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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of sequentially testing individuals for a binary disease outcome whose true risk is governed by an unknown logistic regression 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 prior observations. 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 high probability. 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 never exceeds 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 [46, 9], recommendation systems [2, 49], online ad-allocation [43], and portfolio selection [37]. Bandit algorithms [30] and reinforcement learning [44] 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, determining which problem parameters are vital for decision-making; the second is the decision-making challenge, where effective performance is required concurrently with learning.

Although machine learning systems perform exceptionally well in practice, sometimes even surpassing human performance, when applied in human-centric scenarios safety constraints are paramount [21, 19]. 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 [34, 48, 18]. 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 [28]. In sequential decision making problems learning while satisfying a safety criterion typically makes reward

Submitted to the First Exploration in AI Today Workshop at ICML (EXAIT at ICML 2025). Do not distribute.

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 rule of when to test.

More generally, the objective is to produce accurate guesses of the disease status of an incoming stream of patients while minimizing the number of tests required. This problem belongs to the rich tradition of the active learning and selective sampling literature [41, 23, 33, 6, 17, 42]. These study settings where context information may be abundant but the labels are hard to come by [13]. 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.

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 [33, 40], 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. However, in real-world scenarios like the one in [5], 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 $\alpha \in [0, 1]$. Since testing every patient is expensive, the goal is to minimize the number of tests subject to a misclassification error bound. These objectives can be formalized as:

Is it possible to design a classifier that minimizes the expected number of tests while maintaining a misclassification error below a specified safety threshold?

Contributions: In this work, we provide a logistic bandit algorithm to tackle the aforementioned problem with a regret guarantee of $\mathcal{O}(\sqrt{dT \log(T/\delta)})$ [1] For a more detailed analysis about the constants in Theorem 1 we refer the reader to the appendix. We validate our theoretical results through comprehensive experiments.

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 $A \in \mathbb{R}^{n \times n}$ The minimum eigenvalue of a matrix $A \in \mathbb{R}^{n \times n}$ will be denoted as $\lambda_{\min}(A)$. The set $\{1, 2, \ldots, n\}$ is denoted as $[n]$. $\mu(z) = \frac{1}{1+\exp(z)}$ is the logistic function. $\mathbb{1}$ denotes the indicator function of an

[1] In Theorem 1 we show that the regret is upper bounded by $\tilde{\mathcal{O}}(\sqrt{\frac{dT}{\lambda_0}})$ where λ_0 is the minimum eigenvalue of the covariance matrix of the optimal policy. In case of the uniform distribution over the unit sphere, this quantity is equal to ${}^1/\sqrt{d}$. It is known that in Linear Bandits the dependency on the dimension is linear, so we do not miss any $\sqrt{d}$ factor.

event. For two functions f, g we say that $f(x) \preccurlyeq g(x)$ when there exists a constant $c > 0$ such that $f(x) \leq cg(x)$.

2.1 Problem Definition

We consider the following repeated game scenario between the learner and the environment. At every round $t \in [T]$, the environment generates a context $X_t \in \mathbb{R}^d$ such that $\|X_t\|_2 \leq 1$. 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, $\theta^\star \in \mathbb{R}^d$ 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. $\mathcal{H}_t = \{X_1, Z_1Y_1, X_2, Z_2Y_2, \ldots, X_t\}$ and possibly on internal randomization of the learner. Similarly, $\hat{Y}_t$ must be $\mathcal{F}_t = \sigma\{X_1, Z_1Y_1, X_2, Z_2Y_2, \cdots, X_t, Z_tY_t\}$ 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 high probability. This can be summarized as the safe learning objective below:

$$\min_{\{\hat{Y}_t\}, \{Z_t\}} \sum_{t=1}^{T} \mathbb{E} Z_t \quad \text{s.t.} \quad \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\} \leq \alpha\right) \geq 1 - \delta. \tag{1}$$

2.2 Optimal baseline

First, let us consider the case where the feature distribution P and optimal discriminator $\theta^\star$ are known a priori to the learner. In this case, we can easily devise a threshold decision rule τ that is a function of P, $\theta^\star$ and α. More analytically,

$$Z_t = \mathbb{1}\{|X_t^\top \theta^\star| \leq \tau\} \qquad \hat{Y}_t = \begin{cases} 0 & \text{if } X_t^\top \theta^\star < -\tau, \\ Y_t & \text{if } |X_t^\top \theta^\star| \leq \tau, \\ 1 & \text{if } X_t^\top \theta^\star > \tau. \end{cases} \tag{2}$$

A threshold policy proves to be optimal when the constraint is only required to be met in expectation, as encapsulated in the following proposition.

Proposition 1. *Consider the following variation of our problem where the constraint holds in expectation, and both the batch 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 the appendix.

To analyze the performance of a selected threshold τ, we define the function p_{err} as the probability of misclassification incurred by the the threshold τ, if the true underlying $\theta^\star$ was θ, where the expectation is taken with respect to P:

$$p_{\text{err}}(\theta, P, \tau) = \int (1 + \exp(|x^\top \theta|))^{-1} \mathbb{1}\left\{|x^\top \theta| > \tau\right\} P(dx). \tag{3}$$

The term $(1 + \exp(|x^\top \theta|))^{-1} = \min\left\{\frac{1}{1+\exp(x^\top \theta)}, 1 - \frac{1}{1+\exp(x^\top \theta)}\right\}$ is the optimal misclassification error for fixed x, θ. The term $\mathbb{1}\left\{|x^\top \theta| > \tau\right\}$ takes the value of one only if we use a threshold rule and we predict the label $\hat{y}$ without observing the real label y of the context x.

From this, we can naturally define the optimal $\tau^\star$ for a given α as the minimum value of τ (thus minimizing the number of rejections) that satisfies the α-fraction misclassification constraint. Observe that both p_{err} and $\tau^\star$ can also be evaluated with respect to observed empirical distributions $\hat{P}$, not just the true distribution P. This gives an oracle baseline where any algorithm requires an expected number of tests $p^\star T$, where

$$\tau^\star(\theta, P, \alpha) = \min\{\tau : p_{\text{err}}(\theta, P, \tau) \leq \alpha\} \tag{4}$$

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

This lets us define the "safe regret" of an algorithm as the number of excess tests it takes over this oracle baseline. 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 Equation (6)?

$$\mathbb{E}\left[\sum_{t=1}^{T} Z_t - p^\star\right] \quad \text{s.t.} \quad \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{Y}_t \neq Y_t\} \leq \alpha\right) \geq 1 - \delta. \tag{6}$$

To analyze this quantity, we make a few natural assumptions.

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

In previous related works, [33], [40], $p^\star$ is a quantity analogous to T_ε that describes the number of times the Bayes optimal classifier outputs a label with confidence less than a fixed parameter $\varepsilon > 0$. This serves as a measure to quantify the inherent difficulty of the problem instance.

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

Assumption 2. *We assume that 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* $\lambda_0 > 0$*:*

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

Past adaptive sampling works [40, 23], and those tackling learning halfspaces, commonly assume the Tsybakov noise condition [45, 11]. The Tsybakov noise condition with parameters (α, A) states that for any $0 < t \leq 1/2$, where $\eta(x) = \mathbb{P}(Y(x) = 1)$, that $\mathbb{P}_{x\sim P}[\eta(x) \geq 1/2 - t] \leq A t^{\frac{\alpha}{1-\alpha}}$. 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 [10]). Our assumption is much less stringent, and encompasses standard distributions such as smooth densities of the form $f(x) = g(\|x\|)$, or a truncated Gaussian.

2.3 Logistic Bandits tools

For our algorithm, we leverage existing methods to provide confidence intervals for $\theta^\star$. [14] 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 will choose the first method to keep both the analysis and the algorithm simple. Moreover, we can compute 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 [31], but the results of [14] are sufficient for our needs. Before stating our algorithm, we introduce some necessary notation from [14]. Since in our work we only collect a paired (X_t, Y_t) sample if we test in a given round, we denote the samples collected by the algorithm prior to round t as $\mathcal{S}_\theta^t$, where $|\mathcal{S}_\theta^t| = N_\theta^t$.

We define the regularized log-likelihood as $\mathcal{L}_t^\lambda(\theta)$, the maximum (regularized) likelihood estimator as $\hat{\theta}_t$, the design matrix as V_t, and the objective $g_t(\theta)$. The projection of $\hat{\theta}_t$ to the parameter space is defined as θ_t^L in Equation (7). The confidence ellipsoid for $\theta^\star$ is $\mathcal{C}_t$ in Equation (8) (implicitly a function of δ), which we use solely for the theoretical analysis of our algorithm.

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

$$\hat{\theta}_t = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \mathcal{L}_t^\lambda(\theta)$$

$$V_t = \sum_{s \in \mathcal{S}_\theta^t} X_s X_s^\top + \kappa \lambda \mathbf{I}_d$$

$$g_t(\theta) = \sum_{s \in \mathcal{S}_\theta^t} \mu(\langle X_s, \theta \rangle) X_s + \lambda \theta$$

$$\theta_t^L \triangleq \operatorname*{argmin}_{\theta \in \Theta} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{V_t^{-1}} \tag{7}$$

$$\mathcal{C}_t \triangleq \left\{ \theta \in \Theta, \left\| \theta - \theta_t^L \right\|_{V_t} \leq B_t \right\}, \tag{8}$$

$$B_t \triangleq 2\kappa \left(\sqrt{\lambda} + \sqrt{\log(1/\delta) + 2d \log\left(1 + \frac{N_\theta^t}{\kappa \lambda d}\right)} \right) \tag{9}$$

Any choice of the regularizer $\lambda = \Theta(1)$ yields the same results up to constants, so for simplicity we choose $\lambda = 1$ for our analysis. This form of confidence interval was studied by [14], which provides anytime, high probability guarantees:

Lemma 1. *[Lemma 12 of [14].] For any fixed choice of λ, δ, the confidence intervals defined in Equation* (8) *are valid:*

$$\mathbb{P}\left(\forall t \geq 1, \theta^\star \in \mathcal{C}_t\right) \geq 1 - \delta.$$

3 Algorithm design

With 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 tests the patient if the inner product between their context and the estimated $\theta^\star$ is too close to 0, based on an estimation 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 τ), and data from even samples where a test was performed for $\theta^\star$ estimation.

The testing condition $Z_t = \mathbb{1}\{c_t \leq 0\}$ can be computed as follows: we defer the derivation and details to Appendix C.1.

$$c_t^\star \triangleq |\langle X_t, \theta^\star \rangle| - \tau^\star(\theta^\star, P, \alpha), \tag{10}$$

$$c_t \triangleq |\langle X_t, \theta_t^L \rangle| - \tau^\star(\theta_t^L, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) - 2B_t \|X_t\|_{V_t^{-1}} - \varepsilon_Q. \tag{11}$$

This can be compared to the optimal rule $Z_t^\star = \mathbb{1}\{c_t^\star \leq 0\}$, where we see that the two matches 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 some confidence buffers. ζ_t arises from confidence intervals on our estimates, i.e. that we only have $\hat{P}_t$ and not P, and $B_t \|X_t\|_{V_t^{-1}}$ arises from the fact that we only have the estimate θ_t^L and not $\theta^\star$. ε_Q is a quantization parameter that shows up in our analysis, and should be thought of as some small quantity like $1/T^2$.

$$\zeta_t \triangleq \sqrt{\frac{d \log(3/\varepsilon_Q) + \log(\frac{\pi^2 t^2}{3\delta})}{2t}}. \tag{12}$$

Algorithm 1 SCOUT algorithm

Input: Number of rounds T, target error rate α, confidence level δ
Initialize: $\mathcal{S}_P = \emptyset$, $\mathcal{S}_\theta = \emptyset$. Maintain $N_P = |\mathcal{S}_P|$, $N_\theta = |\mathcal{S}_\theta|$
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 using (7) and c_t as in equation 10
        Set $Z_t = \mathbb{1}\{c_t \leq 0\}$
    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 $\mathcal{S}_\theta = \mathcal{S}_\theta \cup \{(X_t, Y_t)\}$
    end if
    if t is odd **then**
        Set $\mathcal{S}_P = \mathcal{S}_P \cup \{X_t\}$
    end if
end for

4 Regret Analysis

To derive a regret bound, we begin by analyzing the regret at an arbitrary round $t > T_0$, where T_0 is a constant. For more details we refer the reader to Appendix D.

Lemma 2. *For every round $t > T_0$, conditioned on the good event G, the regret is bounded as:*

$$\mathbb{E}[Z_t - Z_t^\star | G_t] \leq 2\pi M \arccos(\tau^\star(\alpha)) \left(1 + \frac{1+e}{2m\pi \arccos(\tau^\star(\alpha))}\right) \left(\lambda^\star \left(2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_0}}\right) + 2\varepsilon_Q\right),$$

where $\lambda^\star(\gamma) \leq (\gamma + \zeta_t)\frac{1+e}{2m\pi \arccos(\tau^\star(\alpha))}$

Theorem 1. *With probability at least $1-\delta$, the algorithm 1 does not exceed error rate α, and has expected number of excess tests $\tilde{\mathcal{O}}\left(\sqrt{\frac{dT}{\lambda_0}}\right)$.*

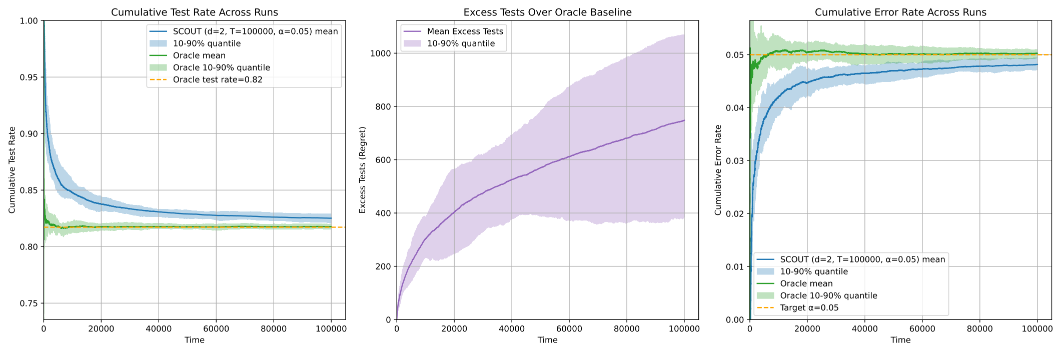
Note that δ can even scale exponentially in T and the algorithm will still have sublinear regret. We need to notice that in linear bandits literature, the dependency in the dimension is $\mathcal{O}(d)$. In our analysis, this extra $\mathcal{O}(\sqrt{d})$ is hidden inside the $\frac{1}{\sqrt{\lambda_0}}$ term where in the case that the distribution of the contexts is the uniform over the unit sphere then this term is equal to $\frac{1}{\sqrt{d}}$.

For a more detailed discussion about future directions we refer the reader to Appendix I.

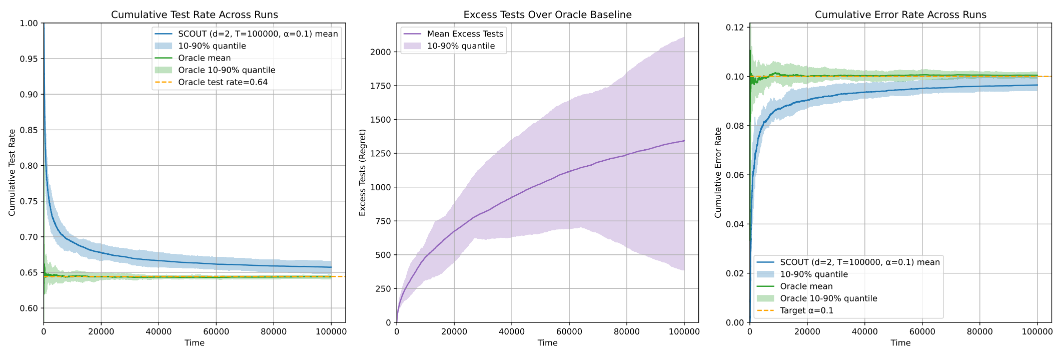
5 Numerical results

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.

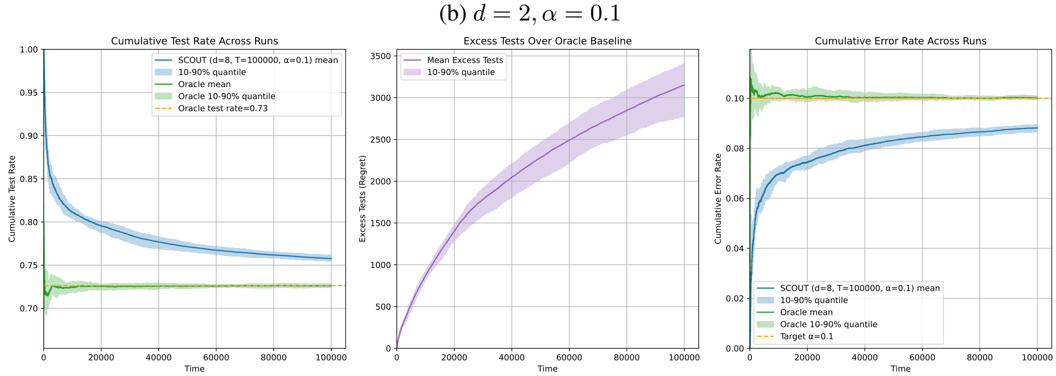
For our simulations we made some slight modifications with respect to the written algorithm. Chiefly, we do not recompute θ_t^L in every iteration, but rather cache its computation, and that of the $\hat{\tau}$, so that at each iteration we simply compute whether the inner product of the context with our estimated θ is above or below a stored threshold.



(a) $d = 2$ simulation, $\alpha = 0.05$.



(b) $d = 2, \alpha = 0.1$



(c) $d = 8, \alpha = 0.1$

Figure 1: Simulation results

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

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

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

[21] 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.

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

[23] 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.

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

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

[26] 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.

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

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

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

[39] 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.

[40] 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.

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

[42] 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.

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

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

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

[46] 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.

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

[48] 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.

[49] 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.

A Baseline policy

A.1 Proof of Proposition 1

Proof. When the value of the parameter θ_* 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 then are following the Bernoulli distribution with parameters p_t, i.e. $Y_t \sim Ber(p_t)$.

To compute the expected error, that is $\mathbb{E}(E_t) = \mathbb{E}(\mathbb{1}\{\hat{Y}_t \neq Y_t\})$, we need to examine the case where we do not test. Otherwise, when we test, we observe the real label and we occur 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 compute the prediction with the least 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.

$$\min_{\{Z_t\}} \sum_{t=1}^{T} 1 - \eta_t \quad \text{s.t.} \quad \frac{1}{T} \sum_{t=1}^{T} \min\{1 - p_t, p_t\}\eta_t \leq \alpha, \quad 0 \leq \eta_t \leq 1. \tag{13}$$

Or equivalently can be written as.

$$\max_{\{Z_t\}} \sum_{t=1}^{T} \eta_t \quad \text{s.t.} \quad \frac{1}{T} \sum_{t=1}^{T} \min\{1 - p_t, p_t\}\eta_t \leq \alpha, \quad 0 \leq \eta_t \leq 1. \tag{14}$$

The solution of this Linear Program is the solution of the *Fractional Knapsack* problem with budget α. In order to solve optimally this problem, we must apply 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.

□

A.2 Discussion of Assumption 3

Assumption 3 requires that the covariance matrix of contexts selected by the optimal policy is 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.

Proof. We have assumed that all contexts have bounded $\mathcal{L}_2$ norm, $\|\mathbf{x}\|_2 \leq B$. Let $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^d \text{ s.t. } \|\mathbf{x}\|_2 \leq 1\}$.

Lemma 3. *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

$$\begin{aligned} \mathbf{v}^\top \Sigma \mathbf{v} &= \mathbb{E}_{\mathbf{x}\sim P} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} && (15) \\ &= \mathbb{E}_{\mathbf{x}\sim P} (\mathbf{x}^\top \mathbf{v})^2, && (16) \end{aligned}$$

and analogously $\mathbf{v}^\top \Sigma_t r \mathbf{v}$ as

$$\begin{aligned} \mathbf{v}^\top \Sigma_t r \mathbf{v} &= \int_{\mathcal{B}} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} m d\mathbf{x} && (17) \\ &= m \int_{\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 d\mathbf{x}. && (18) \end{aligned}$$

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

$$(\mathbf{x}^\top \mathbf{v})^2 p(\mathbf{x}) \geq (\mathbf{x}^\top \mathbf{v})^2 m \tag{19}$$

that implies by integrating all over the domain that

$$\begin{aligned} \implies \int_{x\in\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 p(\mathbf{x}) d\mathbf{x} &\geq \int_{x\in\mathcal{B}} (\mathbf{x}^\top \mathbf{v})^2 m d\mathbf{x} && (20) \\ \mathbf{v}^\top \Sigma \mathbf{v} &\geq \mathbf{v}^\top \Sigma_{tr} \mathbf{v} && (21) \end{aligned}$$

□

The previous lemma applies for any arbitrary vector $\mathbf{v}$, so $\Sigma \succeq \Sigma_{tr}$. Let $(\lambda_{\min}, \mathbf{v}_{\min})$ the eigen-pair of the corresponding minimum eigenvalue of Σ. Let us apply the previous lemma for $\mathbf{v}_{\min}$. Then, we derive that

$$\lambda_{\min} \|\mathbf{v}_{\min}\|_2^2 \geq m \int_{\mathcal{B}} (\mathbf{x}^\top \mathbf{v}_{\min})^2 d\mathbf{x} \tag{22}$$

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. My multiplying and dividing on the right hand side of the previous inequality with $V_d(1)$ we derive that

$$\begin{aligned} \lambda_{\min} \|\mathbf{v}_{\min}\|_2^2 &\geq m V_d(1) \int_{\mathcal{B}} (\mathbf{x}^\top \mathbf{v}_{\min})^2 \frac{1}{V_d(1)} d\mathbf{x} && (23) \\ &= m V_d(1) \int_{\|\mathbf{x}\|_2^2 \leq 1} (\mathbf{x}^\top \mathbf{v}_{\min})^2 \frac{1}{V_d(1)} d\mathbf{x} && (24) \end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[\langle \mathbf{x}, \mathbf{v}_{\min}\rangle^2] &= \mathbb{E}[\mathbf{v}_{\min}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v}_{\min}] \\ &= \mathbf{v}_{\min}^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{v}_{\min}\end{aligned}$$

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.

To see why, consider the $\mathbb{E}[\mathbf{x}_i\mathbf{x}_j]$ for $i \neq j$.

$$\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] = \frac{1}{V_d(1)}\int_{x_1=-1}^{x_1=1}\int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}}\cdots\int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} x_i x_j dx_d\cdots dx_2 dx_1. \tag{25}$$

By a change of variable $\mathbf{x}_i \mapsto -\mathbf{x}_i$:

$$\begin{aligned}\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] &= -\frac{1}{V_d(1)}\int_{x_1=-1}^{x_1=1}\int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}}\cdots\int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}(-x_i)x_j dx_d\cdots d(-x_i)\cdots dx_2dx_1 \quad (26)\\ &= -\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] \quad (27)\end{aligned}$$

As a result we get $\mathbb{E}[\mathbf{x}_i\mathbf{x}_j] = 0$ for $i \neq j$.

To compute the diagonal entries:

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

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

By combining them all we derive

$$\lambda_{\min}\|\mathbf{v}_{\min}\|_2^2 \geq \frac{mV_d(1)S_d(1)}{d(d+2)V_d(1)}\|\mathbf{v}_{\min}\|_2^2 \tag{28}$$

$$\lambda_{\min} \geq \frac{mS_d(1)}{d(d+2)} > 0. \tag{29}$$

□

B Stability of error estimates

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

B.1 Stability with respect to context sampling $\hat{P}_t$

Analyzing Equation (3), we note that we do not know the true distribution P, but only have access to samples from it. For any fixed θ, τ, (3) becomes a sum of i.i.d. $[0, 1/2]$ bounded random variables, enabling us to use standard concentration bounds.

Lemma 4. *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 $\hat{P}_t$:*

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

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. Thus, we utilize an ϵ-net for both $\tau \in [0, 1]$ and $\theta \in \Theta$.

B.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, at every round t, 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 $|\mathcal{Q}_\theta| = O\left(\varepsilon_Q^{-(d-1)}\right)$.

To this end, we define the quantized optimized τ as:

$$\tau_Q^\star(\theta, \hat{P}, \alpha) = \min\{\tau_Q \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, \hat{P}, \tau_Q) \leq \alpha\} \tag{30}$$

$$\tau^\star(\theta, \hat{P}, \alpha) \leq \tau_Q^\star(\theta, \hat{P}, \alpha) \leq \tau^\star(\theta, \hat{P}, \alpha) + \varepsilon_Q \tag{31}$$

as p_{err} is monotonic in τ.

B.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 5. *For all $\theta, \theta' \in \Theta$, $\tau > \|\theta - \theta'\|\|_{V_t} \|x\|_{V_t^{-1}}$, and distribution $\rho(x)$ on $\mathcal{X}$:*

$$p_{err}(\theta, \rho, \tau) - p_{err}(\theta', \rho, \tau - \|\theta - \theta'\|\|_{V_t} \|x\|_{V_t^{-1}}) \leq \|\theta - \theta'\|_{V_t} \|x\|_{V_t^{-1}}.$$

This indicates that as our estimation of θ improves, so will our error probability estimates. To this end, we define the good event $G_{p_{\text{err}}}$ as:

$$G_{p_{err}} = \left\{ \left| p_{\text{err}}(\theta_Q, \hat{P}_t, \tau_Q) - p_{\text{err}}(\theta_Q, P, \tau_Q) \right| \leq \zeta_t \ : \ \forall t \in [T], \forall \theta_Q \in \mathcal{Q}_\theta, \forall \tau_Q \in \mathcal{Q}_\tau \right\}. \tag{32}$$

The following lemma shows that this good event G happens with overwhelming probability.

Lemma 6. *The good event $G_{p_{err}}$ defined in* (32) *holds with high probability:*

$$\mathbb{P}(G_{p_{err}}) \geq 1 - \delta \tag{33}$$

Conditioning on the good event $G_{p_{err}}$, $\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha)$ is close to $\tau^\star(\theta^\star, P)$ when θ_Q is close to $\theta^\star$.

$$\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) \leq \tau^\star\left(\theta^\star, P, \alpha - \zeta_t - \varepsilon_Q - 2B_t \|X_t\|_{V_t^{-1}}\right) + 2B_t \|X_t\|_{V_t^{-1}} + 2\varepsilon_Q, \tag{34}$$

$$\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) \geq \tau^\star(\theta^\star, P, \alpha + \zeta_t + 2B_t \|X_t\|_{V_t^{-1}}). \tag{35}$$

Thus, we can construct an estimator $\hat{\tau}$ as below, which satisfies for all $t \geq 1$:

$$\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) = \tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) + \zeta_t$$

$$\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) \leq \tau^\star\left(\theta^\star, P, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}\right) + 2B_t \|X_t\|_{V_t^{-1}} + \zeta_t + 2\varepsilon_Q. \tag{36}$$

$$\hat{\tau}(\theta_Q, \hat{P}_t, \alpha) \geq \tau^\star(\theta^\star, P, \alpha + \zeta_t + 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t. \tag{37}$$

B.3 Smoothness of $\tau^\star$ with respect to α

As we have seen, we must be able to control τ not just for one α, but for similar α. We show that for small γ, $\tau^\star(\theta^\star, P, \alpha-\gamma)$ is not too much larger than $\tau^\star(\theta^\star, P, \alpha)$. Note that p_{err} is not continuous with respect to α when evaluated at $\hat{P}$, due to the indicator function. However, utilizing Assumption 2, the 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 7 (Stability of $\tau^\star$ with respect to α)**.** *Under Assumptions 1 and 2,*

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

for all θ and $\alpha > \gamma$, where $\lambda^\star(\gamma) = \frac{\gamma(1+\exp(\tau))}{2m\pi \arccos(\tau)} \leq \gamma \frac{1+e}{2m\pi \arccos(\tau^\star(\alpha))}$.

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

C From Stability Analysis to Algorithmic Rules

C.1 Computing Z_t

We design our testing rule based on two main principles. First, our testing rule must be "pessimistic", in that when the baseline police tests, our policy does the same, even for the worst possible $\theta^\star$. Second, our testing rule must be computationally efficient. Recall that the oracle policy is

$$Z_t^\star = \mathbb{1}\{|\langle X_t, \theta^\star\rangle| \leq \tau(\theta^\star, P, \alpha)\}. \tag{39}$$

Our testing rule Z_t must adapt to the data collected, that is $\mathcal{C}_t$ and $\hat{P}$. On the good event G when our estimates are accurate —an event that occurs with high probability— we want to design a policy such that $Z_t \geq Z_t^\star$. To prove so, we will define a dummy testing rule $\tilde{Z}_t$ which considers the worst possible θ in the confidence set $\mathcal{C}_t$, up to the stability analysis terms. We can then show that $Z_t \geq \tilde{Z}_t \geq Z_t^\star$.

Lemma 8. *Let*

$$\tilde{Z}_t = \min_{\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta} \mathbb{1}\{|\langle X, \theta\rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t - \varepsilon_Q \leq 0\}.$$

Then, when G holds, $Z_t^\star = 1 \implies \tilde{Z}_t = 1$, i.e. $\tilde{Z}_t \geq Z_t^\star$ a.s.

Another property of our testing rule is that it makes no additional errors beyond the baseline policy, on the good event. Concretely, 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 and $Y_t^\star$ the one of the oracle baseline policy. On the good event G, when $Z_t = 0$ (which implies that $Z_t^\star = 0$) then $\hat{Y}_t = Y_t^\star$.*

Now we have achieved the first desiderata of our testing rule (pessimism), but are left with a computationally intensive procedure. Naively, computing $\tilde{Z}_t$ is expensive, as even when we relax the optimization domain $\mathcal{C}_t \cap \mathcal{Q}_t$ to only the convex confidence set $\mathcal{C}_t$ we still need to compute the minimization:

$$\min_{\theta \in \mathcal{C}_t} |\langle X, \theta\rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t + \varepsilon_Q. \tag{40}$$

We simplify this in two steps. First, observe that the threshold $\hat{\tau}(\theta, \hat{P}_t, \alpha)$ is not concave in θ, and so maximizing it is highly nontrivial. However, we do not need to precisely compute it: we can simply upper bound $\hat{\tau}(\theta, \hat{P}_t, \alpha)$ for all θ, to yield a more conservative testing condition (testing more often), retaining correctness guarantees and enabling a computationally efficient implementation at the cost of some excess testing. Thus, for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$, $\theta' \in \mathcal{C}_t$, we have:

$$\begin{aligned}
&\hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) \\
&= \tau_Q^\star(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t \\
&\leq \tau^\star(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}} - \varepsilon_Q) + \varepsilon_Q + \zeta_t \\
&\leq \tau^\star(\theta', \hat{P}_t, \alpha - \zeta_t - \varepsilon_Q - 4B_t \|X_t\|_{V_t^{-1}}) + 2B_t \|X_t\|_{V_t^{-1}} + \zeta_t + \varepsilon_Q
\end{aligned} \tag{41}$$

Now, the evaluation of $\tau^\star$ is constant with respect to θ (only depending on θ', e.g. the MLE). Then, the minimization is of $|\langle X, \theta\rangle|$ (a convex function) over a convex set $\mathcal{C}_t$. However, we can simplify this even further, by noting that

$$\left|\min_{\theta\in\mathcal{C}_t} |\langle X, \theta\rangle| - |\langle X, \theta'\rangle|\right| \leq \max_{\theta\in\mathcal{C}_t} |\langle X, \theta' - \theta\rangle| \leq 2B_t \left\|X_t\right\|_{V_t^{-1}} \tag{42}$$

Since $\left\|X_t\right\|_{V_t^{-1}}$ is decaying in t, as given by [31, 1], this allows us to simply utilize $|\langle X, \theta'\rangle|$ as our statistic to threshold instead of the minimization problem described at Equation (40).

To summarize, we have relaxed the testing condition, allowing for efficient computation at the expense of some additional tests. However, as we show in our regret analysis, this is very few additional tests, as we learn $\theta^\star, P$, and $\tau^\star$ sufficiently fast. As we lower bounded the Equation (40), under the good event G it holds that $Z_t = \mathbb{1}\{c_t \leq 0\} \geq \tilde{Z}_t$ and since $\tilde{Z}_t \geq Z_t^\star$, we see that $Z_t \geq Z_t^\star$.

D The good event

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

Our first goal is to prove that the confidence intervals $\mathcal{C}_t$ are valid, i.e., $\theta^\star \in \mathcal{C}_t$ for all t and prove that we have collected enough samples to form them. Although we cannot determine the exact distribution of context, label pair samples to estimate $\theta^\star$, we can demonstrate that our policy is pessimistic and triggers testing whenever the optimal policy would do so. We remind that by the assumption 1 the probability that the optimal policy conducts testing 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 $N_P^t = |\mathcal{S}_P^t|$ for the contexts respectively. The good event comprises the following constituent events.

Definition 1. *At round t the good event G_t holds that*

1. $G_t^{(1)}$*: The confidence sets $\mathcal{C}_t$ are valid, i.e. $\theta^\star \in \mathcal{C}_t$ for all t.*
2. $G_t^{(2)}$*: The estimates $\hat{\tau}(\theta, \hat{P}_t, \alpha)$ are valid, i.e. $|\tau^\star(\theta, P, \alpha) - \tau^\star(\theta, \hat{P}_t, \alpha)| \leq \zeta_t$ for all $\theta \in \Theta, \tau^\star \in \mathcal{Q}_\tau, t \geq 1$.*
3. $G_t^{(3)}$*: the confidence sets $\mathcal{C}_t$ gets enough samples, that is $N_\theta^t \geq p^\star t - \sqrt{\frac{\ln(\pi t^2/3\delta)t}{2}}$.*
4. $G_t^{(4)}$*: The minimum eigenvalue of the empirical covariance matrix formed by our testing policy grows linearly in t. Let $\lambda_{\min}^t \triangleq \lambda_{\min}\left(\sum_{s\in\mathcal{S}_P(t)} X_s X_s^\top\right)$. Then, for all $t \geq 1$:*

$$\lambda_{\min}^t \geq \frac{3}{5}t\lambda_0 - \sqrt{\frac{t}{2}\left(d\log\left(\frac{10}{\lambda_0}+1\right) + \log\left(\frac{2t^2}{\delta}\right)\right)}$$

Let $G^{(i)} = \cap_{t=1}^T G_t^{(i)}$. The good event G is the intersection of $G^{(i)}$, i.e. $G = G^{(1)} \cap G^{(2)} \cap G^{(3)} \cap G^{(4)}$.

The first event, $\mathbb{P}(G^{(1)}) \geq 1 - \delta$, follows from Lemma 1, i.e. the concentration inequality proven by [14]. $\mathbb{P}(G^{(2)}) \geq 1 - \delta$, is proved by Lemma 6. To prove that $G^{(3)}$ holds with high probability, we utilize the fact that when the optimal policy tests, then when $G^{(1)}$ and $G^{(2)}$ hold our policy does the same, as proved in Lemma 8. Observe that on $G_t^{(3)}$, 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, $\mathbb{P}(G^{(4)}) \geq 1 - \delta$ we use a covering argument to bound the minimum eigenvalue of the covariance matrix. For sufficiently large constant T_0 (only a function of δ) we have that for all $T \geq T_0$, that $\lambda_{\min}^t \geq t\lambda_{\min}/4$. We see that G occurs with high probability in the following Lemma.

Lemma 10.

$$\mathbb{P}(G) = \mathbb{P}(G^{(1)} \cap G^{(2)} \cap G^{(3)} \cap G^{(4)}) \geq 1 - 5\delta.$$

E Safety Analysis

Before moving to the regret guarantees of our algorithm we must first show it satisfies the safety constraints. Our primary tools to prove so are 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, $\xi_t \leq \xi_t^\star$ a.s. This implies a total error probabiltiy bound.

Lemma 11. *On the good event G, the total error probability of the algorithm is upper bounded by α with probability at least $1 - \delta$.*

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

F.1 Stability of $\tau^\star$ with respect to $\hat{P}_t$

We remind the reader that for a fixed value of θ, $\hat{P}_t$ represents the empirical distribution of contexts selected from $\mathcal{S}_{\hat{P}}$ to estimate the unknown distribution P, specifically its projection onto the vector θ.

F.1.1 Proof of Lemma 4

Proof. First, we collect a context as a sample at every odd round, so at round t it holds that $\left|\mathcal{S}_{\hat{P}}^t\right| = \lceil t/2 \rceil$.

$$\begin{aligned} p_{\text{err}}(\theta, \hat{P}_T, \tau) - p_{\text{err}}(\theta, P, \tau) &= \int (1 + \exp(|x^\top \theta|))^{-1} \mathbb{1}\left\{|x^\top \theta| > \tau\right\} \hat{P}_t(dx) - p_{\text{err}}(\theta, P, \tau) \\ &= \frac{1}{\lceil t/2 \rceil} \sum_{t=1}^{\lceil t/2 \rceil} \left((1 + \exp(|x_i^\top \theta|))^{-1} \mathbb{1}\left\{|x_i^\top \theta| > \tau\right\} - p_{\text{err}}(\theta, P, \tau)\right) \end{aligned} \tag{43}$$

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

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

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

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

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

□

F.2 Stability of $\tau^\star$ with respect to θ

F.2.1 Proof of Lemma 5

Proof.

$$
\begin{aligned}
p_{\text{err}}(\theta,\rho,\tau) &= \int (1+\exp(|x^\top\theta|))^{-1}\mathbb{1}\left\{|x^\top\theta|>\tau\right\}\rho(dx) \\
&= \int (1+\exp(|x^\top\theta' + x^\top(\theta-\theta')|))^{-1}\mathbb{1}\left\{|x^\top\theta' + x^\top(\theta-\theta')|>\tau\right\}\rho(dx) \\
&\leq \int (1+\exp(|x^\top\theta'| - |x^\top(\theta-\theta')|))^{-1}\mathbb{1}\left\{|x^\top\theta'|>\tau - |x^\top(\theta-\theta')|\right\}\rho(dx) \\
&\leq \int \left((1+\exp(|x^\top\theta'|))^{-1} + |x^\top(\theta-\theta')|\right)\mathbb{1}\left\{|x^\top\theta'|>\tau - |x^\top(\theta-\theta')|\right\}\rho(dx) \\
&= p_{\text{err}}(\theta',\rho,\tau - |x^\top(\theta-\theta')|) + \int |x^\top(\theta-\theta')|\mathbb{1}\left\{|x^\top\theta'|>\tau - |x^\top(\theta-\theta')|\right\}\rho(dx) \\
&\leq p_{\text{err}}(\theta',\rho,\tau - |x^\top(\theta-\theta')|) + \|\theta-\theta'\|_{V_t}\|x\|_{V_t^{-1}}\mathbb{P}_\rho\left(|x^\top\theta'|>\tau - |x^\top(\theta-\theta')|\right) \\
&\leq p_{\text{err}}(\theta',\rho,\tau - \|\theta-\theta'\|_{V_t}\|x\|_{V_t^{-1}}) + \|\theta-\theta'\|_{V_t}\|x\|_{V_t^{-1}}
\end{aligned}
$$

□

F.2.2 Proof of lemma 6

Proof. To extend Lemma 4 to hold universally for all $\theta_Q \in \mathcal{Q}_\theta$ and $\tau_Q \in \mathcal{Q}_\tau$, we define two ε_Q-nets and union bound over them. We need to notice here that there is no need to study the stability of p_{err} with respect to θ, τ now and complete the covering argument argument after taking a union bound.

By Lemma 4 we know that for any fixed θ, τ

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

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 [47] we have that the covering numbers of $\mathcal{S}^{d-1}$ satisfy for any $\varepsilon_Q > 0$;

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

For any $\varepsilon_Q < 1$ it is true that $\mathcal{Q}_\theta \leq (\frac{3}{\varepsilon_Q})^d$. By taking the union bound over all $\theta_Q \in \mathcal{Q}_\theta$ we have

$$
\mathbb{P}\left\{\left|p_{\text{err}}(\theta_Q,\hat{P}_t,\tau) - p_{\text{err}}(\theta_Q,P,\tau)\right| \leq \sqrt{\frac{d\log(\frac{3}{\varepsilon_Q}) + \log(\frac{\pi^2 t^2}{3\delta})}{4t}}, \forall t\geq 1, \theta_Q\in\mathcal{Q}_\theta\right\} \geq 1-\delta.
$$

Now, it remains to union bound over τ_Q. As τ lives in $[0,1]$, an ε-net of the unit segment in the real line is $\{\epsilon, 2\epsilon, \dots, \lfloor\frac{1}{\epsilon}\rfloor\epsilon\}$. It holds that $|\mathcal{Q}_\tau| \leq \frac{1}{\varepsilon_\tau}$. By taking the union bound over all $\tau_Q \in \mathcal{Q}_\tau$ we have

$$
\mathbb{P}\left\{\left|p_{\text{err}}(\theta_Q,\hat{P}_t,\tau_Q) - p_{\text{err}}(\theta_Q,P,\tau_Q)\right| \leq \sqrt{\frac{d\log(\frac{3}{\varepsilon_Q}) + \log(\frac{1}{\epsilon_Q}) + \log(\frac{\pi^2 t^2}{3\delta})}{4t}}, \forall t\geq 1, \theta_Q\in\mathcal{Q}_\theta, \tau_Q\in\mathcal{Q}_\tau\right\} \geq 1-\delta.
$$

We can choose the values of $\varepsilon_\theta, \varepsilon_\tau$ to be arbitrarily small. In fact, any value of order $o(1/T)$ works, although choosing $\varepsilon_Q = \frac{1}{T^2}$ requires the knowledge of the horizon T so we choose $\varepsilon_Q = \frac{1}{t^2}$. At the analysis of Theorem 1 we will see why this choice works.

As stated in the Lemma 6 we use $\zeta_t = \sqrt{\frac{d\log(\frac{3}{\varepsilon_Q})+\log(\frac{1}{\epsilon_Q})+\log(\frac{\pi^2 t^2}{3\delta})}{4t}}$.

Conditioning on the good event G, we have that

$$\begin{aligned}
\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha) &= \min\{\tau \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, \hat{P}_t, \tau_Q) \le \alpha\} \\
&\overset{(a)}{\le} \min\left\{\tau \in \mathcal{Q}_\tau : p_{\text{err}}(\theta_Q, P, \tau) \le \alpha - \zeta_t\right\} \\
&\overset{(b)}{\le} \min\left\{\tau \in \mathcal{Q}_\tau : p_{\text{err}}(\theta, P, \tau - \varepsilon_Q) \le \alpha - \zeta_t - \varepsilon_Q\right\} \\
&\overset{(c)}{\le} \min\left\{\tau \in \mathcal{Q}_\tau : p_{\text{err}}(\theta^\star, P, \tau) \le \alpha - \zeta_t - \varepsilon_Q - \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}}\right\} + \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}} + \varepsilon_Q \\
&\le \min\left\{\tau \in [0,1] : p_{\text{err}}(\theta^\star, P, \tau) \le \alpha - \zeta_t - \varepsilon_Q - \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}}\right\} + \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}} + 2\varepsilon_Q \\
&= \tau^\star\left(\theta^\star, P, \alpha - \zeta_t - \varepsilon_Q - \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}}\right) + \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}} + 2\varepsilon_Q
\end{aligned} \tag{44}$$

Where inequality (a) follows from conditioning on the good event G and using the inequality. For (b) we used that for every $\theta_Q \in \mathcal{Q}_\theta$ there exists a $\theta \in \Theta$ such that $\|\theta_Q - \theta\|_2 \le \varepsilon_Q$, the stability in θ lemma (instead of Holder inequality we used Cauchy-Schwartz) $p_{err}(\theta_Q, P, \tau) \le p_{err}(\theta, P, \tau - \|\theta_Q - \theta\|_2 \|X\|_2) + \|\theta_Q - \theta\|_2 \|X\|_2 \le p_{err}(\theta, P, \tau - \|\theta_Q - \theta\|_2) + \|\theta_Q - \theta\|_2 \le p_{err}(\theta, P, \tau - \varepsilon_Q) + \varepsilon_Q$. Finally, (c) follows from the Lemma 5.

Moreover, with a similar method we can derive a lower bound for $\tau_Q^\star(\theta_Q, \hat{P}_t, \alpha)$.

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

where (a) follows by the good event G, and (b) by the covering argument and Lemma 5. Now, we will lower bound $\tau^\star(\theta, P, \alpha)$ in terms of $\tau^\star(\theta^\star, P, \alpha)$.

$$\tau^\star(\theta, P, \alpha) = \min\{\tau \in [0,1] : p_{err}(\theta, P, \tau) \le \alpha\} \tag{45}$$

$$\overset{(a)}{\ge} \min\{\tau \in [0,1] : p_{err}(\theta^\star, P, \tau - \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}}) - \|\theta - \theta^\star\|_{V_t} \|X_t\|_{V_t^{-1}} \le \alpha\} \tag{46}$$

$$\overset{(b)}{\ge} \min\{\tau \in [0,1] : p_{err}(\theta^\star, P, \tau) - 2B_t \|X_t\|_{V_t^{-1}} \le \alpha\} \tag{47}$$

$$= \tau^\star(\theta^\star, P, \alpha + 2B_t \|X_t\|_{V_t^{-1}}), \tag{48}$$

where (a) follows from Lemma 5, and (b) from monotonicity of p_{err} with respect to τ and the fact that $\theta, \theta^\star \in \mathcal{C}_t$.

Putting all together we have that for all $\theta \in \mathcal{C}_t$:

$$\hat{\tau}(\theta, \hat{P}_t, \alpha) \ge \tau^\star(\theta^\star, P, \alpha + \zeta_t + 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t.$$

□

F.3 Stability of $\tau^\star$ with respect to α

We begin by defining a lemma bounding the probability in the annulus:

Lemma 12 (Probability in annulus). *Under Assumption 2, for all $\tau \in [0,1]$ we have that*

$$m \cdot 2\pi \arccos(\tau + \lambda)\lambda \le \mathbb{P}\left(\tau < |X^\top \theta^\star| \le \tau + \lambda\right) \le M \cdot 2\pi \arccos(\tau)\lambda \tag{49}$$

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

$$\text{Vol}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right) = \text{Vol}\left(|X^\top \theta^\star| > \tau\right) - \text{Vol}\left(|X^\top \theta^\star| \geq \tau + \lambda\right) \tag{50}$$

where $\|\theta^\star\| = 1$, and X lives on the unit sphere.

Geometrically, we see that this is simply the difference between two sphere caps: one with radius $\arccos(\tau)$ and one with $\arccos(\tau + \lambda)$.

The annulus we are trying to study has inner radius $\arccos(\tau)$ and outer radius $\arccos(\tau + \lambda)$. Using the fact that the density is bounded between m and M, we have that we can also bound the surface area of the annulus by the rectangular strip with height λ and width $2\pi \arccos(\tau)$, or $2\pi \arccos(\tau + \lambda)$.

Thus, we have that

$$m \cdot 2\pi \arccos(\tau + \lambda)\lambda \leq \mathbb{P}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right) \leq M \cdot 2\pi \arccos(\tau)\lambda \tag{51}$$

□

Proof of Lemma 7.

Proof.

$$\begin{aligned}
&p_{\text{err}}(\theta^\star, P, \tau) - p_{\text{err}}(\theta^\star, P, \tau + \lambda) \\
&\qquad = \int (1 + \exp(|x^\top \theta^\star|))^{-1} \mathbb{1}\left\{\tau < |x^\top \theta^\star| \leq \tau + \lambda\right\} P(dx) \\
&\qquad \in [(1 + \exp(\tau + \lambda))^{-1}, (1 + \exp(\tau))^{-1}] \cdot \mathbb{P}\left(\tau < |X^\top \theta^\star| \leq \tau + \lambda\right)
\end{aligned} \tag{52}$$

Relating this back to p_{err} yields

$$\begin{aligned}
2m\pi \arccos(\tau + \lambda)\lambda(1 + \exp(\tau + \lambda))^{-1} &\leq p_{\text{err}}(\theta^\star, P, \tau) - p_{\text{err}}(\theta^\star, P, \tau + \lambda) \\
&\leq 2M\pi \arccos(\tau)\lambda(1 + \exp(\tau))^{-1}
\end{aligned}$$

This means that for all θ,$\hat{P}$, and α, on the good event G_T, we have that

$$\begin{aligned}
\tau^\star(\theta, P, \alpha - \gamma) &= \min\{\tau \in \mathcal{N}([0, 1], \varepsilon_Q) : p_{\text{err}}(\theta, P, \tau) \leq \alpha - \gamma\} \\
&\leq \min\{\tau \in \mathcal{N}([0, 1], \varepsilon_Q) : p_{\text{err}}(\theta, P, \tau - \lambda) \leq \alpha - \gamma + 2m\pi \arccos(\tau)\lambda(1 + \exp(\tau))^{-1}\} \\
&= \min\{\tau \in \mathcal{N}([0, 1], \varepsilon_Q) : p_{\text{err}}(\theta, P, \tau - \lambda^\star) \leq \alpha\} \\
&\leq \tau^\star(\theta, P, \alpha) + \lambda^\star + \varepsilon_Q.
\end{aligned}$$

where $\lambda^\star$ is chosen such that $2m\pi \arccos(\tau)(1 + \exp(\tau))^{-1}\lambda^\star = \gamma$, i.e.

$$\lambda^\star = \lambda^\star(\gamma) = \frac{\gamma(1 + \exp(\tau))}{2m\pi \arccos(\tau)} \tag{53}$$

□

G Other proofs

G.1 Proof of Lemma 8

Proof. Using Equation (37), we know that when G holds then for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$ and $\alpha_t \in [0, 1]$:

$$\hat{\tau}(\theta_Q, \hat{P}_t, \alpha_t) \geq \tau^\star(\theta^\star, P, \alpha_t + \zeta_t + 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t.$$

By selecting $\alpha_t \triangleq \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}$ we have that for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$:

$$\hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) - \zeta_t \geq \tau^\star(\theta, P, \alpha).$$

Now, we can lower bound Z_t as

$$\begin{aligned}
Z_t &= \min_{\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta} \mathbb{1}\{|\langle X, \theta \rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t - \varepsilon_Q \le 0\} \\
&\ge \mathbb{1}\{|\langle X, \theta^\star \rangle| - \hat{\tau}(\theta^\star, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t \le 0\} \\
&\ge \mathbb{1}\{|\langle X, \theta^\star \rangle| - \tau^\star(\theta^\star, P, \alpha) \le 0\} \\
&= Z^\star.
\end{aligned}$$

□

G.2 Proof of Lemma 9

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

$$|\langle X_t, \theta \rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t > \varepsilon_Q$$

Using Equation (37), we know that when G holds then for all $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$:

$$\begin{aligned}
&\hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) - \zeta_t \ge \tau^\star(\theta^\star, P, \alpha) \\
&\qquad\qquad \implies |\langle X_t, \theta \rangle| \ge \tau^\star(\theta^\star, P, \alpha) + \varepsilon_Q.
\end{aligned}$$

For any $\theta \in \mathcal{C}_t \cap \mathcal{Q}_\theta$ there exists a $\theta' \in \mathcal{C}_t$ such that $\|\theta' - \theta\| \le \varepsilon_Q$. We can bound then $|\langle X_t, \tilde{\theta} \rangle| \le |\langle X_t, \theta' \rangle| + \varepsilon_Q$.

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

$$\begin{aligned}
|\langle X_t, \theta \rangle| &\ge \tau^\star(\theta^\star, P, \alpha), \\
|\langle X_t, \theta^\star \rangle| &\ge \tau^\star(\theta^\star, P, \alpha) > 0.
\end{aligned}$$

The prediction of our policy is $\hat{Y}_t = \mathbb{1}\{\langle X_t, \theta_t^L \rangle > 0\}$ and $Y_t^\star = \mathbb{1}\{\langle X_t, \theta^\star \rangle > 0\}$. In order to $\hat{Y}_t \ne Y_t^\star$ it must hold $\langle X, \theta_t^L \rangle \langle X, \theta^\star \rangle < 0$. By the Intermediate Value Theorem, or more specifically Bolzano theorem, there exists a $\theta' \in \mathcal{C}_t$ such that $\langle X, \theta' \rangle = 0$. This is a contradiction as for all $\theta \in \mathcal{C}_t$ we have that $|\langle X_t, \theta \rangle| \ge \tau^\star(\theta^\star, P, \alpha) > 0$.

□

G.3 Proof of lemma 13

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 $s > 0$

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

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

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

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

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

□

G.4 Proof of lemma 11

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 $\mu_\xi = \mathbb{E}\left[\sum_{t=1}^{T}\xi_t^\star\right]$, it is true that

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

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

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

Here, taking $\alpha \triangleq \alpha - \sqrt{\frac{\log(4/\delta)}{2T}}$ yields the desired result, where we use Lemma 16 to get that $\mathbb{P}(\bar{G}) \leq \delta/2$.

□

G.5 Proof of Lemma 15

Proof of Lemma 15. Let the random variable $Z_t^v \triangleq v^\top A_t v - \mathbb{E}[v^\top A_t v \mid \mathcal{F}_{t-1}]$, such that $v \in \mathcal{S}^{d-1}$. Notice that Z_t^v is a martingale difference sequence as;

1.

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

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 [7], 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 [47] 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

$$\mathbb{P}\left\{\exists \upsilon_i \in \mathcal{N}(\mathcal{S}^{d-1}, \varepsilon) : \sum_{s=0}^{t}(\upsilon_i^\top A_s \upsilon_i - \mathbb{E}[\upsilon_i^\top A_s \upsilon_i \mid \mathcal{F}_{s-1}]) \leq -\sqrt{\frac{[d\log(2/\varepsilon + 1) + \log(\frac{2t^2}{\delta})]t}{2}} \;\; \forall t \in \mathbb{N}\right\} \leq \delta \tag{54}$$

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.

$$\begin{aligned}
|\upsilon_t^{\star\top}\sum_{s=0}^{t} A_s\upsilon_t^\star - \upsilon_{i_t}^\top\sum_{s=0}^{t} A_s\upsilon_{i_t}| &= |\upsilon_t^{\star\top}\sum_{s=0}^{t} A_s\upsilon_t^\star - \upsilon_t^{\star\top}\sum_{s=0}^{t} A_s\upsilon_{i_t} + \upsilon_t^{\star\top}\sum_{s=0}^{t} A_s\upsilon_{i_t} - \upsilon_{i_t}^\top\sum_{s=0}^{t} A_s\upsilon_{i_t}| \\
&= |\upsilon_t^{\star\top}\sum_{s=0}^{t} A_s(\upsilon_t^\star - \upsilon_{i_t}) + (\upsilon_t^\star - \upsilon_{i_t})^\top\sum_{s=0}^{t} A_s\upsilon_{i_t}| \\
&= |(\upsilon_t^\star - \upsilon_{i_t})^\top\sum_{s=0}^{t} A_s(\upsilon_{i_t} + \upsilon_t^\star)| \\
&\leq \|\upsilon_t^\star - \upsilon_{i_t}\|_2 \left\|\sum_{s=0}^{t} A_s(\upsilon_{i_t} + \upsilon_t^\star)\right\|_2 \\
&\leq \varepsilon\sum_{s=0}^{t}\|A_s\|_{op}\left(\|\upsilon_{i_t}\|_2 + \|\upsilon_t^\star\|_2\right) \\
&= 2t\varepsilon.
\end{aligned} \tag{55}$$

Using inequality 54 we have

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

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

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

This finalizes the result

□

G.6 Proof of Lemma 2.

Proof of Lemma 2. 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 $c_t^\star$ as a function of X_t. We see that

$$\begin{aligned}
c_t &= \min_{\theta \in \mathcal{C}_t} |\langle X_t, \theta \rangle| - \hat{\tau}(\theta, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t \\
&\geq |\langle X_t, \theta^\star \rangle| - \hat{\tau}(\theta^\star, \hat{P}_t, \alpha - \zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \zeta_t \\
&\geq |\langle X_t, \theta^\star \rangle| - \tau^\star(\theta^\star, P, \alpha - 2\zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) - \varepsilon_Q
\end{aligned} \tag{56}$$

In the first line we used the definition of Z_t as in Lemma 8, and in the last line we used Equation (36). We note that in our algorithm we use a relaxation of this minimization problem for computational feasibility, however in our bounds we use its exact definition as it is mathematically equivalent.

$$\begin{aligned}
\mathbb{E}R_t &= \mathbb{E}[Z_t - Z|G] \\
&= \mathbb{P}\left(\{c_t^\star \leq 0\} \cap \{|\langle X_t, \theta^\star \rangle| \geq \tau^\star(\theta^\star, P)\} |G\right) \\
&\overset{a}{\leq} \mathbb{P}\Big(\tau^\star(\theta^\star, P, \alpha) \leq |\langle X_t, \theta^\star \rangle| \leq \tau^\star(\theta^\star, P, \alpha - 2\zeta_t - 2B_t \|X_t\|_{V_t^{-1}}) + \varepsilon_Q |G\Big) \\
&\overset{b}{\leq} \mathbb{P}\Big(\tau^\star(\theta^\star, P, \alpha) \leq |\langle X_t, \theta^\star \rangle| \leq \tau^\star(\theta^\star, P, \alpha - 2\zeta_t - 2\frac{B_t}{\sqrt{\lambda_{\min}^t}}) + \varepsilon_Q |G\Big) \\
&\overset{c}{\leq} \mathbb{P}\Big(\tau^\star(\theta^\star, P, \alpha) \leq |\langle X_t, \theta^\star \rangle| \leq \tau^\star(\theta^\star, P, \alpha - 2\zeta_t - 2\frac{B_t}{\sqrt{t\lambda_{\min}}}) + \varepsilon_Q |G\Big) \\
&\overset{d}{\leq} \mathbb{P}\Big(\tau^\star(\theta^\star, P, \alpha) \leq |\langle X_t, \theta^\star \rangle| \leq \tau^\star(\theta^\star, P, \alpha) + \lambda^\star(2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_{\min}}}) + 2\varepsilon_Q)|G\Big) \\
&\overset{e}{\leq} 2\pi M \arccos(\tau^\star(\alpha)) \left(1 + \frac{1+e}{2m\pi \arccos(\tau^\star(\alpha))}\right) \left(\lambda^\star(2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_{\min}}}) + 2\varepsilon_Q\right)
\end{aligned} \tag{57}$$

a) follows by the upper bounding of the thresholding condition. b) follows by using that $\|X_t\|_{V_t^{-1}} \leq \frac{1}{\sqrt{\lambda^t_{\min}}} \|X_t\|_2 \leq \frac{1}{\sqrt{\lambda^t_{\min}}}$, and the monotonicity of $\tau^\star$. c) follows by the sub-event $G^{(4)}$ of the good event and the bound of lemma 15. d) follows using Lemma 7, with $\gamma = 2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_{\min}}}$. e) follows from Lemma 12.

□

G.7 Proof of Theorem 1.

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

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

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, $\sum_{t=1}^{T} \varepsilon_Q \leq \sum_{t=1}^{\infty} 1/t^2 = \pi^2/6 = o(T)$.

For $\sum_{t=1}^{T} \lambda^\star(2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_{\min}}})$ we have that:

$$\begin{aligned}
\sum_{t=1}^{T} \lambda^\star(2\zeta_t + 2\frac{B_t}{\sqrt{t\lambda_{\min}}}) &= \frac{2(1+e)}{2m\pi \arccos(\tau^\star(\alpha))} \sum_{t=1}^{T} \zeta_t + \frac{2(1+e)}{2m\pi \arccos(\tau^\star(\alpha))} \sum_{t=1}^{T} \frac{B_t}{\sqrt{t\lambda_{\min}}} \\
&\preccurlyeq \sum_{t=1}^{T} \zeta_t + \frac{B_T}{\sqrt{\lambda_{\min}}} \sum_{t=1}^{T} 1\sqrt{t}.
\end{aligned}$$

We remind that $\zeta_t \triangleq \sqrt{\frac{2d\log(3T)+2\log(T)+\log(\frac{\pi^2 t^2}{3\delta})}{4t}} + 2\varepsilon_Q$. As a result $\sum_{t=1}^{T} \zeta_t = \tilde{\mathcal{O}}(\sqrt{dT})$. On the other side, as $B_t \triangleq 2\kappa\left(\sqrt{\lambda} + \sqrt{\log(1/\delta) + 2d\log\left(1 + \frac{t}{\kappa\lambda d}\right)}\right) = \tilde{\mathcal{O}}(\kappa\sqrt{d})$ then $\frac{B_T}{\sqrt{\lambda_{\min}}} \sum_{t=1}^{T} 1\sqrt{t} = \tilde{\mathcal{O}}(\kappa\sqrt{dT})$, as $\sum_{t=1}^{T} 1\sqrt{t} = \mathcal{O}(\sqrt{T})$.

By putting all together we have that $R_T = \tilde{\mathcal{O}}(\kappa\sqrt{\frac{dT}{\lambda_{\min}}})$. □

H Good event proof.

In Lemma 8 we proved that , with high probability, our policy tests whenever the optimal one does, that is $N^t_\Theta \geq N^t_{OPT}$ when $G^{(1)}, G^{(2)}$ 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.

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

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

In lemma Lemma 8 we proved that $\mathbb{P}(\mathcal{E}_1 \mid G^{(1)}, G^{(2)}) \geq 1-\delta$ 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 13. $\mathbb{P}(\mathcal{E}_2 \mid G^{(1)}, G^{(2)}) \geq 1-\delta.$

This implies that, for some T_0, we have that for all $t \geq T_0$

$$N^t_{OPT} \geq p^\star t/2. \tag{58}$$

Lemma 14.

$$\mathbb{P}(G^{(3)} \mid G^{(1)}, G^{(2)}) \geq 1-2\delta.$$

Proof. By taking the union bound

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \mid G^{(1)}, G^{(2)}) \geq 1-2\delta.$$

By using $G^{(3)} \supseteq \mathcal{E}_1 \cap \mathcal{E}_2$ when $G^{(1)}, G^{(2)}$ hold we conclude the proof. □

To show that $\mathbb{P}(G^{(4)}) \geq 1-\delta$ we will use a covering argument to derive a lower bound for the minimum covariance matrix.

Lemma 15. *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 $\lambda^t_{\min} \triangleq \lambda_{\min}\left(\sum_{s=0}^t A_s\right)$. Then, for $\varepsilon > 0$, the following holds:

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

We will apply this lemma for $A_t = X_t X_t^\top$. It is true that $\left\|X_t X_t^\top\right\|_{op} \leq \|X_t\|_2 = 1$. We will make again the same observation, by choosing the covering parameter as $\varepsilon = \frac{\lambda_0}{5}$, then that for some T_0' we have that for all $T \geq T_0'$

$$\lambda^t_{\min} \geq t\lambda_{\min}/4. \tag{59}$$

Lemma 16.

$$\mathbb{P}(G) = \mathbb{P}(G^{(1)} \cap G^{(2)} \cap G^{(3)} \cap G^{(4)}) \geq 1-5\delta.$$

Proof. By using the product rule we have that

$$\mathbb{P}(G^{(1)} \cap G^{(2)} \cap G^{(3)}) = \mathbb{P}(G^{(3)} \mid G^{(1)} \cap G^{(2)})\mathbb{P}(G^{(1)} \cap G^{(2)})$$

As $\mathbb{P}(G^{(1)}) \geq 1-\delta$ from Lemma 1 and $\mathbb{P}(G^{(2)}) \geq 1-\delta$ from Lemma 6, by using the union bound we have $\mathbb{P}(G^{(1)} \cap G^{(2)}) \geq 1-2\delta$. By using also Lemma 14 we have

$$\begin{aligned}\mathbb{P}(G^{(3)} \mid G^{(1)} \cap G^{(2)})\mathbb{P}(G^{(1)} \cap G^{(2)}) &\geq (1-2\delta)^2 \\ &\geq 1-4\delta.\end{aligned}$$

As $\mathbb{P}(G^{(4)}) \geq 1-\delta$ by Lemma 15, by taking the union bound again we have that

$$\mathbb{P}(G^{(1)} \cap G^{(2)} \cap G^{(3)} \cap G^{(4)}) \geq 1-5\delta.$$

□

I 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 $\widetilde{O}(\sqrt{dT/\lambda_0})$ 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 (3). Finally, here we utilized simple confidence bounds for our logistic bandits. Plugging the recent radius-free concentration results of [31] 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 [28] 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.