lceil t/2 \rceil} \left((1+\exp(|x_i^\top\theta|))^{-1}\mathbb{1}\left\{|x_i^\top\theta| > \tau\right\} - p_{\text{err}}(\theta, P, \tau)\right)\right| \ge \sqrt{\frac{\log(2/\delta')}{4t}}\right) \le \delta'.
$$

By taking the union bound over all rounds $t \ge 1$ and setting $\delta' \triangleq \frac{6\delta}{\pi^2 t^2}$ we derive:

$$
\mathbb{P}\left(\left|\frac{1}{\lceil t/2 \rceil}\sum_{i=1}^{\lceil t/2 \rceil} \left((1+\exp(|x_i^\top\theta|))^{-1}\mathbb{1}\left\{|x_i^\top\theta| > \tau\right\} - p_{\text{err}}(\theta, P, \tau)\right)\right| \le \sqrt{\frac{\log\left(\frac{\pi^2 t^2}{3\delta}\right)}{4t}}, \forall t : t \ge 1\right) \ge 1-\delta.
$$

Here, we apply the well-known result for the Basel series: $\sum_{t=1}^{\infty}\frac{1}{t^2} = \frac{\pi^2}{6}$.

□

11.2 Stability of $\tau^\star$ with respect to θ

11.2.1 Proof of Lemma 3

Proof. Here, we use x as a dummy variable for integration:

$$\begin{aligned}
p_{\text{err}}(\theta,\rho,\tau) &= \int (1+\exp(|x^\top\theta|))^{-1}\mathbb{1}\left\{|x^\top\theta|>\tau\right\}\rho(dx)\\
&= \int (1+\exp(|x^\top\theta'+x^\top(\theta-\theta')|))^{-1}\mathbb{1}\left\{|x^\top\theta'+x^\top(\theta-\theta')|>\tau\right\}\rho(dx)\\
&\le \int (1+\exp(|x^\top\theta'|-|x^\top(\theta-\theta')|))^{-1}\mathbb{1}\left\{|x^\top\theta'|>\tau-|x^\top(\theta-\theta')|\right\}\rho(dx)\\
&\le \int \left((1+\exp(|x^\top\theta'|))^{-1}+|x^\top(\theta-\theta')|\right)\mathbb{1}\left\{|x^\top\theta'|>\tau-|x^\top(\theta-\theta')|\right\}\rho(dx)\\
&\le \max_{x'\in\mathcal{X}} \int \left((1+\exp(|x^\top\theta'|))^{-1}+|x'^\top(\theta-\theta')|\right)\mathbb{1}\left\{|x^\top\theta'|>\tau-|x'^\top(\theta-\theta')|\right\}\rho(dx)\\
&= \max_{x'\in\mathcal{X}} p_{\text{err}}(\theta',\rho,\tau-|x'^\top(\theta-\theta')|)+\int |x^\top(\theta-\theta')|\mathbb{1}\left\{|x^\top\theta'|>\tau-|x'^\top(\theta-\theta')|\right\}\rho(dx)\\
&\le \max_{x'\in\mathcal{X}} p_{\text{err}}(\theta',\rho,\tau-\|\theta-\theta'\|_{V_t}\|x'\|_{V_t^{-1}})+\|\theta-\theta'\|_{V_t}\|x'\|_{V_t^{-1}}\mathbb{P}_\rho\left(|x^\top\theta'|>\tau-|x^\top(\theta-\theta')|\right)\\
&\le \max_{x'\in\mathcal{X}} p_{\text{err}}(\theta',\rho,\tau-\|\theta-\theta'\|_{V_t}\|x'\|_{V_t^{-1}})+\|\theta-\theta'\|_{V_t}\|x'\|_{V_t^{-1}}\\
&= p_{\text{err}}(\theta',\rho,\tau-\frac{\|\theta-\theta'\|_{V_t}}{\sqrt{\lambda^t_{\min}}})+\frac{\|\theta-\theta'\|_{V_t}}{\sqrt{\lambda^t_{\min}}}
\end{aligned}$$

The first inequality follows from the triangle inequality, and the second inequality follows from the fact that $1/(1+\exp(z))$ is 1/4-Lipschitz (coarsely upper bounded as 1). The third bounds by looking at the worst case context x'. The fourth inequality utilizes Hölder's inequality, on the worst case context x', and that p_{err} is monotone in τ. The second to last inequality follows from the fact that a probability is always less than or equal to 1. Finally, we apply the following bound for any $x'\in\mathcal{X}$; $\|x'\|_{V_t^{-1}}\le\frac{1}{\sqrt{\lambda^t_{\min}}}$, where we have implicitly used that $\|x'\|\le 1,\forall x'\in\mathcal{X}$.

□

11.2.2 Proof of Lemma 4

Proof. To extend Lemma 2 to hold simultaneously for all $\theta_Q\in\mathcal{Q}_\theta$ and $\tau_Q\in\mathcal{Q}_\tau$, we define an ε_Q-net for each, and union bound over their cartesian product. By Lemma 2 we know that for any fixed θ,τ:

$$\mathbb{P}\left(\left|p_{\text{err}}(\theta,\hat{P}_t,\tau)-p_{\text{err}}(\theta,P,\tau)\right|\le\sqrt{\frac{\log(\frac{\pi^2t^2}{3\delta})}{4t}},\forall t\ge 1\right)\ge 1-\delta.$$

Let $\mathcal{Q}_\theta=\mathcal{N}(\mathcal{S}^{d-1},\varepsilon_\theta)$ an ε_Q-cover of the unit sphere $\mathcal{S}^{d-1}$. By **Corollary 4.2.13** at [53] we have that the covering numbers of $\mathcal{S}^{d-1}$ satisfy for any $\varepsilon_Q>0$;

$$\left(\frac{1}{\varepsilon_Q}\right)^d\le|\mathcal{Q}_\theta|\le\left(\frac{2}{\varepsilon_Q}+1\right)^d.$$

For any $\varepsilon_Q<1$ it is true that $|\mathcal{Q}_\theta|\le(\frac{3}{\varepsilon_Q})^d$. Further, as τ lives in $[0,1]$, an ε-net of the unit segment in the real line is $\{\epsilon,2\epsilon,\dots,\lfloor\frac{1}{\epsilon}\rfloor\epsilon\}$, and so $|\mathcal{Q}_\tau|\le\frac{1}{\varepsilon_\tau}$. By taking the union bound over all $\tau_Q\in\mathcal{Q}_\tau$ and all $\theta_Q\in\mathcal{Q}_\theta$, i.e. taking $\delta'=\delta/(|\mathcal{Q}_\theta|\cdot|\mathcal{Q}_\tau|)$, we have

$$\mathbb{P}\left(\left|p_{\text{err}}(\theta_Q,\hat{P}_t,\tau_Q)-p_{\text{err}}(\theta_Q,P,\tau_Q)\right|\le\zeta_t,\forall t\ge 1,\theta_Q\in\mathcal{Q}_\theta,\tau_Q\in\mathcal{Q}_\tau\right)\ge 1-\delta.$$

Recall that ζ_t is defined in Equation (10) as

$$\zeta_t\triangleq\sqrt{\frac{d\log(3/\varepsilon_Q)+\log(\frac{\pi^2t^2}{3\delta})}{4t}}.$$

We choose ε_Q to be sufficiently small with respect to T, any $\varepsilon_Q = o(1/T)$ suffices. For concreteness, we choose $\varepsilon_Q = 1/T^2$ as this simplifies the analysis in Theorem 2, but anytime choices like $\varepsilon_Q^t = 1/t^2$ work as well. □

11.3 Proof of Lemma 5

Proof. Conditioning on the good event $G_{p_{\text{err}}}$, we have that

$$\begin{aligned}
\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) &= \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, \hat{P}_t, \tau_Q) \leq \alpha\} \\
&\overset{(a)}{\leq} \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, P, \tau_Q) \leq \alpha - \zeta_t\} \\
&\overset{(b)}{\leq} \min\left\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, P, \tau_Q) \leq \alpha - \zeta_t - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right\} + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} \\
&\leq \min\left\{\tau \in [0,1] : p_{\text{err}}(\theta, P, \tau) \leq \alpha - \zeta_t - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right\} + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} + \varepsilon_Q \\
&= \tau^\star\left(\theta, P, \alpha - \zeta_t - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} + \varepsilon_Q \qquad (35)
\end{aligned}$$

Where inequality (a) follows from conditioning on the good event $G_{p_{\text{err}}}$, and (b) follows from the Lemma 3.

The lower bound for $\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha)$ follows more simply:

$$\begin{aligned}
\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) &= \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, \hat{P}_t, \tau_Q) \leq \alpha\} \\
&\overset{(a)}{\geq} \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, P, \tau_Q) \leq \alpha + \zeta_t\} \\
&= \tau_Q^\star(\theta_Q, P, \alpha + \zeta_t) \\
&\geq \tau^\star(\theta_Q, P, \alpha + \zeta_t),
\end{aligned}$$

where (a) follows by the good event $G_{p_{\text{err}}}$. Now, we will lower bound $\tau^\star(\theta_Q, P, \alpha)$ in terms of $\tau^\star$.

$$\begin{aligned}
\tau^\star(\theta_Q, P, \alpha) &= \min\{\tau \in [0,1] : p_{\text{err}}(\theta_Q, P, \tau) \leq \alpha\} \\
&\overset{(a)}{\geq} \min\left\{\tau \in [0,1] : p_{\text{err}}\left(\theta, P, \tau + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} \leq \alpha\right\} \\
&\geq \min\left\{\tau \in [0,1] : p_{\text{err}}(\theta, P, \tau) \leq \alpha + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right\} - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}} \\
&= \tau^\star\left(\theta, P, \alpha + \frac{\|\theta_Q - \theta^\star\|_{V_t}}{\sqrt{\lambda_{\min}^t}}\right) - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}},
\end{aligned}$$

where (a) follows from Lemma 3, where we lower bound $p_{\text{err}}(\theta_Q, P, \tau) \geq p_{\text{err}}(\theta, P, \tau + \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}) - \frac{\|\theta_Q - \theta\|_{V_t}}{\sqrt{\lambda_{\min}^t}}$.

□

Putting all together we have that on $G_{p_{\text{err}}}, G_\theta$, evaluating at $\theta = \theta^\star$,

$$\begin{aligned}
\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) &= \tau_Q^\star\left(\theta_Q, \hat{P}_t, \alpha - \zeta_t - 2B_t/\sqrt{\lambda_{\min}^t}\right) + 2B_t/\sqrt{\lambda_{\min}^t} \\
&\overset{(a)}{\geq} \tau^\star(\theta^\star, P, \alpha).
\end{aligned}$$

where (a) leverages Lemma 5.

11.4 Stability of $\tau^\star$ with respect to α

We begin by proving a lemma bounding the probability in the annulus $\tau < |X^\top \theta^\star| \leq \tau + \lambda$. With our pessimistic policy, we want to always test when the optimal baseline tests (at τ), and need to ensure that with a slightly larger threshold (at $\tau + \lambda$) we do not test that much more.

Lemma 13. *Under Assumption 2, for all $\lambda > 0$ we have that*

$$m \cdot 2\pi \arccos(\min(1, \tau + \lambda))\lambda \leq \mathbb{P}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right) \leq M \cdot 2\pi \arccos(\tau)\lambda$$

where $\tau_{outer} = \min(\tau + \lambda, 1)$.

Proof. Since the contexts are in $\mathbb{R}^d$ and the density is bounded between m and M, we simply need to upper and lower bound

$$\text{Vol}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right) = \text{Vol}\left(|X^\top \theta^\star| > \tau\right) - \text{Vol}\left(|X^\top \theta^\star| \geq \tau + \lambda\right) \tag{36}$$

where $\|\theta^\star\| = 1$, and X lives on the unit sphere. If $\lambda + \tau > 1$, then the outer edge of the annulus is exactly at 1. While we proceed with the analysis assuming that $\lambda < 1 - \tau$, we concretely take the outer edge of the annulus to be $\tau_{\text{outer}} = \min(\tau + \lambda, 1)$.

Geometrically, this probability is the difference between two sphere caps: one with radius $\arccos(\tau)$ and one with $\arccos(\tau_{\text{outer}})$. Using the fact that the density is bounded between m and M, we can upper bound the surface area of the annulus by the rectangular strip with height λ and width $2\pi \arccos(\tau)$, or lower bound by $2\pi \arccos(\tau_{\text{outer}})$.

Thus, we have that

$$m \cdot 2\pi \arccos(\tau_{\text{outer}})\lambda \leq \mathbb{P}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right) \leq M \cdot 2\pi \arccos(\tau)\lambda \tag{37}$$

□

11.4.1 Proof of Lemma 6

Restating the lemma:

Lemma (Restating Lemma 6)**.** *Under Assumptions 1 and 2,*

$$\tau^\star(\theta^\star, P, \alpha - \gamma) \leq \tau^\star(\theta^\star, P, \alpha) + K_{\max}\gamma + \varepsilon_Q, \tag{38}$$

for $\gamma \leq \gamma_{\max}$ (defined in Equation (41)*), where $K_{\max}$ is defined in Equation* (42)*.*

Proof. We begin by studying the difference between p_{err} evaluated at thresholds τ and $\tau - \lambda$ for arbitrary $\tau < 1$.

$$\begin{aligned}
&p_{\text{err}}(\theta^\star, P, \tau - \lambda) - p_{\text{err}}(\theta^\star, P, \tau) \\
&\qquad = \int (1 + \exp(|x^\top \theta^\star|))^{-1} \mathbb{1}\left\{\tau - \lambda < |x^\top \theta^\star| \leq \tau\right\} P(dx) \\
&\qquad \geq (1 + \exp(\tau))^{-1} \mathbb{P}\left(\tau - \lambda < |X^\top \theta^\star| \leq \tau\right) \\
&\qquad \geq 2m\pi(1 + \exp(\tau))^{-1} \arccos(\tau)\lambda
\end{aligned} \tag{39}$$

where the last inequality follows from Lemma 13.

To show that only a small λ is necessary, we note that τ is a continuous function of α when evaluated at the true distribution P. Since $\tau^\star(\theta^\star, P, \alpha) < 1$ by Assumption 1 we have that $\tau^\star(\theta^\star, P, \alpha - \gamma) < (1 + \tau^\star)/2 \triangleq \tau_{\max} < 1$ for all $\gamma \leq \gamma_{\max}$. This lets us define the maximum proportionality constant $K(\tau)$, which maximizes $\frac{(1+\exp(\tau))}{2m\pi \arccos(\tau)}$ over $\tau \in [0, \tau_{\max}]$. Concretely:

$$\tau_{\max} \triangleq (1 + \tau^\star)/2 \tag{40}$$

$$\gamma_{\max} \triangleq \operatorname*{argmax}_{\gamma < \alpha} \left\{\gamma : \tau^\star(\theta^\star, P, \alpha - \gamma) \leq \tau_{\max}\right\} \tag{41}$$

$$K_{\max} \triangleq \frac{(1 + e)}{2m\pi \arccos(\tau_{\max})} \tag{42}$$

With the proportionality constant $K(\tau)$, we show that $\tau_Q^\star(\theta^\star, P, \alpha-\gamma) \leq \tau^\star(\theta^\star, P, \alpha)+K(\tau)\gamma+\varepsilon_Q$. This enables a uniform bound as:

$$K(\tau) = \frac{(1+\exp(\tau))}{2m\pi \arccos(\tau)} \leq K_{\max}$$

This means that, leveraging Equation (39) with $\lambda = K_{\max}\gamma$, we have that for all $\gamma \leq \gamma_{\max}$:

$$\begin{aligned}
\tau_Q^\star(\theta^\star, P, \alpha-\gamma) &= \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta^\star, P, \tau_Q) \leq \alpha-\gamma\} \\
&\overset{(a)}{\leq} \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, P, \tau_Q - K_{\max}\gamma) \leq \alpha-\gamma+2m\pi(1+\exp(\tau_Q))^{-1}\arccos(\tau_Q)K_{\max}\gamma\} \\
&\overset{(b)}{\leq} \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, P, \tau_Q - K_{\max}\gamma) \leq \alpha\} \\
&\leq \min(1, \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, P, \tau_Q) \leq \alpha\} + K_{\max}\gamma + \varepsilon_Q) \\
&= \min(1, \tau_Q^\star(\theta, P, \alpha) + K_{\max}\gamma + \varepsilon_Q).
\end{aligned}$$

In (a) we leveraged the p_{err} difference bound derived in Equation (39). (b) follows from the definition of $K_{\max}$ in Equation (42), and the monotonicity of τ in α. This expression can be further bounded using that $\arccos(\tau_{\max}) \geq \sqrt{2(1-\tau_{\max})}$.

□

12 Other proofs

12.1 Proof of Lemma 8

Proof. Lower bounding the threshold τ_t we use:

$$\tau_t = \hat{\tau}(\theta_t^L, \hat{P}_t, \alpha_t) + B_t/\sqrt{\lambda_{\min}^t} \geq \tau^\star(\theta^\star, P, \alpha) + B_t/\sqrt{\lambda_{\min}^t}$$

where we use the monotonicity of $\tau^\star$ with respect to α, and that $\alpha_t < \alpha$, in addition to Lemma 5. Then, we upper bound the inner product computed:

$$|\langle X_t, \theta_t^L\rangle| \leq |\langle X_t, \theta^\star\rangle| + \|\theta_t^L - \theta^\star\|_{V_t}\|X_t\|_{V_t^{-1}} \leq |\langle X_t, \theta^\star\rangle| + B_t/\sqrt{\lambda_{\min}^t}$$

By Holder. Combining these together yields:

$$|\langle X_t, \theta^\star\rangle| \leq \tau^\star \quad \Longrightarrow \quad |\langle X_t, \theta_t^L\rangle| \leq \tau_t. \tag{43}$$

□

12.2 Proof of Lemma 9

Proof. When $Z_t = 0$ on $G_{p_{\text{err}}}$ it holds that for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$:

$$|\langle X_t, \theta\rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha) > \varepsilon_Q$$

Using Lemma 4, we know that when $G_{p_{\text{err}}}, G_\theta$ holds, then for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$:

$$\begin{aligned}
&\hat{\tau}(\theta, \hat{P}_t, \alpha) \geq \tau^\star(\theta^\star, P, \alpha) \\
\Longrightarrow\ &|\langle X_t, \theta\rangle| \geq \tau^\star(\theta^\star, P, \alpha) + \varepsilon_Q.
\end{aligned}$$

For any $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$ there exists a $\theta' \in \mathcal{C}_t$ such that $\|\theta' - \theta\| \leq \varepsilon_Q$. Similarly to the previous proof, we can bound then $|\langle X_t, \tilde{\theta}\rangle| \leq |\langle X_t, \theta'\rangle| + \varepsilon_Q$ by the triangle inequality and Cauchy-Schwarz.

Then, it is true that for any $\theta \in \mathcal{C}_t$

$$|\langle X_t, \theta\rangle| \geq \tau^\star(\theta^\star, P, \alpha) > 0.$$

Under G_θ we have that $\theta^\star \in \mathcal{C}_t, \forall t$, and as a consequence.

$$|\langle X_t, \theta^\star\rangle| \geq \tau^\star(\theta^\star, P, \alpha) > 0.$$

□

12.3 Proof of lemma 10

Proof. We analyze the four possible outcomes of the binary random variables $(Z_t^\star, Z_t)$, under the good event G.

Case 1: $(Z_t^\star, Z_t) = (1, 1)$**.** In this case, both our policy and the oracle baseline observe the true label and $\xi_t = \xi_t^\star = 0$, i.e. neither method makes an error.

Case 2: $(Z_t^\star, Z_t) = (1, 0)$**.** Under the good event G, by Lemma 8 this cannot occur.

Case 3: $(Z_t^\star, Z_t) = (0, 1)$**.** When, $Z_t^\star = 0$ and $Z_t = 1$, our policy tests and observes the true label while the optimal baseline predicts $\hat{Y}_t^\star$, in which case $0 = \xi_t \leq \xi_t^\star$ a.s.

Case 4: $(Z_t^\star, Z_t) = (0, 0)$**.** When, $Z_t^\star = 0$ and $Z_t = 0$, from Lemma 9 it holds that $\hat{Y}_t = \hat{Y}_t^\star$ a.s., and so $\xi_t = \xi_t^\star$ a.s.

Combining these 4 cases together, we have shown that $\xi_t \leq \xi_t^\star$ a.s. Utilizing this, we have that for any $\gamma > 0$

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t \geq \alpha+\gamma\right) &\leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t \geq \alpha+\gamma \,\middle|\, G\right) + \mathbb{P}(\bar{G}) \\ &\leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^\star \geq \alpha+\gamma \,\middle|\, G\right) + \mathbb{P}(\bar{G})\end{aligned}$$

To bound $\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^\star \geq \alpha+\gamma \,\middle|\, G\right)$ we will use $\mathbb{P}(X|G) = \mathbb{P}(X \cap G)/\mathbb{P}(G)$. $\mathbb{P}(X \cap G) \leq \mathbb{P}(X)$, and $\mathbb{P}(G) \geq 1/2$. Thus, $\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^\star \geq \alpha+\gamma \,\middle|\, G\right) \leq 2\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^\star \geq \alpha+\gamma\right)$. Now, $\xi_t^\star$ are binary i.i.d. random variables with $\mathbb{E}(\xi_t^\star) \leq \alpha$. Let $\mu_\xi = \mathbb{E}\left[\sum_{t=1}^{T}\xi_t^\star\right]$, then at any time $\tilde{T} \leq T$:

$$\begin{aligned}\mathbb{P}\left(\frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}}\xi_t^\star \geq \alpha+\gamma\right) &\leq \mathbb{P}\left(\frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}}(\xi_t^\star - \mathbb{E}\xi_t^\star) \geq \gamma\right) \\ &\leq \exp(-2\tilde{T}\gamma^2).\end{aligned}$$

By choosing $\gamma = \sqrt{\frac{\log(\frac{4\pi^2\tilde{T}^2}{6\delta})}{2\tilde{T}}}$, we get that

$$2\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(\xi_t^\star - \mathbb{E}\xi_t^\star) \geq \sqrt{\frac{\log(\frac{4\pi^2T^2}{6\delta})}{2T}}, \forall T\right) \leq \delta/2.$$

Here, taking $\alpha' \triangleq \alpha - \sqrt{\frac{\log(\frac{4\pi^2T^2}{6\delta})}{2T}}$ yields the desired result, where we use Lemma 18 to get that $\mathbb{P}(\bar{G}) \leq \delta/2$. This enables us to take:

$$\alpha_t = \alpha - \sqrt{\frac{\log(7t^2/\delta)}{2t}}$$

to satisfy this constraint.

□

12.4 Lemma 14

Lemma 14. *For every round $t > T_0$, conditioned on the good event G, the regret is bounded as:*

$$\mathbb{E}[Z_t - Z_t^\star|G] \leq 2\pi M \arccos(\tau^\star)\left(1 + \frac{1+e}{2m\pi\arccos(\tau^\star)}\right)\left(K_{\max}\zeta_t + 2(K_{\max}+1)\frac{B_t}{\sqrt{\frac{p^\star t\lambda_0}{8}}} + 3\varepsilon_Q\right).$$

Proof of Lemma 14. For $t \leq T_0$ we can bound each term of the regret by one, $\mathbb{E}[Z_t - Z] \leq 1$. For $t > T_0$ this requires analyzing $\mathbb{E}[Z_t - Z]$. For this, we need to essentially lower bound c_t as a function of X_t. We see that

$$\begin{aligned}
\{c_t \leq 0\} &= \{|\langle X_t, \theta_t^L \rangle| - \tau_t \leq 0\} \\
&\subseteq \{\min_{\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta} |\langle X_t, \theta \rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha) + \varepsilon_Q \leq 0\} \\
&\subseteq \{\min_{\theta \in \mathcal{C}_t} |\langle X_t, \theta \rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha) + \varepsilon_Q \leq 0\} \\
&\subseteq \{|\langle X_t, \theta^\star \rangle| - \hat{\tau}(\theta^\star, \hat{P}_t, \alpha) + \varepsilon_Q \leq 0\} \\
&\subseteq \{|\langle X_t, \theta^\star \rangle| - \tau^\star(\theta^\star, P, \alpha) + \varepsilon_Q \leq 0\}
\end{aligned} \tag{44}$$

We relaxed the minimization domain and used Equation (13) and (11).

$$\begin{aligned}
\mathbb{E}R_t &= \mathbb{E}[Z_t - Z | G] \\
&= \mathbb{P}\left(\{c_t \leq 0\} \cap \{|\langle X_t, \theta^\star \rangle| \geq \tau^\star\} | G\right) \\
&\overset{a}{\leq} \mathbb{P}\Big(\tau^\star \leq |\langle X_t, \theta^\star \rangle| \leq \tau(\theta^\star, P, \alpha_t - \zeta_t - 2B_t/\sqrt{\lambda_{\min}^t}) + 2B_t/\sqrt{\lambda_{\min}^t} + 2\varepsilon_Q | G\Big) \\
&\overset{b}{\leq} \mathbb{P}\Big(\tau^\star \leq |\langle X_t, \theta^\star \rangle| \leq \tau(\theta^\star, P, \alpha - 2\zeta_t - 2\frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}}) + 2\frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}} + 2\varepsilon_Q | G\Big) \\
&\overset{c}{\leq} \mathbb{P}\Big(\tau^\star \leq |\langle X_t, \theta^\star \rangle| \leq \tau^\star + 2K_{\max}(\zeta_t + \frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}}) + 2\frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}} + 3\varepsilon_Q | G\Big) \\
&\overset{d}{\leq} 2\pi M \arccos(\tau^\star) \left(1 + \frac{1+e}{2m\pi \arccos(\tau^\star)}\right) \left(2K_{\max}\zeta_t + 2(K_{\max}+1)\frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}} + 3\varepsilon_Q\right).
\end{aligned} \tag{45}$$

a) follows by the upper bounding of the thresholding condition. b) follows by the sub-event $G^{(4)}$ of the good event and the bound of lemma 17, and by using $\sqrt{\log(7t^2/\delta)/2t} \preccurlyeq \zeta_t$. c) follows using Lemma 6, with $\gamma = \zeta_t + 2\frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}}$. d) follows from Lemma 13.

An important technical detail in step (c) is that $\gamma \leq \gamma_{\max}$. As we discussed in the Section 11.4.1 $p^\star > 0 \implies \tau^\star < 1$. As a consequence, it holds that $\gamma_{\max} > 0$. Finally, $\gamma \triangleq \gamma_t = \mathcal{O}(\frac{1}{\sqrt{t}})$, as a result there exists a constant, w.l.o.g. T_0 such that for all $t \geq T_0$ we have that $\gamma_t \leq \gamma_{\max}$.

□

12.5 Proof of Theorem 2

Proof of Theorem 2. Let $A(m, M, \alpha) \triangleq 2\pi M \arccos(\tau^\star)\left(1 + \frac{1+e}{2m\pi \arccos(\tau^\star)}\right)$. By using the lemma 14, and by conditioning on the good event we have that with probability at least $1 - \delta$:

$$\begin{aligned}
Regret(T) &\leq T_0 + \sum_{t=T_0}^{T} \mathbb{E}R_t \\
&= T_0 + 2A(m, M, \alpha)K_{\max} \sum_{t=T_0}^{T} \zeta_t + 2(K_{\max}+1) \sum_{t=T_0}^{T} \frac{B_t}{\sqrt{\frac{p^\star t \lambda_0}{8}}} + 3A(m, M, \alpha) \sum_{t=T_0}^{T} \varepsilon_Q.
\end{aligned}$$

To control $\sum_{t=1}^{T} \varepsilon_Q$ we can either choose ε_Q to be small, e.g. $\varepsilon_Q = \frac{1}{T^2}$. However, that requires the knowledge of the horizon T. In order to surpass this obstacle, we can choose $\{\varepsilon_Q^t\}_{t=1}^{\infty} = \{1/t^2\}_{t=1}^{\infty}$. In that case, $\sum_{t=1}^{T} \varepsilon_Q \leq \sum_{t=1}^{\infty} 1/t^2 = \pi^2/6 = O(1)$.

For $2(K_{\max}+1)\sum_{t=T_0}^{T}\frac{B_t}{\sqrt{\frac{p^\star t\lambda_0}{8}}}$ we have that:

$$2(K_{\max}+1)\sum_{t=T_0}^{T}\frac{B_t}{\sqrt{\frac{p^\star t\lambda_0}{8}}} \leq 2(\frac{(1+e)}{2m\pi\arccos(\tau_{\max})}+1)\sum_{t=1}^{T}\frac{B_t}{\sqrt{\frac{p^\star t\lambda_0}{8}}}$$
$$\preccurlyeq \frac{B_T}{\sqrt{p^\star\lambda_0}}\sum_{t=1}^{T}1/\sqrt{t} = \mathcal{O}(\frac{B_T}{\sqrt{p^\star\lambda_0}}\sqrt{T}).$$

We remind that $\zeta_t \triangleq \sqrt{\frac{d\log(3/\varepsilon_Q)+\log(\frac{\pi^2t^2}{3\delta})}{2t}}$. As a result $\sum_{t=1}^{T}\zeta_t = \tilde{\mathcal{O}}(\sqrt{dT})$. On the other side, as $B_t \triangleq 2\kappa\left(1+\sqrt{\log(1/\delta)+2d\log\left(1+\frac{N_\theta^t}{\kappa\lambda d}\right)}\right) = \mathcal{O}(\kappa\sqrt{\log(1/\delta)+d\log(p^\star T)})$ then $B_T\sum_{t=1}^{T}1\sqrt{t} = \mathcal{O}(\sqrt{\log(1/\delta)T+dT\log(p^\star T)})$, as $\sum_{t=1}^{T}1/\sqrt{t} = \mathcal{O}(\sqrt{T})$.

By putting all together we have that $R_T = \tilde{\mathcal{O}}\big(\sqrt{\frac{(\log(1/\delta)+d)T}{p^\star\lambda_0}}\big)$. □

13 Good event proof

In Lemma 8 we proved that, with high probability, our policy tests whenever the optimal one does, that is $N_\Theta^t \geq N_{OPT}^t$ when $G_\theta, G_{p_{\text{err}}}$ hold. We must collect enough samples so as the confidence set provide tight estimates about the value of $\theta^\star$. Let define the following auxiliary good events.

- $\mathcal{E}_1 = \{\forall t\geq 1 : N_\Theta^t \geq N_{OPT}^t\}$.
- $\mathcal{E}_2 = \{\forall t\geq 1 : N_{OPT}^t \geq N(t,\delta)\}$.

It is true that $G^{(2)} = \{\forall t\geq 1 : N_\theta^t \geq N(t,\delta)\} \supseteq \mathcal{E}_1\cap\mathcal{E}_2$ when $G_\theta, G_{p_{\text{err}}}$ hold, where $N(t,\delta) = p_\star t - \sqrt{\frac{\ln(\pi t^2/3\delta)t}{2}}$.

In Lemma 8 we proved that $\mathbb{P}(\mathcal{E}_1 \mid G_\theta, G_{p_{\text{err}}}) \geq 1-\delta$ due to pessimism. Now, it remains to prove the same for the event $\mathcal{E}_2$. As the number of samples of the optimal policy follows the binomial distribution with parameter $p^\star$ we can use standard concentration inequalities to derive such a bound.

Lemma 15. $\mathbb{P}(\mathcal{E}_2 \mid G_\theta, G_{p_{err}}) \geq 1-\delta$.

Proof. As the contexts arrive in an i.i.d. fashion, then $N_{OPT}^t \sim \text{Binom}(p_\star, t)$. By a Chernoff-Hoeffding bound, for $s>0$

$$\mathbb{P}(|N_{OPT}^t - p_\star t| \geq s) \leq 2\exp(-\frac{2s^2}{t}).$$

By choosing $s \triangleq \sqrt{\frac{\ln(\pi t^2/3\delta)t}{2}}$ we derive

$$\mathbb{P}(|N_{OPT}^t - p_\star t| \geq \sqrt{\frac{\ln(\pi t^2/3\delta)t}{2}}) \leq \delta\frac{6}{\pi}\frac{1}{t^2}.$$

Now, by using the union bound for all $t\geq 1$,

$$\mathbb{P}\Big(\forall t\geq 1 : |N_{OPT}^t - p_\star t| \geq \sqrt{\frac{\ln(\pi t^2/3\delta)t}{2}}\Big) \leq \delta\frac{6}{\pi}\sum_{t=1}^{\infty}\frac{1}{t^2} = \delta.$$

□

δ is a constant, we simply require that $\delta = \Omega(T^2e^{-T})$. Then, for some T_0, we have that for all $t\geq T_0$ with probability at least $1-\delta$;

$$N_{OPT}^t \geq p^\star t/2. \tag{46}$$

Lemma 16.

$$\mathbb{P}(G^{(2)} \mid G_\theta, G_{p_{err}}) \geq 1-2\delta.$$

Proof. By taking the union bound

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \mid G_\theta, G_{p_{\text{err}}}) \geq 1 - 2\delta.$$

By using $G^{(2)} \supseteq \mathcal{E}_1 \cap \mathcal{E}_2$ when $G_\theta, G_{p_{\text{err}}}$ hold we conclude the proof. □

To show that $\mathbb{P}(G^{(3)}) \geq 1 - \delta$ we will use a covering argument to derive a lower bound for the minimum covariance matrix. Then, we will use Lemma 15 as a lower bound on the number of samples collected to construct the empirical covariance matrix. Finally, we will union bound these two events to complete the proof. We note some constants may have changes due to the union bound.

Lemma 17. *Let $\delta \in (0,1)$. Consider a random $d \times d$ dimensional matrix valued process $\{A_t\}_{t=0}^{\infty}$ adapted to a filtration $\mathcal{F}_t = \sigma(A_k \mid k \leq t)$, where each $A_t \in \mathbb{R}^{d\times d}$ is symmetric ($A_t = A_t^\top$), positive semi-definite, satisfies $\|A_t\|_{op} \leq 1$ almost surely and such that there is a constant $\lambda_0 > 0$ satisfying*

$$\mathbb{P}\left(\lambda_{\min}(\mathbb{E}[A_t|\mathcal{F}_{t-1}]) \geq \lambda_0 \ \forall t \in \mathbb{N}\right) \geq 1 - \delta.$$

Let $\lambda_{\min}^t \triangleq \lambda_{\min}\left(\sum_{s=0}^t A_s\right)$. Then, for $\varepsilon > 0$, the following holds:

$$\mathbb{P}\left\{\lambda_{\min}^t \geq t(\lambda_0 - 2\varepsilon) - \sqrt{\frac{t}{2}\left(d\log\left(\frac{2}{\varepsilon}+1\right) + \log\left(\frac{2t^2}{\delta}\right)\right)} \ \forall t \in \mathbb{N}\right\} \geq 1 - \delta.$$

The proof is in Section 13.1.

We will apply this lemma for $A_t = X_t X_t^\top$. It is true that $\left\|X_t X_t^\top\right\|_{op} = \|X_t\|_2 = 1$. We will make again the same observation, by choosing the covering parameter as $\varepsilon = \frac{\lambda_0}{5}$, then we have that for all $t \geq T_0$

$$\lambda_{\min}^t \geq N_\theta^t \lambda_0/4. \tag{47}$$

In Lemma 15 we proved that with probability at least $1-\delta$, it holds that $N_\theta^t \geq \frac{p^\star t}{2}$. By taking the union bound over the two events, we have that with probability at least $1 - 2\delta$

$$\lambda_{\min}^t \geq p^\star t \frac{\lambda_0}{8}.$$

Lemma 18.

$$\mathbb{P}(G) = \mathbb{P}(G_\theta \cap G^{(2)} \cap G^{(3)} \cap G_{p_{err}}) \geq 1 - 6\delta.$$

Proof. By using the product rule we have that

$$\mathbb{P}(G_\theta \cap G^{(2)} \cap G_{p_{\text{err}}}) = \mathbb{P}(G^{(2)} \mid G_\theta \cap G^{p_{\text{err}}})\mathbb{P}(G_\theta \cap G_{p_{\text{err}}})$$

As $\mathbb{P}(G_\theta) \geq 1 - \delta$ from Lemma 1 and $\mathbb{P}(G_{p_{\text{err}}}) \geq 1 - \delta$ from Lemma 4, by using the union bound we have $\mathbb{P}(G_\theta \cap G^{(2)}) \geq 1 - 2\delta$. By using also Lemma 16 we have

$$\begin{aligned}\mathbb{P}(G^{(2)} \mid G_\theta \cap G_{p_{\text{err}}})\mathbb{P}(G_\theta \cap G_{p_{\text{err}}}) &\geq (1-2\delta)^2 \\ &\geq 1 - 4\delta.\end{aligned}$$

As $\mathbb{P}(G^{(3)}) \geq 1 - 2\delta$ by Lemma 17, by taking the union bound again we have that

$$\mathbb{P}(G_\theta \cap G^{(2)} \cap G^{(3)} \cap G_{p_{\text{err}}}) \geq 1 - 6\delta.$$

□

13.1 Proof of Lemma 17

Proof of Lemma 17. Let the random variable $Z_t^{\upsilon} \triangleq \upsilon^\top A_t \upsilon - \mathbb{E}[\upsilon^\top A_t \upsilon \mid \mathcal{F}_{t-1}]$, such that $\upsilon \in \mathcal{S}^{d-1}$. Notice that Z_t^{υ} is a martingale difference sequence as;

1.
$$\begin{aligned}\mathbb{E}[|Z_t^{\upsilon}|] &\leq \mathbb{E}[|\upsilon^\top A_t \upsilon|] + \mathbb{E}|\mathbb{E}[\upsilon^\top A_t \upsilon \mid \mathcal{F}_{t-1}]| \\ &\leq \mathbb{E}[\upsilon^\top A_t \upsilon] + \mathbb{E}\mathbb{E}[\upsilon^\top A_t \upsilon \mid \mathcal{F}_{t-1}] \\ &\leq 1 + 1 = 2 < \infty.\end{aligned}$$

2.
$$\mathbb{E}[Z_t^{\upsilon} \mid \mathcal{F}_{t-1}] = \mathbb{E}[\upsilon^\top A_t \upsilon \mid \mathcal{F}_{t-1}] - \mathbb{E}[\upsilon^\top A_t \upsilon \mid \mathcal{F}_{t-1}] = 0.$$

By the Azuma-Hoeffding Inequality [9], as $Z_t^{\upsilon} \in [0, 1]$ a.s., for a fixed $t \in [T]$ we have, $c \geq 0$;

$$\mathbb{P}\left\{\sum_{s=0}^{t}(\upsilon^\top A_s \upsilon - \mathbb{E}[\upsilon^\top A_s \upsilon \mid \mathcal{F}_{s-1}]) \leq -c\right\} \leq \exp\left(-\frac{2c^2}{t}\right).$$

Setting the error probability to δ_t,

$$\mathbb{P}\left\{\sum_{s=0}^{t}(\upsilon^\top A_s \upsilon - \mathbb{E}[\upsilon^\top A_s \upsilon \mid \mathcal{F}_{s-1}]) \leq -\sqrt{\frac{\log(\frac{1}{\delta_t})t}{2}}\right\} \leq \delta_t.$$

Thus, substituting $\delta_t = \frac{\delta}{2t^2}$ and using the union bound we get,

$$\mathbb{P}\left\{\sum_{s=0}^{t}(\upsilon^\top A_s \upsilon - \mathbb{E}[\upsilon^\top A_s \upsilon \mid \mathcal{F}_{s-1}]) \leq -\sqrt{\frac{\log(\frac{2t^2}{\delta})t}{2}} \;\; \forall t \in \mathbb{N}\right\} \leq \sum_{t=1}^{\infty}\delta_t \leq \delta.$$

Let $\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$ an ε-cover of $\mathcal{S}^{d-1}$. By **Corollary 4.2.13** at [53] we have that the covering numbers of $\mathcal{S}^{d-1}$ satisfy for any $\varepsilon > 0$;

$$\left(\frac{1}{\varepsilon}\right)^d \leq \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^d.$$

By taking the union bound over all $\upsilon_i \in \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$ we have

$$\mathbb{P}\left\{\exists \upsilon_i \in \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon) : \sum_{s=0}^{t}(\upsilon_i^\top A_s \upsilon_i - \mathbb{E}[\upsilon_i^\top A_s \upsilon_i \mid \mathcal{F}_{s-1}]) \leq -\sqrt{\frac{[d\log(2/\varepsilon + 1) + \log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \leq \delta \tag{48}$$

Let $\upsilon_t^\star \triangleq \operatorname{argmin}_{\upsilon \in \mathcal{S}^{d-1}} \upsilon^\top \sum_{s=0}^{t} A_s \upsilon$, then there exists an $\upsilon_{i_t} \in \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$ such that $\|\upsilon_{i_t} - \upsilon_t^\star\|_2 \leq \varepsilon$ We are going to bound $|\upsilon_t^{\star\top} \sum_{s=0}^{t} A_s \upsilon_t^\star - \upsilon_{i_t}^\top \sum_{s=0}^{t} A_s \upsilon_{i_t}|$ by a function of ε.

$$\begin{aligned}|\upsilon_t^{\star\top} \sum_{s=0}^{t} A_s \upsilon_t^\star - \upsilon_{i_t}^\top \sum_{s=0}^{t} A_s \upsilon_{i_t}| &= |\upsilon_t^{\star\top} \sum_{s=0}^{t} A_s \upsilon_t^\star - \upsilon_t^{\star\top} \sum_{s=0}^{t} A_s \upsilon_{i_t} + \upsilon_t^{\star\top} \sum_{s=0}^{t} A_s \upsilon_{i_t} - \upsilon_{i_t}^\top \sum_{s=0}^{t} A_s \upsilon_{i_t}| \\ &= |\upsilon_t^{\star\top} \sum_{s=0}^{t} A_s (\upsilon_t^\star - \upsilon_{i_t}) + (\upsilon_t^\star - \upsilon_{i_t})^\top \sum_{s=0}^{t} A_s \upsilon_{i_t}| \\ &= |(\upsilon_t^\star - \upsilon_{i_t})^\top \sum_{s=0}^{t} A_s (\upsilon_{i_t} + \upsilon_t^\star)| \\ &\leq \|\upsilon_t^\star - \upsilon_{i_t}\|_2 \left\|\sum_{s=0}^{t} A_s (\upsilon_{i_t} + \upsilon_t^\star)\right\|_2 \\ &\leq \varepsilon \sum_{s=0}^{t} \|A_s\|_{op} (\|\upsilon_{i_t}\|_2 + \|\upsilon_t^\star\|_2) \\ &= 2t\varepsilon.\end{aligned} \tag{49}$$

Using inequality 48 we have

$$\mathbb{P}\left\{\sum_{s=0}^{t} v_{i_t}^\top A_s v_{i_t} \geq \sum_{s=0}^{t} \mathbb{E}[v_{i_t}^\top A_s v_{i_t} \mid \mathcal{F}_{s-1}] - \sqrt{\frac{[d\log(2/\varepsilon+1)+\log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \geq 1-\delta$$

where i_t is a point in the cover $\mathcal{N}(\mathcal{S}^{d-1}, \varepsilon)$ such that $\|v_{i_t} - v_t^\star\|_2 \leq \varepsilon$. Equation 49 can be used to relate $\sum_{s=0}^{t} v_{i_t}^\top A_s v_{i_t}$ and $\lambda_{\min}^t$,

$$\mathbb{P}\left\{\underbrace{\sum_{s=0}^{t} v_t^{\star\top} A_s v_t^\star}_{\lambda_{\min}^t} + 2t\varepsilon \geq \sum_{s=0}^{t} \mathbb{E}[v_{i_t}^\top A_s v_{i_t} \mid \mathcal{F}_{s-1}] - \sqrt{\frac{[d\log(2/\varepsilon+1)+\log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \geq 1-\delta.$$

Using the fact that $\mathbb{E}[v_{i_t}^\top A_s v_{i_t} \mid \mathcal{F}_{s-1}] \geq \lambda_{\min}(\mathbb{E}[A_s \mid \mathcal{F}_{s-1}])$ we conclude that,

$$\mathbb{P}\left\{\lambda_{\min}^t + 2t\varepsilon \geq \sum_{s=0}^{t} \lambda_{\min}(\mathbb{E}[A_s \mid \mathcal{F}_{s-1}]) - \sqrt{\frac{[d\log(2/\varepsilon+1)+\log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \geq 1-\delta.$$

Finally, the assumption that $\mathbb{P}\left(\lambda_{\min}(\mathbb{E}[A_t|\mathcal{F}_{t-1}]) \geq \lambda_0 \; \forall t \in \mathbb{N}\right) \geq 1-\delta$ and a union bound allows us to conclude that,

$$\begin{aligned}
&\mathbb{P}\left\{\lambda_{\min}^t \geq t(\lambda_0 - 2\varepsilon) - \sqrt{\frac{[d\log(2/\varepsilon+1)+\log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \\
&\geq \mathbb{P}\left\{\lambda_{\min}^t + 2t\varepsilon \geq \sum_{s=0}^{t} \lambda_{\min}(\mathbb{E}[A_s \mid \mathcal{F}_{s-1}]) - \sqrt{\frac{[d\log(2/\varepsilon+1)+\log(\frac{2t^2}{\delta})]t}{2}} \cap \lambda_{\min}(\mathbb{E}[A_t|\mathcal{F}_{t-1}]) \geq \lambda_0 \;\; \forall t \in \mathbb{N}\right\} \\
&\geq 1-\delta'.
\end{aligned}$$

This finalizes the result for $\delta' = 2\delta$.

□