Active Learning with Safety Constraints

Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, Kevin Jamieson
University of Washington, Seattle, WA
{camilr,ajwagen,jamiemmt,jamieson}@cs.washington.edu,lalitj@uw.edu

June 23, 2022

Abstract

Active learning methods have shown great promise in reducing the number of samples necessary for learning. As automated learning systems are adopted into real-time, real-world decision-making pipelines, it is increasingly important that such algorithms are designed with *safety* in mind. In this work we investigate the complexity of learning the best *safe* decision in interactive environments. We reduce this problem to a constrained linear bandits problem, where our goal is to find the best arm satisfying certain (unknown) safety constraints. We propose an adaptive experimental design-based algorithm, which we show efficiently trades off between the difficulty of showing an arm is unsafe vs suboptimal. To our knowledge, our results are the first on best-arm identification in linear bandits with safety constraints. In practice, we demonstrate that this approach performs well on synthetic and real world datasets.

1 Introduction

In many problems in online decision-making, the goal of the learner is to take measurements in such a way as to learn a near-optimal policy. Oftentimes, though the space of policies may be large, the set of feasible, or safe policies could be much smaller, effectively constraining the search space of the learner. Furthermore, these constraints may themselves depend on unknown problem parameters.

For example, consider the problem of bidding sequentially in a series of auctions where the bidder bids a price s_t, the value of winning an item t is denoted v_t, and the utility of winning that item and paying price p_t is $v_t - p_t$. The goal of the bidder is to choose an optimal strategy amongst bidding strategies $s \in S, s : \mathbb{R} \to \mathbb{R}$. When a bidder is deciding how to choose these strategies, they often face constraints: they may have a budget B they must abide to; they may wish to have those auctions they win be well-distributed across time (e.g. in the case of advertising campaigns); they may want to ensure the set of items they win satisfy some other property (e.g. for advertisements, they might want to ensure they are not over-targeting any demographic group).

As another example, inventory management systems may face similar issues of deciding amongst strategies, where there is some objective function (such as revenue) and a variety of constraints at play in this choice (e.g. capacity of a set of warehouses, employee scheduling constraints, or limits on the duration of delivery lag). They also operate in markets with changing demand and other uncertainties, leading to uncertainty about which strategies are feasible or safe (satisfy constraints) and uncertainty about the revenue they generate.

Both of these scenarios motivate understanding the sample complexity of selecting an action or strategy which approximately maximizes an objective while also satisfying some constraints, where samples are needed to both learn the objective value of actions and whether or not they satisfy said constraints. In this work, we study the *active* sample complexity of this task—if the learner

can choose which examples to observe and have labeled, how many fewer samples might they need compared to the number needed in a passive setting? We pose this as a best-arm identification problem in the setting of linear bandits with safety constraints, where the goal is to estimate the best arm, subject to it meeting certain (initially unknown) safety constraints. We propose an experiment design-based algorithm which efficiently learns the best safe decision, and show the efficacy of this approach in practice through several experimental examples. To the best of our knowledge, ours is the first approach to handle best-arm identification in linear bandits with safety constraints.

1.1 Linear Bandits with Safety Constraints

Let $\delta \in (0,1)$ be a confidence parameter, $\mathcal{X}, \mathcal{Z} \subseteq \mathbb{R}^d$ be finite known sets of vectors, and assume there exists $\theta_* \in \mathbb{R}^d$, $\mu_* \in \mathbb{R}^{m \times d}$ unknown to the learner. For simplicity, we assume that $\|\theta_*\|_2 \leq 1$, and $\|\mu_{*,i}\|_2 \leq 1, i \in [m]$ and $\|x\|_2 \leq 1$, $\|z\|_2 \leq 1$, $\forall x \in \mathcal{X}, z \in \mathcal{Z}$. The learner plays according to the following protocol: at each time step t the learner chooses some action $x_t \in \mathcal{X}$, observes $(r_t, \{s_{t,i}\}_{i=1}^m)$ where $r_t = \theta_*^\top x_t + w_t^\theta$ and $s_{t,i} = \mu_{*,i}^\top x_t + w_{t,i}^\mu$ for all $i \in [m]$, where $w_t^\theta, w_{t,i}^\mu$ are i.i.d. mean zero 1-subGaussian noise. The choice of action x_t is measurable with respect to the history $\mathcal{F}_t = \{(x_j, r_j, \{s_{j,i}\}_{i=1}^m)\}_{j=1}^{t-1}$. The learner stops at a stopping time τ_δ which is measurable with respect to the filtration generated by $\mathcal{F}_{t \leq \tau}$, and returns $\widehat{z}_\tau \in \mathcal{Z}$. In general, when referring to any expectation $\mathbb{E}$ or probability $\mathbb{P}$, the underlying measure will be with respect to the actions, observed rewards, and internal randomness of the algorithm.

We are interested in the *safe transductive best-arm identification problem* (**STBAI**), where the goal of the learner is to identify

$$z_* := \max_{z \in \mathcal{Z}} z^\top \theta_* \quad \text{s.t.} \quad z^\top \mu_{*,i} \leq \gamma, \forall i \in [m]$$

for some (known) threshold γ. In words, our goal is to identify the best *safe* arm in $\mathcal{Z}$, z_*, where we say an arm z is safe if it satisfies every linear constraint: $z^\top \mu_{*,i} \leq \gamma, \forall i \in [m]$. We are interested in obtaining learners that take the fewest number of samples possible to accomplish this. In practice, we will consider a slightly easier objective. Fix some tolerance $\epsilon > 0$ and let

$$\mathcal{Z}_\epsilon := \{z \in \mathcal{Z} \ : \ z^\top \theta_* \geq z_*^\top \theta_* - \epsilon, \ z^\top \mu_{*,i} \leq \gamma + \epsilon, \forall i \in [m]\}.$$

Then our goal is to obtain an (ϵ, δ)-PAC learner defined as follows:

Definition 1.1 ((ϵ, δ)-PAC Learner)**.** A learner is (ϵ, δ)-PAC if for any instance it returns $\widehat{z}_\tau$ such that $\mathbb{P}[\widehat{z}_\tau \in \mathcal{Z}_\epsilon] \geq 1 - \delta$.

We define the *optimality gap* for any $z \in \mathcal{Z}$ as $\Delta(z) := \theta_*^\top (z_* - z)$, and the *safety gap* for constraint i as $\Delta_{\text{safe}}^i(z) := \gamma - \mu_{*,i}^\top z$. Note that either $\Delta(z)$ or $\Delta_{\text{safe}}^i(z)$ can be negative. If $\Delta(z) < 0$, it follows that z has larger *value*—$z^\top \theta_*$—than the best safe arm z_*, which implies it must be unsafe. If $\Delta_{\text{safe}}^i(z) < 0$ for some i, then arm z is unsafe. We also define the *ϵ-safe optimality gap* as:

$$\Delta^\epsilon(z) = \max_{z' \in \mathcal{Z}} (z' - z)^\top \theta_* \quad \text{s.t.} \quad \min_{i \in [m]} \Delta_{\text{safe}}^i(z) \geq \epsilon. \tag{1.1}$$

$\Delta^\epsilon(z)$ is then the gap in value between arm z and the best arm with minimum safety gap at least ϵ.

Mathematical Notation. Let $\|x\|_A^2 = x^\top A x$ and $\mathfrak{p}(x) := \max\{x, 0\}$. $\widetilde{\mathcal{O}}(\cdot)$ hides factors that are logarithmic in the arguments. $\lesssim$ denotes inequality up to constants. We denote the simplex as $\triangle_{\mathcal{X}} := \{\lambda \in \mathbb{R}_{\geq 0}^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} \lambda_x = 1\}$.

2 Safe Best-Arm Identification in Linear Bandits

2.1 Algorithm Definition

Algorithm 1 **Be**st **S**afe Arm **Id**entification (BESIDE)

input: tolerance ϵ, confidence δ
$\iota_\epsilon \leftarrow \lceil \log(\frac{20}{\epsilon}) \rceil$, $\widehat{\Delta}^{i,0}_{\text{safe}}(z) \leftarrow 0, \widehat{\Delta}^{0}(z) \leftarrow 0$ for all $z \in \mathcal{Z}$
for $\ell = 1, 2, \ldots, \iota_\epsilon$ **do**
    $\epsilon_\ell \leftarrow 20 \cdot 2^{-\ell}$
// Phase 1: Solve design to reduce uncertainty in safety constraints
    Define

$$c_\ell(z) = \min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))$$

    Let τ_ℓ be the minimal value of $\tau \in \mathbb{R}_+$ which is greater than $4 \log \frac{4m|\mathcal{Z}|\ell^2}{\delta}$ such that the objective to the following is no greater than $\epsilon_\ell/100$, and λ_ℓ the corresponding optimal distribution

$$\inf_{\lambda \in \triangle_\mathcal{X}} \max_{z \in \mathcal{Z}} -\frac{1}{100}\left(c_\ell(z) + \epsilon_\ell\right) + \sqrt{\tau^{-1} \cdot \|z\|^2_{A(\lambda)^{-1}} \cdot \log(\tfrac{4m|\mathcal{Z}|\ell^2}{\delta})}$$

    Sample $x_t \sim \lambda_\ell$, collect τ_ℓ observations $\{(x_t, r_t, s_{t,1}, \ldots, s_{t,m})\}_{t=1}^{\tau_\ell}$
// Phase 2: Estimate safety constraints
    $\{\widehat{\mu}^{i,\ell}\}_{i=1}^m \leftarrow \mathsf{RIPS}(\{(x_t, s_{t,i})\}_{t=1}^{\tau_\ell}, \mathcal{Z}, \frac{\delta}{2m\ell^2})$
    $\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) \leftarrow \gamma - z^\top \widehat{\mu}^{i,\ell} + \|z\|_{A(\lambda_\ell)^{-1}} \sqrt{\tau_\ell^{-1} \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}$
// Phase 3: Refine estimates of optimality gaps
    $\{\widehat{\Delta}^{\ell}(z)\}_{z \in \mathcal{Z}} \leftarrow \text{RAGE}^\epsilon\Big(\mathcal{Z}, \mathcal{Y}_\ell, \epsilon_\ell, \frac{\delta}{4\ell^2}, \{\widehat{\Delta}_{\text{safe}}(z) \leftarrow \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell}_{\text{safe}}(z))\}_{z \in \mathcal{Z}}\Big)$
// Perform final round of exploration to ensure we find ϵ-good arm
$\mathcal{Y}_{\text{end}} \leftarrow \{z \in \mathcal{Z} : c_\ell(z) \lesssim \widehat{\Delta}^{i,\ell}_{\text{safe}}(z) + \epsilon\}$
$\{\widehat{\Delta}^{\text{end}}(z)\}_{z \in \mathcal{Y}_{\text{end}}} \leftarrow \text{RAGE}^\epsilon(\mathcal{Y}_{\text{end}}, \mathcal{Y}_{\text{end}}, \epsilon, \delta, \{\widehat{\Delta}_{\text{safe}}(z) \leftarrow \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell}_{\text{safe}}(z))\}_{z \in \mathcal{Z}})$
return $\widehat{z} = \arg\min_{z \in \mathcal{Y}_{\text{end}}} \widehat{\Delta}^{\text{end}}(z)$

The main challenge in algorithm design for the safe best-arm identification problem is ensuring that we are efficiently balancing our exploration between refining our estimates of both the safety gaps, as well as the optimality gaps. Our approach is given in Algorithm 1, BESIDE.

BESIDE relies on a round-based adaptive experimental design approach. In each round BESIDE consists of three phases. In the first phase, it solves an experimental design over $\lambda_\ell \in \triangle_\mathcal{X}$, with the goal of refining our estimates of the safety gaps. It then takes τ_ℓ samples from λ_ℓ. In the second phase these samples are used to estimate the safety constraints, $\widehat{\mu}^{i,\ell}$, and the safety gaps of each arm, $\widehat{\Delta}^{i,\ell}_{\text{safe}}(z)$. Finally, in Phase 3, an additional experimental design is solved which now aims to refine our estimates of the optimality gaps, and the estimates of the optimality gaps $\widehat{\Delta}^{\ell}(z)$ for each $z \in \mathcal{Z}$ are then computed. We encapsulate Phase 3 in a subroutine, RAGE^ϵ, which we outline in the following. We now carefully describe each phase—we begin with Phase 2 to explain how our estimator works.

Phase 2: In Phase 2 the algorithm would like to use the τ_ℓ samples drawn from the design λ_ℓ to estimate the constraints for each $z \in \mathcal{Z}$: $z^\top \mu_{*,i}$ for each $i \in [m]$. Past works using adaptive experimental design in the linear bandits literature have utilized the least-squares estimator along with complicated rounding schemes Fiez et al. [2019] which may require an additional poly(d) samples each round (this poly(d) factor could be prohibitively large—for example, in active classification problems, d is the total number of data points). We instead utilize the RIPS estimator of Camilleri et al. [2021a] which gives us a guarantee of the form: with probability greater than $1-\delta$, for all $z \in \mathcal{Z}$,

$$|z^\top(\widehat{\mu}^{i,\ell} - \mu_{*,i})| \lesssim \|z\|_{A(\lambda_\ell)^{-1}} \cdot \sqrt{\tau_\ell^{-1} \log(\tfrac{4m|\mathcal{Z}|\ell^2}{\delta})}. \tag{2.1}$$

We describe the RIPS estimator in more detail in Appendix B.

Phase 1: By our definition of the experimental design on Line 6, our safety gap estimation error bound in (2.1) satisfies, for each $z \in \mathcal{Z}$:

$$|z^\top(\widehat{\mu}^{i,\ell} - \mu_{*,i})| \lesssim \|z\|_{A(\lambda_\ell)^{-1}} \cdot \sqrt{\tau_\ell^{-1} \log(\tfrac{4m|\mathcal{Z}|\ell^2}{\delta})} \lesssim c_\ell(z) + \epsilon_\ell. \tag{2.2}$$

Note that our design chooses an allocation that minimizes the variance in our estimate of each safety constraint (up to some tolerance), which scales as $\|z\|^2_{A(\lambda)^{-1}}$. This can be thought of as a form of $\mathcal{XY}$*-design*—a design of the form $\inf_{\lambda \in \triangle_\mathcal{X}} \max_{y \in \mathcal{Y}} \|y\|^2_{A(\lambda)^{-1}}$—where here $\mathcal{Y} \leftarrow \mathcal{Z}$ is chosen to reduce our uncertainty in estimating the safety value for each $z \in \mathcal{Z}$. We refer to such a design objective henceforth as $\mathcal{XY}_{\text{safe}}$. Assume that at round $\ell - 1$, we can guarantee

$$\begin{aligned} c_\ell(z) &= \min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell \\ &\lesssim \min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\epsilon_{\ell-1}}(z)) + \epsilon_\ell. \end{aligned} \tag{2.3}$$

Then combining the above inequalities, we see that the experiment design on Line 6 aims to minimize the uncertainty in our estimate of $z^\top \mu_{*,i}$ up to a tolerance that scales as the maximum of the four terms in (2.3). It follows that if any of these terms is large, we will only allocate a small number of samples to refining our estimate of arm z. Each one of these terms can be intuitively motivated by thinking through what is needed to prove that an arm $z \neq z_*$.

- z **has small safety gap** $\min_j |\Delta^j_{\text{safe}}(z)|$**:** if this term is large, it implies that minimum safety gap for z is large. To show an arm is safe or unsafe, it suffices to learn each safety gap up to a tolerance a constant factor from its value—regularizing by this term ensures we do just that.
- z **fails some safety constraint** $\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z))$**:** if this term is large, it implies that arm z is very unsafe for some constraint. In this case, we can easily determine z is unsafe, and therefore do not need to reduce our uncertainty in the safety gap any more.
- z **is sub-optimal** $\mathfrak{p}(\Delta^{\epsilon_{\ell-1}}(z))$**:** if this term is large, it implies that z is very suboptimal compared to some safe arm with safety gap at least $\epsilon_{\ell-1}$. In this case, we do not need to estimate z's safety gap, as we will have already eliminated it.

It remains to ensure that (2.3) holds. As we show in Appendix D through a careful inductive argument, combining (2.2) with our guarantee on the estimates of the optimality gaps obtained in Phase 3, $\widehat{\Delta}^\ell(z)$, is sufficient to guarantee (2.3) holds. In particular, if any gap is greater than ϵ_ℓ it is estimated up to a constant factor, and otherwise it is estimated up to $\mathcal{O}(\epsilon_\ell)$. This ensures that our gaps are estimated at the correct rate while guaranteeing we do not collect too many samples in each round.

Phase 3: In this phase we estimate the suboptimality gaps using RAGE$^\epsilon$. RAGE$^\epsilon$ is inspired by the RAGE algorithm of Fiez et al. [2019] for best-arm identification. In the interest of space, we defer the full definition of RAGE$^\epsilon$ to Appendix C but provide some intuition here. After Phase 2, by (2.2) the set of arms $\mathcal{Y}_\ell := \{z \in \mathcal{Z} : c_s(z) \lesssim \widehat{\Delta}^{i,s}(z), \forall i \in [m]\}$ for $s \leq \ell$ are precisely the ones that we can certify are safe (note that we do not need to ever explicitly construct such a set—we can instead maintain an implicit definition through the constraints). RAGE$^\epsilon$ uses an adaptive experimental design procedure to sample in such a way as to optimally estimate the gaps $(z - \widehat{y})^\top \theta_*, \forall z \in \mathcal{Z}$ and some $\widehat{y} \in \mathcal{Y}_\ell$ up to some (sufficient) tolerance. In particular, it also solves an $\mathcal{XY}$-design, but now on the set $\mathcal{Y} \leftarrow \{z - \widehat{y} : z \in \mathcal{Z}\}$. Thus, rather than minimizing $\|z\|^2_{A(\lambda)^{-1}}$, we minimize $\|z - \widehat{y}\|^2_{A(\lambda)^{-1}}$. This design reduces uncertainty on the *differences* between arms, which allows us to refine our estimates of their optimality gaps. Henceforth we refer to such a design as $\mathcal{XY}_{\text{diff}}$. We describe the importance of the choice of design in more detail in Section 2.4. Ultimately, if an arm z has value within a factor of ϵ_ℓ of the best safe arm in $\mathcal{Y}_\ell$, and if we have not yet shown arm z is unsafe, then we will estimate its optimality gap up to a constant factor of ϵ_ℓ. If we were maintaining arm sets explicitly (similar to the original RAGE algorithm of Fiez et al. [2019]) we would eliminate arms at this point.

Remark 2.1 (Computational Complexity)**.** The main computational challenge in BESIDE and RAGE$^\epsilon$ is the calculation of the experimental designs (i.e. Line 6 and the corresponding design in RAGE$^\epsilon$). In general, the presence of the square root implies that the resulting optimization problem may not be convex in λ. To handle this issue we note that $2\sqrt{xy} = \min_{\alpha>0} \alpha x + \frac{y}{\alpha}$—thus we can replace the existing design with $\inf_{\lambda \in \triangle_\mathcal{X}} \max_{z \in \mathcal{Z}} \min_{\alpha > 0} -\frac{1}{100}\left(c_\ell(z) + \epsilon_\ell\right) + \alpha \|z\|^2_{A(\lambda)^{-1}} + \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})/(\alpha\tau)$. By appropriately discretizing the space we search over for τ and α we can then apply the Frank-Wolfe algorithm to minimize over λ. While computationally efficient in theory, this procedure is quite complicated and impractical for large problems. In the experiments section we provide a practical heuristic that is motivated by the above algorithm and is computationally efficient for larger problems.

2.2 Main Result

BESIDE achieves the following complexity.

Theorem 1. BESIDE *is* (ϵ, δ)*-PAC. In other words, with probability at least* $1 - \delta$*,* BESIDE *returns an arm* $\widehat{z} \in \mathcal{Z}$ *such that*

$$\widehat{z}^\top \theta_* \geq z_*^\top \theta_* - \epsilon, \quad \min_{i \in [m]} \Delta^i_{\text{safe}}(\widehat{z}) \geq -\epsilon$$

and terminates after collecting at most

$$C \cdot \sup_{\widetilde{\epsilon} \geq \epsilon} \inf_{\lambda \in \triangle_\mathcal{X}} \max_{z \in \mathcal{Z}} \frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{m|\mathcal{Z}|}{\delta})}{\left(\min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\widetilde{\epsilon}}(z)) + \widetilde{\epsilon}\right)^2} \qquad \text{(safety)}$$

$$+ C \cdot \sup_{\widetilde{\epsilon} \geq \epsilon} \inf_{\lambda \in \triangle_\mathcal{X}} \max_{z \in \mathcal{Z}} \frac{\|z - z_*\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{|\mathcal{Z}|}{\delta})}{\left(\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\widetilde{\epsilon}}(z)) + \widetilde{\epsilon}\right)^2} + C_0 \qquad \text{(optimality)}$$

samples for some $C = \operatorname{poly}\log(\frac{1}{\epsilon})$ *and* $C_0 = \operatorname{poly}\log(\frac{1}{\epsilon}, |\mathcal{Z}|) \cdot \log\frac{1}{\delta}$.

The complexity bound given in Theorem 1 may, at first glance, appear rather opaque, yet it in fact yields a very intuitive interpretation. The first term in the complexity, the safety term, is the complexity needed to show each arm is safe or unsafe, *if they have not otherwise been eliminated.* As

described in the previous section, if $\mathfrak{p}(\Delta^{\widetilde{\epsilon}}(z))$ is large, this implies we have found an arm better than z, so learning its safety value is irrelevant.

The second term in the complexity, the optimality term, corresponds to the difficulty of showing an arm is worse *than the best arm we can guarantee is safe*. Note that we can only guarantee an arm is suboptimal if we can find a safe arm with higher value. Recall the definition of $\Delta^{\widetilde{\epsilon}}(z)$ given in (1.1). Intuitively, $\Delta^{\widetilde{\epsilon}}(z)$ denotes the gap in value between arm z and the best arm with safety gap at least $\widetilde{\epsilon}$. As we make $\widetilde{\epsilon}$ smaller, we can show additional arms are safe, which increases $\Delta^{\widetilde{\epsilon}}(z)$. While this makes it easier to show z is suboptimal, it comes at a cost—the extra samples necessary to decrease our safety tolerance, given by the first term in the complexity. BESIDE trades off between optimizing for each of these terms—gradually decreasing its tolerance on both the safety and optimality terms to more easily eliminate suboptimal arms, while not allocating too many samples to guarantee safety.

To help illustrate this complexity, we consider a simple example with orthogonal arms, i.e. a multi-armed bandit example.

Example 2.1 (BESIDE on Multi-Armed Bandits)**.** In the multi-armed bandit setting, we have $\mathcal{X} = \mathcal{Z} = \{e_1, \dots, e_d\}$. Let $m = 1$, $d = 3$, and consider the settings of θ_* and μ_* given in Figure 1. Here we see that arm e_1 is safe and has value much higher than any other arm, so $z_* = e_1$, and can be shown to be safe relatively easily; arm e_2 has near-optimal value but is very unsafe; and arm e_3 is unsafe with very small safety gap, but has the smallest value.

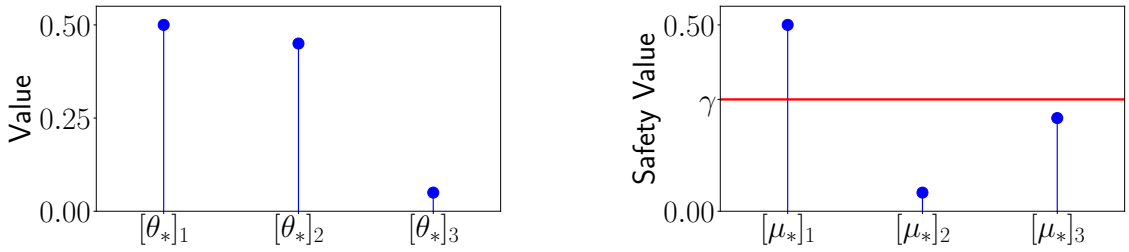

Figure 1: Multi-Armed Bandit Instance

Showing e_2 is Suboptimal. As e_2 has near-optimal value, $\Delta(e_2)$ is very small and it is very difficult to show e_2 is suboptimal. However, $-\Delta_{\text{safe}}(e_2) = \mathcal{O}(1)$, so it is very easy to show e_2 is unsafe. It follows that $\mathfrak{p}(-\Delta_{\text{safe}}(e_2)) = \mathcal{O}(1)$ so both denominators in our complexity will always be $\mathcal{O}(1)$ for $z = e_2$—BESIDE does not attempt to show e_2 is suboptimal, but instead shows it is unsafe, and therefore does not pay for the small optimality gap of $\Delta(e_2)$ in the complexity.

Showing e_3 is Suboptimal. Recall the definition of $\Delta^{\epsilon}(z) = \max_{z' : \Delta_{\text{safe}}(z') \geq \epsilon} \theta_*^\top (z' - z)$. In this case, for $\epsilon = \mathcal{O}(1)$, we will have $\Delta_{\text{safe}}(e_1) \geq \epsilon$, which implies that $\Delta^{\epsilon}(e_3) = \theta_*^\top (e_1 - e_2) = \Delta(e_3) = \mathcal{O}(1)$. To show e_3 is suboptimal, we could either show it is unsafe (which is very difficult) or suboptimal (which is very easy). Observing the sample complexity of Theorem 1, we see that the denominator of both terms will always be $\mathcal{O}(1)$ for $z = e_3$ since $\Delta^{\epsilon}(e_2) = \mathcal{O}(1)$—BESIDE never pays for the small safety gap of e_3, it instead takes advantage of the fact that e_3 can easily be shown to be suboptimal, and uses this to eliminate it.

In both of these cases we see that BESIDE does the "right" thing, always using the easier of the two criteria—either showing an arm is unsafe or suboptimal—to show that $z \neq z_*$. Combining the above observations, for $\epsilon \approx \min\{\Delta(e_3), -\Delta_{\text{safe}}(e_2), \Delta_{\text{safe}}(e_1)\}$, it follows that on this example the

total sample complexity of BESIDE given by Theorem 1 scales as:

$$\widetilde{\mathcal{O}}\left(\left(\frac{1}{\Delta_{\text{safe}}(e_1)^2}+\frac{1}{\Delta_{\text{safe}}(e_2)^2}+\frac{1}{\Delta(e_3)^2}\right)\cdot\log\frac{1}{\delta}\right)$$

where the $1/\Delta_{\text{safe}}(e_1)^2$ arises because we must also show e_1 is safe.

2.3 Optimality of BESIDE

Optimality in Best-Arm Identification. Consider applying BESIDE to a problem instance where $m = 1$, $\mu_{*,1} = 0$, and $\gamma = 1$. In this case, every arm is safe, and the safety constraints are essentially vacuous—every arm can easily be shown safe. We can therefore think of this as simply an instance of the best-arm identification problem. In this setting, we obtain the following corollary.

Corollary 1. *Consider running* BESIDE *on a problem instance where* $m = 1$, $\mu_{*,1} = 0$, *and* $\gamma = 1$, *and set* $\epsilon = \frac{1}{2}\max_{z\neq z_*}\theta_*^\top(z_* - z)$. *Then with probability at least* $1-\delta$, BESIDE *returns* z_* *and has sample complexity bounded by:*

$$\widetilde{\mathcal{O}}\left(\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{\Delta(z)^2}\cdot\log\frac{|\mathcal{Z}|}{\delta}+\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\|z\|^2_{A(\lambda)^{-1}}\cdot\log\frac{|\mathcal{Z}|}{\delta}\right).$$

Up to lower-order terms, this exactly matches the lower bound on best-arm identification given in Fiez et al. [2019]. Thus, in settings where the safety constraint is vacuous, BESIDE hits the optimal rate.

Worst-Case Performance of BESIDE. We next consider the worst-case performance of BESIDE in settings when $\mathcal{X} = \mathcal{Z}$. We have the following result.

Corollary 2. *Assume that* $\mathcal{X} = \mathcal{Z}$. *Then for any* θ_* *and* $(\mu_{*,i})_{i=1}^m$, *the sample complexity of* BESIDE *necessary to return an* ϵ*-good and* ϵ*-safe arm is bounded as* $\widetilde{\mathcal{O}}(\frac{d}{\epsilon^2}\cdot(\log(m|\mathcal{X}|)+\log\frac{1}{\delta}))$.

Theorem 2 of Wagenmaker et al. [2022] shows a worst-case lower bound of $\Omega(d^2/\epsilon^2)$ on the sample complexity of identifying an ϵ-optimal arm in the standard linear bandit setting. Safe best-arm identification problems in which the safety constraint is vacuous are at least as hard as the standard best-arm identification problem, since at minimum we need to find the best arm out of every safe arm. Thus, $\Omega(d^2/\epsilon^2)$ is also a worst-case lower bound for the safe best-arm identification problem. The hard instance of Wagenmaker et al. [2022] has $|\mathcal{X}| = \mathcal{O}(2^d)$, so it follows that on this instance, BESIDE achieves a complexity of $\widetilde{\mathcal{O}}(\frac{d}{\epsilon^2}\cdot(d+\log\frac{1}{\delta}))$, and therefore BESIDE has optimal dimensionality dependence. In addition, this also implies that safe best-arm identification, in the worst-case, is no harder than the standard best-arm identification problem—it is no harder to find the best *safe* arm, regardless of the number of safety constraints, than to find the best arm, ignoring safety constraints.

2.4 The Role of Experiment Design

We can think of the safe best-arm identification problem, in some sense, as an interpolation of the standard best-arm identification problem, as well as the level-set estimation problem, where the goal is to identify $z \in \mathcal{Z}$ satisfying $z^\top\mu_* \leq \gamma$ [Mason et al., 2021]. In the former problem, Fiez et al. [2019] shows that the instance-optimal rate can be attained by running a round-based algorithm and at every round solving an instance of the $\mathcal{XY}_{\text{diff}}$ experiment design, as defined in Section 2.1. In the latter problem, [Mason et al., 2021] also show that a round-based algorithm can hit the

instance-optimal rate, but instead solving the $\mathcal{XY}_{\text{safe}}$ problem at each round. It is natural to ask whether either of these strategies could be applied to the safe best-arm identification problem directly, or if it is necessary to alternate between them. The following results show that, on their own, each of these designs is unable to hit the optimal rate.

Proposition 1. *Fix some small enough $\epsilon > 0$. Then there exist instances of the safe best-arm identification problem, $\mathcal{I}_i = (\theta_*^i, \mu_*^i, \mathcal{X}^i, \mathcal{Z}^i)$, $i = 1, 2$, with $d = |\mathcal{X}^i| = |\mathcal{Z}^i| = 2$, $m = 1$, such that:*

- *On $\mathcal{I}^1$, any (ϵ, δ)-PAC algorithm which plays only allocations minimizing $\mathcal{XY}_{\text{diff}}$ must have $\mathbb{E}[\tau_\delta] \geq \Omega\left(\frac{1}{\epsilon^3} \cdot \log \frac{1}{\delta}\right)$, while* BESIDE *identifies an ϵ-optimal arm after $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2} \cdot \log 1/\delta)$ samples.*
- *On $\mathcal{I}^2$, any (ϵ, δ)-PAC algorithm which plays only allocations minimizing $\mathcal{XY}_{\text{safe}}$ must have $\mathbb{E}[\tau_\delta] \geq \Omega\left(\frac{1}{\epsilon^{3/2}} \cdot \log \frac{1}{\delta}\right)$, while* BESIDE *identifies an ϵ-optimal arm after $\widetilde{\mathcal{O}}(\frac{1}{\epsilon} \cdot \log 1/\delta)$ samples.*

Proposition 1 implies that, to solve the safe best-arm identification problem optimally, more care must be taken in exploring than either standard experiment design induces—we must trade off between $\mathcal{XY}_{\text{diff}}$ and $\mathcal{XY}_{\text{safe}}$ as BESIDE does. We remark briefly on the instance $\mathcal{I}^1$. On this instance we have $\mathcal{X} = \{e_1, e_2\}$ and $\mathcal{Z} = \{z_1, z_2\}$ with $z_1 = [1/4, 1/2]$ and $z_2 = [3/4, 1/2 + \alpha]$. We set $\theta_*^1 = [1, 0]$, $\mu_*^1 = [0, 1]$, and $\gamma = 1/2 + \alpha/2$. Here z_2 is unsafe while z_1 is safe, so it follows that $z_* = z_1$. As $z_2^\top \theta_*^1 > z_1^\top \theta_*^1$, to show $z_2 \neq z_*$, we must show it is unsafe. However, if we solve the design $\mathcal{XY}_{\text{diff}}$, we see that it places nearly all of the mass on the first coordinate. While this would be optimal if both z_1 and z_2 were safe and we simply wished to determine which has a higher value, to show z_2 is unsafe, the optimal strategy places (roughly) the same mass on each coordinate, since each coordinate could contribute to the safety value. This is precisely the allocation BESIDE will play, so it is able to show that z_2 is unsafe much more efficiently than a naive $\mathcal{XY}_{\text{diff}}$ approach.

3 Experiments for Safe Best Arm Identification in Linear Bandits

We next present experimental results on BESIDE to demonstrate the advantage of experimental design—especially combining $\mathcal{XY}_{\text{diff}}$ and $\mathcal{XY}_{\text{safe}}$ designs. As there are no existing algorithms that consider safe best-arm identification, as a benchmark we consider the naive adaptive approach BASELINE that first solves the problem of estimating the safety gap of each arm up to a desired tolerance, and then solves the problem of finding the best (safe) arm among the arms that were found to be safe. We first describe instances on which we test BESIDE. Our experimental details and precise implementation of BESIDE using elimination are described in Section F.

Multi-Armed Bandit. We consider a best-arm identification problem in which every arm is safe, but the arm with highest value is very difficult to identify as safe, while the second-best arm can easily be shown safe. We vary the total number of arms and run BESIDE and BASELINE with $\epsilon = 0.5$ and $\delta = 0.1$. From Figure 2, we observe that the sample complexity of BESIDE is smaller (up to about two times for 100 arms) than the sample complexity of its baseline.

Linear Response Model. *Random Instance*: We also consider the more general setup where $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, $\theta \in \mathbb{R}^d$ and $\mu \in \mathbb{R}^d$ are randomly generated from independent Gaussian random variables with mean 0 and variance 1. We set $|\mathcal{X}| = 50$ and vary the size of $|\mathcal{Z}|$. In Figure 3, we see again that BESIDE significantly outperforms the baseline.

Hard Instance: We last consider the instance of Proposition 1 and benchmark against the strategy playing only allocations minimizing $\mathcal{XY}_{\text{diff}}$. In Figure 4, we see again that BESIDE significantly outperforms this baseline, corroborating the theoretical result of Proposition 1.

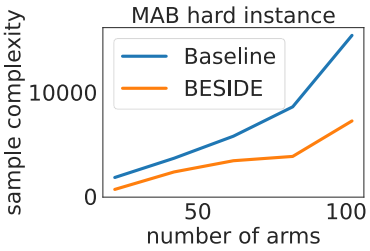

Figure 2: Total arm pulls to termination vs. number of arms

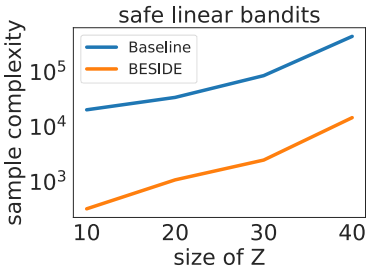

Figure 3: Total arm pulls to termination vs. $|\mathcal{Z}|$

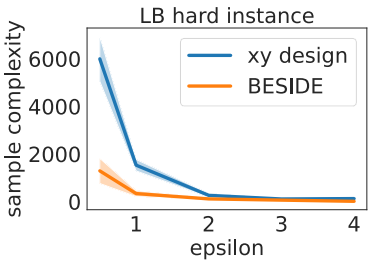

Figure 4: Total arm pulls to termination vs. ϵ

3.1 Practical Algorithms for Active Classification Under Constraints

Next, we provide an application of the above ideas to pool-based active classification with constraints—namely, adaptive sampling to learn the highest accuracy classifier with a constraint on the false discovery rate (FDR). We first explain how this problem maps to the linear bandit setting. Precisely, let $\mathcal{X}$ be the example space and $\mathcal{Y} = \{0, 1\}$ the label space. Fix a hypothesis class $\mathcal{H}$ such that each $h \in \mathcal{H}$ is a classifier $h : \mathcal{X} \to \mathcal{Y}$. We represent each h with an associated indicator vector $z_h \in \{0, 1\}^{|\mathcal{X}|}$ where $z_h(x) = 1 \iff h(x) = 1$. Similarly, let $\eta \in [0, 1]^{|\mathcal{X}|}$ represent the label distribution, i.e. $\eta(x) = \mathbb{P}(Y = 1|X = x)$. Then the risk of a classifier $R(h) := \mathbb{E}_{x\sim\text{Unif}(\mathcal{X}),Y\sim\mathbf{Ber}(\eta(x))}[\mathbb{1}[h(x) \neq Y]] = z_h^\top(2\eta - 1)$ and the FDR is defined as $\text{FDR}(h) := (\mathbf{1} - \eta)^\top z / \mathbf{1}^\top z$. In the case when $\eta \in \{0, 1\}^{|\mathcal{X}|}$, $\text{FDR}(h)$ is the proportion of examples that h incorrectly labels as 1 out of all examples h labels as 1. Our goal is to solve the following constrained best arm identification problem:

$$\widehat{h} = \min_{h\in\mathcal{H}} R(h) \quad \text{s.t.} \quad \text{FDR}(h) \leq q \iff \min_{h\in\mathcal{H}} z_h^\top \eta \quad \text{s.t.} \quad ((\mathbf{1} - \eta)^\top - q\mathbf{1}^\top)^\top z \leq 0. \tag{3.1}$$

The main challenge in running BESIDE on this problem directly is a potentially high computational cost from computing a design over an extremely large hypothesis class $\mathcal{H}$ (e.g. neural networks of a bounded width). In this section we provide an alternative approach motivated by BESIDE. Algorithm 2 follows a similar design as BESIDE and relies on an oracle, `CERM`, that can solve (3.1), i.e. given a dataset it returns the highest accuracy classifier under an FDR constraint. Such oracles are available in, for example in Agarwal et al. [2018], Cotter et al. [2018]. In each round of Algorithm 2 we perform *randomized exploration* by perturbing the labels on our existing dataset with mean zero Gaussian noise, and then training k classifiers $\widehat{h}_i, i \in [k]$, on the resulting datasets. Implicitly, we are making the assumption that the loss function in the training of `ERM` can handle continuous labels, such as the MLE of logistic regression. As described in Kveton et al. [2019], randomized exploration emulates sampling from a posterior distribution on our possible set of classifiers. We then use the labels generated from these classifiers to compute safe classifiers $h_i, i \in [k]$. Finally, mimicking the strategy of BESIDE, we compute $\mathcal{XY}_{\text{safe}}$ and $\mathcal{XY}_{\text{diff}}$ designs on these k safe classifiers and repeat (note that the designs computed on Line 5 are equivalent to $\mathcal{XY}_{\text{safe}}$ and $\mathcal{XY}_{\text{diff}}$ in the classification setting).

To validate Algorithm 2, we experiment against a passive baseline that selects points uniformly at randoms from the pool of examples $\mathcal{X}$, retrains the model using the same Constrained Empirical Risk Minimization oracle (`CERM`) as Algorithm 2 on its current samples, and report the accuracy and FDR. We evaluate on two real world datasets next and provide an additional experiment on a

Algorithm 2 `Active constrained classification with randomized exploration`

Require: Batch size n, initial (labeled) data $x_1^{(0)}, \dots, x_n^{(0)}$, number of rounds L, number of classifiers per round k, perturbation variance σ

1: **for** $\ell = 1, \dots, L$ **do**
2: **for** $i = 1, \dots, k$ **do**
3: $\widehat{h}_i = \texttt{ERM}(\{(x_t^{(\ell)}, y_t^{(\ell)} + \epsilon_t^{(i)})\}_{t=1}^n)$, where $\{\epsilon_t^{(i)}\}_{1 \le t \le n} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
4: $h_i = \texttt{CERM}(\{(x, \widehat{h}_i(x))\}_{x \in \mathcal{X}})$
5: Compute designs: $\lambda_{\text{safe}} = \arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{1 \le i \le k} \sum_{x \in \mathcal{X}} \frac{\mathbb{1}\{h_i(x) \neq 0\}}{\lambda_x}$, $\lambda_{\text{diff}} = \arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{1 \le i \ne j \le k} \sum_{x \in \mathcal{X}} \frac{\mathbb{1}\{h_i(x) \neq h_j(x)\}}{\lambda_x}$
6: Sample $x_1^{(\ell)}, \dots, x_n^{(\ell)}$ from a uniform mixture of $\lambda_{\text{safe}}, \lambda_{\text{diff}}$
7: Observe corresponding labels $y_1^{(\ell)}, \dots, y_n^{(\ell)}$
return $\widetilde{h} = \texttt{CERM}(\{(x_t^{(\ell)}, y_t^{(\ell)})\}_{1 \le t \le n, 0 \le \ell \le L})$

synthetic dataset in Section F.

Adult dataset. We evaluate on the adult income data set Lichman [2013] (48,842 examples) where the goal is to predict whether someone's income is above \$50k per year. We report in Figure 5 the accuracy and the FDR obtained when varying the number of labels given to each method. We observe that for any desired accuracy Algorithm 2 allows us to provide a classifier with lower FDR. Also, for any chosen number of total labels—such as $500, 750, 2000$ as reported in Figure 5—the Algorithm 2 gives a classifier with higher accuracy and lower FDR. In general we found that the active method needed half the number of samples as the passive sampling to achieve a given FDR. This demonstrates the effectiveness of Algorithm 2 to learn simultaneously the objective (risk) and the constraint (FDR), in a similar favorable way as characterized by our theoretical findings.

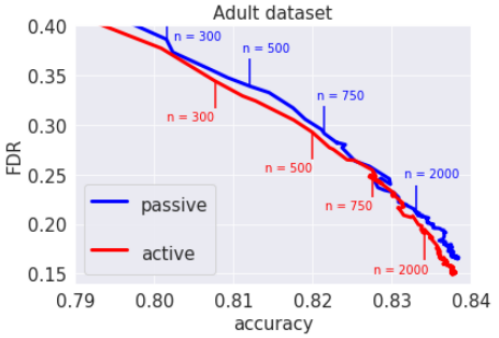

Figure 5: FDR vs accuracy for active (Algorithm 2) and passive sampling, ticks report number of samples. FDR and accuracy are averaged over 5 trials

German Credit dataset. We consider the German Credit Dataset originally from the Staflog Project Databases Keogh et al. [1998]. The goal is to predict whether someone's credit is 'bad' or 'good'. We report in Figure 6 the recall (TPR) and the precision $(1 - \text{FDR})$ obtained when varying the number of labels given to each method. We observe that for any desired precision Algorithm 2 allows us to provide a classifier with higher recall. Also, for any chosen number of total labels—such as $170, 270, 330, 450, 600$ as reported in Figure 6—the Algorithm 2 gives a classifier with higher precision and higher recall. As for the Adult dataset we found that the active method needed half the number of samples as the passive sampling to achieve a given precision.

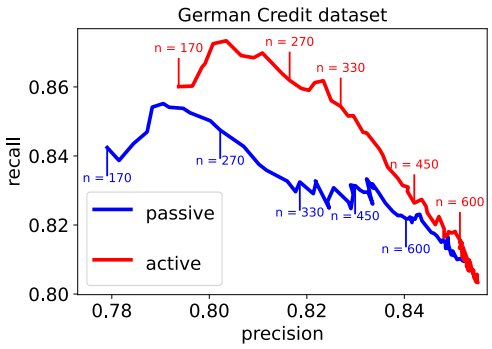

Figure 6: TPR vs FDR for active active (Algorithm 2) and passive sampling, ticks report number of samples. Precision is $1 - \text{FDR}$, recall is TPR. Precision and recall are averaged over 25 trials

4 Related works

Constrained Bandits. A growing body of work seeks to address the question of safe learning in interactive environments. In particular, the majority of such works have considered the problem of regret minimization in linear bandits with linear safety constraints. Here, the goal is to maximize online reward, $x_t^\top \theta_*$, by choosing actions $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$, while ensuring a safety constraint of the form $x_t^\top \mu_* \leq \gamma$ is met at all times (either in expectation or with high probability). A variety of algorithms have been proposed, including UCB-style [Kazerouni et al., 2017, Amani et al., 2019, Pacchiano et al., 2020], and Thompson Sampling [Moradipari et al., 2019, 2020]. While these works show that $\sqrt{T}$ regret is attainable, they only provide worst-case bounds (while we obtain instance-dependent bounds) and do not study the pure-exploration best-arm identification problem.

To our knowledge, only several existing works consider the question of best-arm identification with safety constraints [Sui et al., 2015, 2018, Wang et al., 2022]. [Sui et al., 2015, 2018] consider a general constrained optimization setting where the goal of the learner is to minimize some function $f(x)$ over a domain $x \in \mathcal{D}$, while only having access to noisy samples of $f(x)$, $f(x_t) + w_t$, and guaranteeing that a safety constraint $g(x_t) \geq h$ is met for every query point x_t. While they do provide a sample complexity upper bound, they give no lower bound, and, as shown in [Wang et al., 2022], their approach can be very suboptimal. [Wang et al., 2022] considers the setting of best-arm identification in multi-armed bandits. In their setting, at every step t they query a value $a_t \in \mathcal{A}$ for a particular coordinate i_t, and their goal is to identify the coordinate i^* such that $a_{i^*}^* \theta_{i^*} \geq \max_i a_i^* \theta_i$, where a_i^* is the largest value respecting the safety constraint: $a_i^* = \arg\max_{a \in \mathcal{A}} a\theta_i$ s.t. $a\mu_i \leq \gamma$. Similar to [Sui et al., 2015, 2018], they require that the safety constraint $a_t \mu_{i_t} \leq \gamma$ must be met while learning. Though they do show matching upper and lower bounds, and in addition consider a slightly more general setting that allows for nonlinear (but monotonic) response functions, they treat every coordinate as independent, and do not allow for information-sharing between coordinates—the key generalization the linear bandit setting targets. We remark as well that in our setting, unlike these works, we allow the learner to query unsafe points during exploration, and only require that they output a safe decision at termination.

Best-Arm Identification in Linear Bandits. The best-arm identification problem in multi-armed bandits (without safety constraints) is a classical and well-studied problem [Bechhofer, 1958, Paulson, 1964, Even-Dar et al., 2002, Bubeck et al., 2009], and near-optimal algorithms exist [Jamieson et al., 2014, Kaufmann et al., 2016]. More recently, there has been a growing interest in understanding the sample complexity of best-arm identification in linear bandits [Soare et al., 2014, Karnin, 2016, Xu et al., 2018, Fiez et al., 2019, Katz-Samuels et al., 2020, Degenne et al., 2020]. We highlight in particular the work of Fiez et al. [2019] which proposes an experiment-design based algorithm, RAGE, that our approach takes inspiration from. While much progress has been made in understanding best-arm identification in linear bandits, to our knowledge, no existing works consider the setting of best-arm identification in linear bandits with safety constraints, the setting of this work.

Active Classification under FDR constraints We finally mention one other related body of work—the problem of actively sampling to find a classifier with high accuracy or recall under precision constraints. Motivated by the experimental design approach of our main algorithm, BESIDE, we provide a heuristic algorithm for this problem with good empirical performance in Section 3.1. There is an extensive body of work on active learning (see the survey Hanneke [2014]) but only recently have works made the connection between best-arm identification for linear bandits and

classification Katz-Samuels et al. [2021], Jain and Jamieson [2020], Camilleri et al. [2021b]. Precision constraints has been less studied in the adaptive context, we only know of Jain and Jamieson [2020], Bennett et al. [2017].

5 Conclusion

In this work we have shown that it is possible to efficiently find the best *safe* arm in linear bandits with a carefully designed adaptive experiment design-based approach. Our results open up several interesting directions for future work.

Instance Optimality. While BESIDE is worst-case optimal, in Appendix A we show an instance-dependent lower bound which BESIDE does not, in general, seem to hit. We conjecture that this lower bound may be loose—addressing this discrepancy and showing matching instance-dependent upper and lower bounds is an exciting direction for future work.

Safety During Exploration. Though there are many interesting applications where we may not require safety during exploration (i.e. only querying safe arms), in other cases we may need to ensure safety is met during exploration. Extending our work to this setting is an interesting open problem.

Potential Impacts. As with any algorithm making stochastic assumptions, if assumptions are not met we can not guarantee the performance. In this case, one limitation is that if the underlying environment is changing (i.e. the constraints vary over time) the algorithm could have unexpected behavior with unintended consequences. Such a situation could lead to harmful results in examples such as the online advertising bidding example from the introduction. To mitigate this limitation of our setting, practitioners are encouraged to monitor many metrics, both short and long-term.

Acknowledgements

The work of AW was supported by an NSF GFRP Fellowship DGE-1762114. The work of JM was supported by an NSF Career award, and NSF AI institute (IFML) and the Simons collaborative grant on the foundations of fairness. The work of KJ was funded in part by the AFRL and NSF TRIPODS 2023166.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification, 2018.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Edward A. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9252–9262, 2019. URL http://papers.nips.cc/paper/9124-linear-stochastic-bandits-under-safety-constraints.

Robert E Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

Paul N Bennett, David M Chickering, Christopher Meek, and Xiaojin Zhu. Algorithms for active classifier selection: Maximizing recall with precision constraints. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 711–719, 2017.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Romain Camilleri, Julian Katz-Samuels, and Kevin Jamieson. High-dimensional experimental design and kernel bandits, 2021a.

Romain Camilleri, Zhihan Xiong, Maryam Fazel, Lalit Jain, and Kevin Jamieson. Selective sampling for online best-arm identification, 2021b.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints, 2018. URL https://arxiv.org/abs/1807.00028.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits, 2019.

Steve Hanneke. Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3), 2014.

Elad Hazan and Satyen Kale. Projection-free online learning. *arXiv preprint arXiv:1206.4657*, 2012.

Lalit Jain and Kevin Jamieson. A new perspective on pool-based active classification and false-discovery control, 2020.

Kevin Jamieson and Lalit Jain. Interactive machine learning. 2022.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Zohar S Karnin. Verification based solution for structured mab problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

Julian Katz-Samuels, Jifan Zhang, Lalit Jain, and Kevin Jamieson. Improved algorithms for agnostic pool-based active classification, 2021. URL https://arxiv.org/abs/2105.06499.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models, 2016.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

E. Keogh, C.; Blake, and C. J. Merz. Uci repository of machine learning databases,, 1998. URL http://archive.ics.uci.edu/ml.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits, 2019. URL https://arxiv.org/abs/1906.08947.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

M Lichman. Uci machine learning repository,, 2013. URL http://archive.ics.uci.edu/ml.

Blake Mason, Romain Camilleri, Subhojyoti Mukherjee, Kevin Jamieson, Robert Nowak, and Lalit Jain. Nearly optimal algorithms for level set estimation. *arXiv preprint arXiv:2111.01768*, 2021.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information, 2019. URL https://arxiv.org/abs/1911.02156.

Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. 2020. doi: 10.48550/ARXIV.2010.00081. URL https://arxiv.org/abs/2010.00081.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints, 2020. URL https://arxiv.org/abs/2006.10185.

Edward Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, pages 174–180, 1964.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834. PMLR, 2017.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27:828–836, 2014.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, pages 997–1005. PMLR, 2015.

Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4781–4789. PMLR, 2018.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. *arXiv preprint arXiv:2201.11206*, 2022.

Zhenlin Wang, Andrew Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

Contents

A Lower Bounds

A.1 Oracle Lower Bound

Theorem 2 (Oracle Lower Bound). *Let τ denote the stopping time for any $(0,\delta)$-PAC algorithm for pure exploration in safe linear bandits. Then*

$$\frac{\mathbb{E}_{\theta_*,\mu_*}[\tau]}{\log\frac{1}{2.4\delta}} \geq \min_{\lambda\in\triangle_{\mathcal{X}}} \max\left\{ \max_{z\in\mathcal{Z}\setminus z_*} \min\left\{ \frac{\|z\|^2_{A(\lambda)^{-1}}}{\mathfrak{p}(-\Delta_{\text{safe}}(z))^2}, \frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{\mathfrak{p}(\Delta(z))^2} \right\}, \frac{\|z_*\|^2_{A(\lambda)^{-1}}}{(z_*^\top\mu_* - \alpha)^2} \right\}.$$

Comparing Complexity with Theorem 1. In the single-constraint setting, the complexity of BESIDE reduces to

$$C\cdot\sup_{\widetilde{\epsilon}\geq\epsilon}\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{m|\mathcal{Z}|}{\delta})}{\left(|\Delta_{\text{safe}}(z)|+\mathfrak{p}(\Delta^{\widetilde{\epsilon}}(z))+\widetilde{\epsilon}\right)^2} + C\cdot\sup_{\widetilde{\epsilon}\geq\epsilon}\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{|\mathcal{Z}|}{\delta})}{\left(\mathfrak{p}(-\Delta_{\text{safe}}(z))+\mathfrak{p}(\Delta^{\widetilde{\epsilon}}(z))+\widetilde{\epsilon}\right)^2} + C_0$$

Consider the case when $\Delta^{\widetilde{\epsilon}}(z)$ is "smooth" in $\widetilde{\epsilon}$, in the sense that $\Delta^{\widetilde{\epsilon}}(z)\geq\Delta(z)-\widetilde{\epsilon}$. This condition corresponds to the case, for example, where z_* has a large safety gap (in which case we simply have $\Delta^{\widetilde{\epsilon}}(z)=\Delta(z)$ for moderate values of $\widetilde{\epsilon}$), or where z_* might have a small safety gap, but where there are arms placed at even intervals so that, as we let the safety gap get smaller, we are always able to find better arms. Under this assumption, the complexity can be upper bounded as

$$\begin{aligned}
&C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{m|\mathcal{Z}|}{\delta})}{\left(|\Delta_{\text{safe}}(z)|+\mathfrak{p}(\Delta(z))\right)^2} + C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{|\mathcal{Z}|}{\delta})}{\left(\mathfrak{p}(-\Delta_{\text{safe}}(z))+\mathfrak{p}(\Delta(z))\right)^2} + C_0\\
&\leq C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{m|\mathcal{Z}|}{\delta})}{\left(|\Delta_{\text{safe}}(z)|+\mathfrak{p}(\Delta(z))\right)^2} + C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{|\mathcal{Z}|}{\delta})}{\left(\mathfrak{p}(-\Delta_{\text{safe}}(z))+\mathfrak{p}(\Delta(z))\right)^2} + C_0
\end{aligned}$$

which can be upper bounded as

$$C\log(\tfrac{m|\mathcal{Z}|}{\delta})\cdot\left(\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}}{\max\{\Delta_{\text{safe}}(z)^2,\mathfrak{p}(\Delta(z))^2\}} + \inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{\max\{\mathfrak{p}(-\Delta_{\text{safe}}(z))^2,\mathfrak{p}(\Delta(z))^2\}}\right) + C_0.$$

While this does not match the lower bound of Theorem 2 exactly, it scales in a similar manner. As in Theorem 2, we pay only for the larger of the optimality gap, $\mathfrak{p}(\Delta(z))$, and safety gap $\mathfrak{p}(-\Delta_{\text{safe}}(z))$ (if the arm is unsafe). The primary difference between Theorem 2 and this complexity are the terms in the numerator—in Theorem 2, the numerator scales as $\|z-z_*\|^2_{A(\lambda)^{-1}}$ only if an arm is easier to eliminate by showing it is suboptimal, while in our complexity it could scale this way in either case.

The primary difficulty in hitting the lower bound exactly is that Theorem 2 is a *verification* lower bound. It assumes knowledge of the best arm, and is told whether every other arm has smaller safety gap (if the arm is unsafe) or optimality gap. It can therefore simply use this knowledge to focus all samples on verifying an arm is either unsafe, or suboptimal.

In practice, we do not have access to such information. Without knowing whether it is easier to eliminate an arm by showing it is unsafe or suboptimal, the best we can hope to do is to seek to estimate both the safety value and reward value of every arm, until we have estimated one well enough to show the arm is suboptimal or unsafe.

We conjecture that the lower bound of Theorem 2 is loose, and that Theorem 1 is nearly optimal. We believe the gap arises because, as noted, lower bound proof techniques, such as those proposed in Kaufmann et al. [2016], which is what we rely on to prove Theorem 2, are lower bounding only the complexity of verifying the optimal solution. In problem settings such as ours where the *order* matters—where we will obtain a very different rate if we focus our attention on one arm versus another, to show it is safe or unsafe—such techniques appear insufficient to obtain a tight lower bound. Indeed, we conjecture that a "moderate-confidence" lower bound can be shown using techniques from Simchowitz et al. [2017], and that such a lower bound may have a complexity nearly matching that of Theorem 1. We leave proving this for future work.

Proof of Theorem 2. Following the proof of Theorem 1 of Fiez et al. [2019] and applying the Transportation Lemma of Kaufmann et al. [2016], we have that any δ-PAC algorithm must satisfy

$$\sum_{x\in\mathcal{X}} \mathbb{E}[T_x] \geq \log \frac{1}{2.4\delta} \cdot \inf_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{\min_{(\theta,\mu)\in\mathcal{C}_{\text{alt}}} \sum_{x\in\mathcal{X}} \lambda_x \text{KL}(\nu_{(\theta_*,\mu_*),i}||\nu_{(\theta,\mu),i})}$$

there T_x denotes the number of pulls to arm x, and $\mathcal{C}_{\text{alt}}$ is the set of alternate instances defined in Lemma A.1. As we assume that the noise is $\mathcal{N}(0,1)$, and since the noise is independent for the safety observations and reward observations, we have

$$\text{KL}(\nu_{(\theta_*,\mu_*),i}||\nu_{(\theta,\mu),i}) = \frac{1}{2}(x_i^\top(\theta_* - \theta))^2 + \frac{1}{2}(x_i^\top(\mu_* - \mu))^2.$$

Some algebra shows that

$$\sum_{x\in\mathcal{X}} \lambda_x \text{KL}(\nu_{(\theta_*,\mu_*),i}||\nu_{(\theta,\mu),i}) = \frac{1}{2}\|\theta_* - \theta\|^2_{A(\lambda)} + \frac{1}{2}\|\mu_* - \mu\|^2_{A(\lambda)}.$$

The result then follows by applying Lemma A.1 to compute

$$\min_{(\theta,\mu)\in\mathcal{C}_{\text{alt}}} \frac{1}{2}\|\theta_* - \theta\|^2_{A(\lambda)} + \frac{1}{2}\|\mu_* - \mu\|^2_{A(\lambda)}.$$

□

Lemma A.1. *Define the alternate set:*

$$\mathcal{C}_{\text{alt}} = \{(\theta,\mu) \text{ s.t. } \mu^\top z^* > \alpha\} \cup \{(\theta,\mu) \text{ s.t. } \exists z' \neq z^*, \mu^\top z' \leq \alpha, \theta^\top(z^* - z') \leq 0\},$$

Then the projection to the alternate is

$$\min_{(\theta,\mu)\in\mathcal{C}_{\text{alt}}} \|\theta - \theta_*\|^2_{A(\lambda)} + \|\mu - \mu_*\|^2_{A(\lambda)} = \min\left\{ \min_{z\neq z^*} \frac{\mathfrak{p}(z^\top\mu_* - \alpha)^2}{\|z\|^2_{A(\lambda)^{-1}}} + \frac{\mathfrak{p}((z_* - z)^\top\theta_*)^2}{\|z - z_*\|^2_{A(\lambda)^{-1}}}, \frac{(z_*^\top\mu_* - \alpha)^2}{\|z_*\|^2_{A(\lambda)^{-1}}} \right\}.$$

Proof. For each arm x the associated and we want to solve

$$\min_{(\theta,\mu)\in\mathcal{C}_{\text{alt}}} \|\theta - \theta_*\|^2_{\sum_{x\in\mathcal{X}} \lambda_x xx^\top} + \|\mu - \mu_*\|^2_{\sum_{x\in\mathcal{X}} \lambda_x xx^\top}.$$

To do so, we use that $\min_{x\in A\cup B} f(x) = \min_{S\in\{A,B\}} \min_{x\in S} f(x)$ on the quadratic objective by defining the sets

$$A := \{(\theta,\mu) \text{ s.t. } \mu^\top z^* > \alpha\}\ ,\ B = \{(\theta,\mu) \text{ s.t. } \exists z' \neq z^*, \mu^\top z' \leq \alpha, \theta^\top(z^* - z') \leq 0\},$$

such that their union is $A \cup B = \mathcal{C}_{\text{alt}}$.

Note that we know from Mason et al. [2021] that

$$\min_{(\theta,\mu)\in A} \|\theta-\theta_*\|^2_{\sum_{x\in\mathcal{X}}\lambda_x xx^\top} + \|\mu-\mu_*\|^2_{\sum_{x\in\mathcal{X}}\lambda_x xx^\top} = \frac{(z_*^\top\mu_*-\alpha)^2}{\|z_*\|^2_{A(\lambda)^{-1}}}.$$

We now lift B to a set $\text{lift}(B)$ that is defined as

$$\text{lift}(B) = \{[\theta,\mu] \text{ s.t. } \exists z' \neq z^*, [\theta,\mu]^\top[(z^*-z'),0;0,z'] \leq [0,\alpha]\}.$$

Thus we can focus on $D_z = \{\kappa \in \mathbb{R}^{2n} \text{ s.t. } A_z\kappa \leq b\}$ where $A_z = [(z^*-z'),0;0,z'] \in R^{2\times 2n}$. Now we want to solve

$$\min_{z\in\mathcal{Z}\setminus\{z_*\}} \min_{\kappa\in\mathbb{R}^{2n} \text{ s.t. } A_z\kappa\leq b} \|\kappa-\kappa_*\|_\Gamma,$$

where $\Gamma = I_2 \bigotimes \left(\sum_{x\in\mathcal{X}}\lambda_x xx^\top\right)$.

Lemma A.2. *The optimal solution of*

$$\min_{\kappa\in\mathbb{R}^{2n} \text{ s.t. } A\kappa\leq b} \frac{\|\kappa-\kappa_*\|_\Gamma}{2}$$

is $\kappa_0 = \kappa_* - \Gamma^{-1}A^\top(A\Gamma^{-1}A^\top)^{-1}\{A\kappa_*-b\}_+$ *and the optimal value is*

$$\frac{1}{2}\mathfrak{p}(A\kappa_*-b)^\top(A\Gamma^{-1}A^\top)^{-1}\mathfrak{p}(A\kappa_*-b),$$

where $\mathfrak{p}(\cdot)$ *is applied element-wise to* $A\kappa_*-b$.

This translate to

$$\min_{(\theta,\mu)\in B} \|\theta-\theta_*\|^2_{\sum_{x\in\mathcal{X}}\lambda_x xx^\top} + \|\mu-\mu_*\|^2_{\sum_{x\in\mathcal{X}}\lambda_x xx^\top} = \min_{z\neq z^*} \frac{\mathfrak{p}(z^\top\mu_*-\alpha)^2}{\|z\|^2_{A(\lambda)^{-1}}} + \frac{\mathfrak{p}((z_*-z)^\top\theta_*)^2}{\|z-z_*\|^2_{A(\lambda)^{-1}}},$$

and we get the desired result.

□

Proof of Lemma A.2. Consider the Lagrangian

$$\mathcal{L}(\kappa,\mu) = \frac{1}{2}(\kappa-\kappa_*)^\top\Gamma(\kappa-\kappa_*) + \mu^\top(A\kappa-b)$$
$$\mathcal{L}(\kappa',\mu) = \frac{1}{2}\kappa'^\top\Gamma\kappa' + \mu^\top(A\kappa'-b+A\kappa_*)$$

minimized at $\kappa_0' = -\Gamma^{-1}A^\top\mu$. We have

$$\begin{aligned}\max_{\mu\geq 0}\min_{\kappa'} \mathcal{L}(\kappa',\mu) &= \max_{\mu\geq 0} \frac{1}{2}(\Gamma^{-1}A^\top\mu)^\top\Gamma(\Gamma^{-1}A^\top\mu) + \mu^\top(-A\Gamma^{-1}A^\top\mu - b + A\kappa_*) \\ &= \max_{\mu\geq 0} -\frac{1}{2}\mu^\top A\Gamma^{-1}A^\top\mu + \mu^\top(A\kappa_*-b)\end{aligned}$$

maximized at $\mu_0 = (A\Gamma^{-1}A\top)^{-1}\{A\kappa_*-b\}_+$ where $\{[b_1,b_2]\}_+ = [\max\{b_1,0\},\max\{b_2,0\}]$. Plugging μ_0 back in the solution κ_0', we get the solution κ_0

$$\kappa_0 = \kappa_* - \Gamma^{-1}A^\top(A\Gamma^{-1}A^\top)^{-1}\{A\kappa_*-b\}_+$$

and the optimal value follows. □

A.2 Proof of Proposition 1

Proof of Proposition 1.

Proof for $\mathcal{I}^1$. Fix $\alpha \in (0, 0.1)$ and consider the following instance with $m = 1$:

$$\mathcal{X} = \{e_1, e_2\}, \quad \mathcal{Z} = \{z_1, z_2\}, \quad z_1 = [1/4, 1/2], \quad z_2 = [3/4, 1/2 + \alpha]$$
$$\theta_* = e_1, \quad \mu_* = [0, 1], \quad \gamma = 1/2 + \alpha/2.$$

On this example, z_1 is safe and z_2 is unsafe with $\Delta_{\text{safe}}(z_2) = -\alpha/2$.

Let $A(\lambda) = \lambda_1 e_1 e_1^\top + \lambda_2 e_2 e_2^\top$ denote the design matrix. Then the allocation that minimizes $\mathcal{XY}_{\text{diff}}$:

$$\max_{z,z' \in \mathcal{Z}} \|z - z'\|^2_{A(\lambda)^{-1}} = \frac{1}{4\lambda_1} + \frac{\alpha^2}{\lambda_2}$$

is

$$\lambda_1 = \frac{1}{1 + 2\alpha}, \quad \lambda_2 = \frac{2\alpha}{1 + 2\alpha}.$$

Denote this allocation as $\widetilde{\lambda}$.

Applying the Transportation Lemma of Kaufmann et al. [2016], this implies that any δ-PAC strategy must have

$$\mathbb{E}[T_1]\text{KL}(\nu_{(\theta_*,\mu_*),1}||\nu_{(\theta,\mu),1}) + \mathbb{E}[T_2]\text{KL}(\nu_{(\theta_*,\mu_*),2}||\nu_{(\theta,\mu),2}) \geq \log \frac{1}{2.4\delta}$$

for all $(\theta, \mu) \in \mathcal{C}_{\text{alt}}$, where $\mathcal{C}_{\text{alt}}$ is defined as in Lemma A.1. If a learner plays $\widetilde{\lambda}$ for T steps, they will have $\mathbb{E}[T_1] = \lambda_1 \mathbb{E}[T], \mathbb{E}[T_2] = \lambda_2 T$. In this case, the above can be rewritten as

$$\begin{aligned}
\mathbb{E}[T] &\geq \log \frac{1}{2.4\delta} \cdot \frac{1}{\lambda_1 \text{KL}(\nu_{(\theta_*,\mu_*),1}||\nu_{(\theta,\mu),1}) + \lambda_2 \text{KL}(\nu_{(\theta_*,\mu_*),2}||\nu_{(\theta,\mu),2})} \\
&= \log \frac{1}{2.4\delta} \cdot \frac{2}{\|\theta_* - \theta\|^2_{A(\widetilde{\lambda})} + \|\mu_* - \mu\|^2_{A(\widetilde{\lambda})}}.
\end{aligned}$$

where the equality follows by the same calculation as in the proof of Theorem 2. Take (θ, μ) to be $\theta = \theta_*$, $\mu = [0, 1 - \frac{\alpha}{1+2\alpha}]$ and note that $(\theta, \mu) \in \mathcal{C}_{\text{alt}}$ since with this choice of μ, arm z_2 is now safe. Now,

$$\|\mu_* - \mu\|^2_{A(\widetilde{\lambda})} = (\frac{\alpha}{1 + 2\alpha})^2 \cdot \frac{2\alpha}{1 + 2\alpha} = \frac{2\alpha^3}{(1 + 2\alpha)^3}.$$

This gives a lower bound of

$$\mathbb{E}[T] \geq \log \frac{1}{2.4\delta} \cdot \frac{2(1 + 2\alpha)^3}{2\alpha^3} \geq \log \frac{1}{2.4\delta} \cdot \frac{1}{\alpha^4}.$$

This lower bound is for best-arm identification ($\epsilon = 0$), but setting $\alpha \leftarrow 2\epsilon - a$ for a arbitrarily small, identifying an ϵ-optimal, ϵ-safe arm is equivalent to identifying the best arm, so this therefore holds as a lower bound on (ϵ, δ)-PAC algorithms.

The upper bound on the performance of BESIDE follows trivially since by setting $\lambda_1 = \lambda_2$, we can make the numerator in both terms of the complexity $\mathcal{O}(1)$, and the denominator of each term will be at least ϵ^2.

Proof for $\mathcal{I}^2$. Fix $\alpha \in (0, 0.1)$ and consider the following instance with $m = 1$:

$$\mathcal{X} = \{e_1, e_2\}, \quad \mathcal{Z} = \{z_1, z_2\}, \quad z_1 = [1/2 + \alpha^2/2, 0], \quad z_2 = [1/2, \alpha/2]$$
$$\theta_* = [1/2, 0], \quad \mu_* = [0, 0], \quad \gamma = 1.$$

On this instance, both z_1 and z_2 are safe, and z_1 is optimal.

The $\mathcal{XY}_{\text{safe}}$ design minimizes:

$$\max_{z \in \mathcal{Z}} \|z\|^2_{A(\lambda)^{-1}} = \max\left\{\frac{1 + 2\alpha + \alpha^2}{4\lambda_1}, \frac{1}{4\lambda_1} + \frac{\alpha^2}{4\lambda_2}\right\}.$$

Some computation shows that, for α small, the optimal settings are $\lambda_1 = \mathcal{O}(1)$ and $\lambda_2 = \mathcal{O}(\alpha)$ (where here $\mathcal{O}(\cdot)$ hides terms that are $o(\alpha)$). Denote this allocation as $\widetilde{\lambda}$. Following the same argument as above, we have

$$\mathbb{E}[T] \geq \log \frac{1}{2.4\delta} \cdot \frac{2}{\|\theta_* - \theta\|^2_{A(\widetilde{\lambda})} + \|\mu_* - \mu\|^2_{A(\widetilde{\lambda})}}$$

for any $(\mu, \theta) \in \mathcal{C}_{\text{alt}}$. Let $\theta = [1/2, 2\alpha]$ and note that $(\mu_*, \theta) \in \mathcal{C}_{\text{alt}}$ since z_2 is now the optimal arm with this θ. We then have

$$\|\theta_* - \theta\|^2_{A(\widetilde{\lambda})} + \|\mu_* - \mu\|^2_{A(\widetilde{\lambda})} = \mathcal{O}(\alpha^3)$$

which gives a lower bound of

$$\mathbb{E}[T] \geq \Omega\left(\frac{1}{\alpha^3} \cdot \log \frac{1}{\delta}\right).$$

This lower bound holds for the best-arm identification problem, but setting $\alpha \leftarrow \sqrt{2\epsilon} - a$ for a arbitrarily small, finding an ϵ-optimal arm is equivalent to finding the best arm, so the lower bound applies in that setting as well.

To compute the sample complexity of BESIDE, we note that $\Delta_{\text{safe}}(z_1) = \Delta_{\text{safe}}(z_2) = 1$, so the first term in the complexity is negligible. We also have that $\Delta^{\widetilde{\epsilon}}(z_2) = \alpha^2/2 = \mathcal{O}(\epsilon)$ for $\widetilde{\epsilon} \leq 1$. Thus, the second term in the complexity scales as

$$\begin{aligned}
\widetilde{\mathcal{O}}\left(\inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z, z' \in \mathcal{Z}} \frac{\|z - z'\|^2_{A(\lambda)^{-1}} \cdot \log 1/\delta}{\epsilon^2}\right) &= \widetilde{\mathcal{O}}\left(\inf_{\lambda \in \triangle_{\mathcal{X}}} \frac{(\alpha^4/\lambda_1 + \alpha^2/\lambda_2) \cdot \log 1/\delta}{\epsilon^2}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\alpha^2 \cdot \log 1/\delta}{\epsilon^2}\right) \\
&= \widetilde{\mathcal{O}}\left(\frac{\log 1/\delta}{\epsilon}\right).
\end{aligned}$$

□

B Robust Mean Estimation

In order to form estimates of $z^\top \theta_*$ and $z^\top \mu_{*,i}$, we will rely on the RIPS procedure proposed in Camilleri et al. [2021a], instantiated with the robust Catoni estimator Catoni [2012].

Catoni Estimation. The robust Catoni mean estimator proposed in Catoni [2012] is defined as follows.

Definition B.1 (Catoni Estimator)**.** Consider real values $X_1, \ldots, X_T$. Then the *robust Catoni mean estimator*, $\mathsf{cat}_\alpha[\{X_t\}_{t=1}^T]$, with parameter $\alpha > 0$ is the unique root z of the function

$$f_{\mathsf{cat}}(z; \{X_i\}_{i=1}^T, \alpha) := \sum_{t=1}^T \psi_{\mathsf{cat}}(\alpha(X_t - z)) \quad \text{for} \quad \psi_{\mathsf{cat}}(y) := \begin{cases} \log(1 + y + y^2) & y \geq 0 \\ \log(1 - y + y^2) & y < 0 \end{cases}.$$

The Catoni estimator satisfies the following guarantee.

Proposition 2. *Let $X_1, \ldots, X_T$ be independent, identically distributed random variables with mean ζ and variance $\sigma^2 < \infty$. Fix $\delta \in (0, 1)$ and assume $T \geq 4 \log(1/\delta)$. Then the Catoni estimator $\mathsf{cat}_\alpha[\{X_t\}_{t=1}^T]$ with parameter*

$$\alpha = \sqrt{\frac{2 \log 1/\delta}{T\sigma^2}} \tag{B.1}$$

satisfies, with probability at least $1 - 2\delta$,

$$|\mathsf{cat}_\alpha[\{X_t\}_{t=1}^T] - \zeta| \leq \sqrt{\frac{8\sigma^2 \log 1/\delta}{T}}.$$

Notably, the estimation error given by Proposition 2 scales only with the variance of the random variables, and not with their magnitude.

Robust Inverse Propensity Score (RIPS) Estimator. We apply the Catoni estimator with the RIPS estimator of Camilleri et al. [2021a]. In particular, consider running the following procedure.

Algorithm 3 Robust Inverse Propensity Score Estimation (RIPS)

1: **input:** samples $\{(x_t, r_t)\}_{t=1}^T$ for $x_t \sim \lambda$ and $r_t = \theta^\top x_t + w_t$, active set $\mathcal{Y}$, confidence δ

2: For each $y \in \mathcal{Y}$, set $W^y \leftarrow \mathsf{cat}_\alpha[\{y^\top A(\lambda)^{-1} x_t r_t\}_{t=1}^T]$, for α chosen as in (B.1) with $\delta \leftarrow \frac{\delta}{2|\mathcal{Y}|}$, and $A(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^\top$.

3: Set

$$\widehat{\theta} = \arg\min_\theta \max_{y \in \mathcal{Y}} \frac{|\theta^\top y - W^y|}{\|y\|_{A(\lambda)^{-1}}}.$$

4: **return** $\widehat{\theta}$

We have the following guarantee on this procedure.

Proposition 3 (Theorem 1 of Camilleri et al. [2021a])**.** *If $T \geq 4 \log \frac{2|\mathcal{Z}|}{\delta}$, then with probability at least $1 - \delta$, for all $z \in \mathcal{Z}$, the RIPS estimator of Algorithm 3 returns an estimate $\widehat{\theta}$ which satisfies:*

$$|y^\top(\widehat{\theta} - \theta_*)| \leq \|y\|_{A(\lambda)^{-1}} \cdot \sqrt{\frac{8 \log(2|\mathcal{Z}|/\delta)}{T}}.$$

The use of the RIPS estimator allows us to avoid sophisticated rounding procedures often found in the linear bandit literature. Note that the RIPS estimator can be computed in time scaling polynomially in $|\mathcal{Y}|$, d, and T.

C RAGE$^\epsilon$

A note on constants. Throughout our algorithm definitions, in both this section and the following, we use generic constants rather than precise numerical settings, and carry these generic constants through our proofs. At various points in the proofs, we require that these constants satisfy certain constraints. The following result shows that there exist suitable settings for all constants such that these constraints are satisfied.

Lemma C.1. *There exist settings of $c_a, c_b, c_d, c_e, c_f, c_e, c_g, c_1, c_2, c_3, c_4, c_\Delta$ and c_0 such that Equations (D.2), (D.5), (D.6), (D.7), (D.8), (D.9), (D.10), (C.1), (C.2), and (C.3) are satisfied, and*

$$\frac{c_3(1+c_g)}{1-c_3} \leq 0.2, \quad c_g \leq 0.2, \quad c_0 \geq 0.0001.$$

Proof. First, note that in addition to the conditions listed above, we must also have

$$c_1 \leq c_f, \quad 3(c_d + c_e) \leq c_2.$$

Furthermore, by Lemma D.6, it suffices to always take $c_\Delta = 3c_d + 3c_e - c_g$. Direct computation then shows that the following settings suffice, up to machine precision:

$$\begin{aligned}
c_1 &= 0.05978841810030329 \\
c_2 &= 0.0600087370242953 \\
c_3 &= 0.1 \\
c_4 &= 0.1 \\
c_a &= 0.0013004532984432395 \\
c_b &= 0.41043329378840077 \\
c_d &= 0.01 \\
c_e &= 0.01 \\
c_c &= 0.0014065949472697806 \\
c_g &= 0.178 \\
c_f &= c_1
\end{aligned}$$

Given these settings, we can bound

$$\frac{c_3(1+c_g)}{1-c_3} \leq 0.5.$$

□

C.1 Preliminaries

Assumptions and Definitions. For all $y \in \mathcal{Y}$, $\widehat{\Delta}_{\text{safe}}(y) \geq -c_\Delta \epsilon$. We will also assume that $\mathcal{Y} \subseteq \mathcal{X}$. We define

$$y_\star = \arg\min_{y \in \mathcal{Y}} y^\top \theta_*$$

and

$$\Delta(z) = \theta_*^\top (z - y_\star).$$

We will take $\gamma = 0$, so we set $A(\lambda) = \sum_{x \in \mathcal{X}} \lambda_x x x^\top$.

C.2 Algorithm and Main Results

At a high-level, RAGE$^\epsilon$ attempts to estimate the difference between the performance of each $z \in \mathcal{Z}$ and the *best* $y \in \mathcal{Y}$. The safety gap estimate, $\widehat{\Delta}_{\text{safe}}(z)$, acts as a regularizer: if $\widehat{\Delta}_{\text{safe}}(z) < 0$, then we do not seek to estimate the gap of z with as high accuracy, since we can already eliminate it by showing it is unsafe. The proof in this section follow closely the proof given in Section 6.4.4 of Jamieson and Jain [2022].

Algorithm 4 RAGE$^\epsilon$

input: active set $\mathcal{Z}$, optimal set $\mathcal{Y}$, tolerance ϵ, confidence δ, safety gap estimate $\{\widehat{\Delta}_{\text{safe}}(z)\}_{z\in\mathcal{Z}}$
Choose $\widehat{y}_0$ arbitrarily from $\mathcal{Y}$, set $\widehat{\Delta}^0(z) \leftarrow 0$ for all $z \in \mathcal{Z}$
for $\ell = 1, 2, \ldots, \lceil \log(2/c_f\epsilon) \rceil$ **do**
    $\epsilon_\ell \leftarrow \frac{2}{c_f} \cdot 2^{-\ell}$
    Let τ_ℓ be the minimal value of $\tau = 2^j \geq 4 \log \frac{4|\mathcal{Z}|^2\ell^2}{\delta}$ such that the objective to the following is no greater than $c_c\epsilon_\ell$, and λ_ℓ the corresponding optimal distribution

$$\inf_{\lambda\in\triangle_\mathcal{X}} \max_{z\in\mathcal{Z}} -c_a(\mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + \sqrt{\frac{\|z - \widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau}}.$$

    Sample $x_t \sim \lambda_\ell$, collect observations $\{(x_t, r_t, s_{t,1}, \ldots, s_{t,m})\}_{t=1}^{\tau_\ell}$
    $\mathcal{W} \leftarrow \{z - z' \ : \ z, z' \in \mathcal{Z}\}$
    $\widehat{\theta}^\ell \leftarrow \mathsf{RIPS}(\{(x_t, r_t)\}_{t=1}^{\tau_\ell}, \mathcal{W}, \frac{\delta}{2\ell^2})$
    Set

$$\widehat{y}_\ell \leftarrow \arg\min_{y\in\mathcal{Y}} y^\top\widehat{\theta}^\ell + 8\sqrt{\frac{\|y - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$

$$\widehat{\Delta}^\ell(y) \leftarrow (y - \widehat{y}_\ell)^\top\widehat{\theta}^\ell + \sqrt{\frac{\|y - \widehat{y}_\ell\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$

return $\{\widehat{\Delta}^\ell(z)\}_{z\in\mathcal{Z}}$

Theorem 4. *With probability at least* $1 - \delta$, RAGE$^\epsilon$ *will terminate after collecting at most*

$$C \cdot \sum_{\ell=1}^{\lceil \log 2/c_f\epsilon \rceil} \inf_{\lambda\in\triangle_\mathcal{X}} \max_{z\in\mathcal{Z}} \frac{\|z - y_\star\|^2_{A(\lambda)^{-1}}}{(\mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z)) + \mathfrak{p}(\Delta(z)) + \epsilon_\ell)^2} \cdot \log(\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta}) + 4\lceil \log \tfrac{2}{c_f\epsilon} \rceil \log(\tfrac{4|\mathcal{Z}|^2\lceil \log \frac{2}{c_f\epsilon} \rceil}{\delta})$$

samples, for a universal constant C, *and will output estimates of the gaps* $\widehat{\Delta}(z)$ *such that, for all* $z \in \mathcal{Z}$,

$$|\widehat{\Delta}(z) - \Delta(z)| \leq c_f \left(\epsilon + \mathfrak{p}(\Delta(z)) + \mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z))\right).$$

C.3 Estimating the Gaps

Lemma C.2. *Let* $\mathcal{E}_{\text{RAGE}^\epsilon}$ *denote the event that for all* ℓ *and all* $z, z' \in \mathcal{X}$, *we have:*

$$|(\widehat{\theta}^\ell - \theta_*)^\top(z - z')| \leq \sqrt{8\|z - z'\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log \frac{4|\mathcal{Z}|^2\ell^2}{\delta}}{\tau_\ell}}.$$

Then $\mathbb{P}[\mathcal{E}_{\mathrm{RAGE}^\epsilon}] \geq 1-\delta$.

Proof. Since $\tau_\ell \geq 4\log\frac{4|\mathcal{Z}|^2\ell^2}{\delta}$, we can apply Proposition 3 to get that, with probability at least $1-\delta/2\ell^2$, for all $w \in \mathcal{W}$,

$$|(\widehat{\theta}^\ell - \theta_*)^\top w| \leq \sqrt{8\|w\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log\frac{4|\mathcal{Z}|^2\ell^2}{\delta}}{\tau_\ell}}.$$

The result then follows by a union bound since

$$\sum_{\ell=1}^{\infty} \frac{\delta}{2\ell^2} = \frac{\pi^2}{12}\delta \leq \delta.$$

□

Lemma C.3. *On* $\mathcal{E}_{\mathrm{RAGE}^\epsilon}$*, for all* $z \in \mathcal{Z}$ *and all* ℓ*,*

$$|\widehat{\Delta}^\ell(z) - \theta_*^\top(z - \widehat{y}_\ell)| \leq 8c_a\left(\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_\ell))\right) + 8(c_c + c_a + 2c_ac_\Delta)\epsilon_\ell.$$

Proof. By construction, we have that

$$\max_{z\in\mathcal{Z}} -c_a(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + \sqrt{\frac{\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} \leq c_c\epsilon_\ell.$$

This implies that, for all $z \in \mathcal{Z}$:

$$\sqrt{\frac{\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} \leq c_a(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))) + (c_c + c_a)\epsilon_\ell.$$

On $\mathcal{E}_{\mathrm{RAGE}^\epsilon}$, we have

$$\begin{aligned}
|\widehat{\Delta}_\ell(z) - \theta_*^\top(z-\widehat{y}_\ell)| &\leq \sqrt{8\|z-\widehat{y}_\ell\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} \\
&\leq \sqrt{16\|\widehat{y}_{\ell-1}-\widehat{y}_\ell\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} + \sqrt{16\|\widehat{y}_{\ell-1}-\widehat{y}_\ell\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} \\
&\leq 8c_a\left(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_\ell)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_\ell))\right) + 8(c_c+c_a)\epsilon_\ell.
\end{aligned}$$

By construction we have that $\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_\ell) \geq -c_\Delta\epsilon \geq -2c_\Delta\epsilon_\ell$, so $\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_\ell)) \leq 2c_\Delta\epsilon_\ell$, which proves the result. □

Lemma C.4. *On* $\mathcal{E}_{\mathrm{RAGE}^\epsilon}$ *and the event that* $\widehat{\Delta}^{\ell-1}(y_\star) \leq c_b\epsilon_\ell$*, we have*

$$\Delta(\widehat{y}_\ell) \leq 6(c_c + c_a(1 + c_b + 2c_\Delta))\epsilon_\ell.$$

Proof. By the definition of $\mathcal{E}_{\mathrm{RAGE}^\epsilon}$ and $\widehat{y}_\ell$, we can bound

$$\theta_*^\top(\widehat{y}_\ell - \widehat{y}_{\ell-1}) \leq (\widehat{\theta}^\ell)^\top(\widehat{y}_\ell - \widehat{y}_{\ell-1}) + \sqrt{8\|\widehat{y}_\ell - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$

$$= \min_{y \in \mathcal{Y}} (\widehat{\theta}^\ell)^\top (y - \widehat{y}_{\ell-1}) + \sqrt{8\|y - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$
$$\leq (\widehat{\theta}^\ell)^\top (y_\star - \widehat{y}_{\ell-1}) + \sqrt{8\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$
$$\leq \theta_*^\top (y_\star - \widehat{y}_{\ell-1}) + 2\sqrt{8\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}.$$

By the definition of τ_ℓ and λ_ℓ, we have

$$c_c\epsilon_\ell \geq \max_{z \in \mathcal{Z}} -c_a(\mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + \sqrt{\frac{\|z - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$
$$\geq -c_a(\mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(y_\star)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(y_\star)) + \epsilon_\ell) + \sqrt{\frac{\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$
$$\overset{(a)}{\geq} -c_a(\mathfrak{p}(\widehat{\Delta}^{\ell-1}(y_\star)) + (1 + 2c_\Delta)\epsilon_\ell) + \sqrt{\frac{\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$
$$\overset{(b)}{\geq} -c_a(1 + c_b + 2c_\Delta)\epsilon_\ell + \sqrt{\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}}$$

where (a) uses that $\widehat{\Delta}_{\text{safe}}(y_\star) \geq -c_\Delta\epsilon \geq -2c_\Delta\epsilon_\ell$, by definition, and (b) follows by our assumption on $\widehat{\Delta}^{\ell-1}(y_\star)$. This implies that

$$\sqrt{\|y_\star - \widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4|\mathcal{Z}|^2\ell^2}{\delta})}{\tau_\ell}} \leq (c_c + c_a(1 + c_b + 2c_\Delta))\epsilon_\ell.$$

Combining this with the above we have that

$$\theta_*^\top (\widehat{y}_\ell - \widehat{y}_{\ell-1}) \leq \theta_*^\top (y_\star - \widehat{y}_{\ell-1}) + 6(c_c + c_a(1 + c_b + 2c_\Delta))\epsilon_\ell.$$

Rearranging this proves the result. □

Lemma C.5. *For all $z \in \mathcal{Z}$ and all ℓ, on the event $\mathcal{E}_{\text{RAGE}^\epsilon}$,*

$$|\widehat{\Delta}^\ell(z) - \Delta(z)| \leq c_f \left(\epsilon_\ell + \mathfrak{p}(\Delta(z)) + \mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z))\right).$$

Proof. We prove this by induction. Assume that at $\ell - 1$, for all $z \in \mathcal{Z}$,

$$|\widehat{\Delta}^{\ell-1}(z) - \Delta(z)| \leq c_f \left(\epsilon_{\ell-1} + \mathfrak{p}(\Delta(z)) + \mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z))\right).$$

On $\mathcal{E}_{\text{RAGE}^\epsilon}$ and by Lemma C.3 we can bound

$$|\widehat{\Delta}^\ell(z) - \Delta(z)| = |\widehat{\Delta}^\ell(z) - (R(z) - R(\widehat{y}_\ell) + R(\widehat{y}_\ell) - R(y_\star))|$$
$$\leq |\widehat{\Delta}^\ell(z) - (R(z) - R(\widehat{y}_\ell))| + \Delta(\widehat{y}_\ell)$$
$$\leq 8c_a \left(\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_\ell))\right) + 8(c_c + c_a + 2c_ac_\Delta)\epsilon_\ell + \Delta(\widehat{y}_\ell).$$

By the inductive hypothesis, we can bound

$$\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \le (1+c_f)\mathfrak{p}(\Delta(z)) + c_f\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_f\epsilon_{\ell-1}$$
$$\mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_\ell)) \le (1+c_f)\mathfrak{p}(\Delta(\widehat{y}_\ell)) + c_f\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_\ell)) + c_f\epsilon_{\ell-1}.$$

By construction $\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_\ell)) \ge -c_\Delta\epsilon \ge -2c_\Delta\epsilon_\ell$, so

$$|\widehat{\Delta}^{\ell}(z) - \Delta(z)| \le 8c_a(1+c_f)\mathfrak{p}(\Delta(z)) + 8c_a(1+c_f)\mathfrak{p}(\widehat{\Delta}_{\mathrm{safe}}(z)) + (8c_a(1+c_f)+1)\Delta(\widehat{y}_\ell) + 8(c_ac_f(1+c_\Delta)+c_c+c_a+2c_ac_\Delta)\epsilon_\ell.$$

It remains to bound $\Delta(\widehat{y}_\ell) = R(\widehat{y}_\ell) - R(y_\star)$. On the inductive hypothesis, we have that

$$|\widehat{\Delta}^{\ell-1}(y_\star) - \Delta(y_\star)| \le c_f\left(\epsilon_{\ell-1} + \mathfrak{p}(\Delta(y_\star)) + \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(y_\star))\right).$$

By definition, $\Delta(y_\star) = 0$ and $\widehat{\Delta}_{\mathrm{safe}}(y_\star) \ge -c_\Delta\epsilon \ge -2c_\Delta\epsilon_\ell$, which implies that $\widehat{\Delta}^{\ell-1}(y_\star) \le 2c_f(1+c_\Delta)\epsilon_\ell$. It follows that the conditions of Lemma C.4 are met as long as

$$2c_f(1+c_\Delta) \le c_b, \tag{C.1}$$

so we can bound $\Delta(\widehat{y}_\ell) \le 6(c_c + c_a(1+c_b+2c_\Delta))\epsilon_\ell$. Thus,

$$|\widehat{\Delta}^{\ell}(z) - \Delta(z)| \le 8c_a(1+c_f)\mathfrak{p}(\Delta(z)) + 8c_a(1+c_f)\mathfrak{p}(\widehat{\Delta}_{\mathrm{safe}}(z)) + 8(c_ac_f(1+c_\Delta)+c_c+c_a+2c_ac_\Delta)\epsilon_\ell + (8c_a(1+c_f)+1)(6(c_c+c_a(1+c_b+2c_\Delta)))\epsilon_\ell.$$

which proves the inductive hypothesis as long as

$$\begin{aligned} &8(c_ac_f(1+c_\Delta)+c_c+c_a+2c_ac_\Delta) + (8c_a(1+c_f)+1)(6(c_c+c_a(1+c_b+2c_\Delta))) \le c_f \\ &8c_a(1+c_f) \le c_f \end{aligned} \tag{C.2}$$

For the base case, we need to show that

$$|\widehat{\Delta}^{0}(z) - \Delta(z)| \le c_f\left(\epsilon_0 + \mathfrak{p}(\Delta(z)) + \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z))\right).$$

By construction $\widehat{\Delta}^0(z) = 0$ for all z, and $\mathfrak{p}(\Delta(z)) \ge 0$, $\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) \ge 0$. Thus, it suffices to show $|\Delta(z)| \le c_f\epsilon_0$. However, by construction $|\Delta(z)| \le 1$, and $c_f\epsilon_0 = 1$, which proves the base case. □

C.4 Bounding the Sample Complexity

Lemma C.6. *On the event $\mathcal{E}_{\mathrm{RAGE}^\epsilon}$, RAGE^ϵ will terminate after collecting at most*

$$C\cdot\sum_{\ell=1}^{\lceil\log(2/c_f\epsilon)\rceil}\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z-y_\star\|^2_{A(\lambda)^{-1}}}{(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z))+\mathfrak{p}(\Delta(z))+\epsilon_\ell)^2}\cdot\log(\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta}) + 8\lceil\log\tfrac{2}{c_f\epsilon}\rceil\log(\tfrac{4|\mathcal{Z}|^2\lceil\log\frac{2}{c_f\epsilon}\rceil^2}{\delta})$$

samples, for a universal constant C.

Proof. If, for all $z\in\mathcal{Z}$,

$$\tau \ge \frac{\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}}}{(c_a(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z))+\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))+\epsilon_\ell)+c_c\epsilon_\ell)^2}\cdot\log(\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta})$$

we will have that the objective on Line 5 of RAGE$^\epsilon$ is less than $c_c\epsilon_\ell$. Since we can take the best-case $\lambda \in \triangle_{\mathcal{X}}$, and since we have that τ_ℓ will be at most a factor of 2 from the optimal τ, it follows that

$$\begin{aligned}\tau_\ell &\le \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{z\in\mathcal{Z}} \frac{2\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}}}{(c_a(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + c_c\epsilon_\ell)^2} \cdot \log(\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta}) \vee 8\log\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta} \\ &\le \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{z\in\mathcal{Z}} \frac{2\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}}}{(c_a(\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + c_c\epsilon_\ell)^2} \cdot \log(\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta}) + 8\log\tfrac{4|\mathcal{Z}|^2\ell^2}{\delta}\end{aligned}$$

where the additional $8\log\frac{4|\mathcal{Z}|^2\ell^2}{\delta}$ factor arises since we always require $\tau_\ell \ge 4\log\frac{4|\mathcal{Z}|^2\ell^2}{\delta}$.

We can upper bound

$$\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}} \le 2\|z-y_\star\|^2_{A(\lambda)^{-1}} + 2\|y_\star-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}}.$$

By construction, $\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_{\ell-1})) \le 2c_\Delta\epsilon_\ell$, so for any z, $\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_{\ell-1})) - 2c_\Delta\epsilon_\ell \le \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z))$. Furthermore, by definition,

$$\widehat{\Delta}^{\ell-1}(\widehat{y}_{\ell-1}) = 0$$

so $\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \ge \mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_{\ell-1}))$. Thus,

$$\begin{aligned}&\inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{z\in\mathcal{Z}} \frac{\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}}}{(c_a\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_a\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + (c_a+c_c)\epsilon_\ell)^2} \\ &\le \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{z\in\mathcal{Z}} \frac{2\|z-y_\star\|^2_{A(\lambda)^{-1}}}{(c_a\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_a\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + (c_a+c_c)\epsilon_\ell)^2} \\ &\quad + \frac{2\|\widehat{y}_{\ell-1}-y_\star\|^2_{A(\lambda)^{-1}}}{(c_a\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(\widehat{y}_{\ell-1})) + c_a\mathfrak{p}(\widehat{\Delta}^{\ell-1}(\widehat{y}_{\ell-1})) + (c_a+c_c-2c_ac_\Delta)\epsilon_\ell)^2} \\ &\le \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{z\in\mathcal{Z}} \frac{4\|z-y_\star\|^2_{A(\lambda)^{-1}}}{(c_a\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_a\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + (c_a+c_c-2c_ac_\Delta)\epsilon_\ell)^2}.\end{aligned}$$

By Lemma C.5, we can lower bound

$$\widehat{\Delta}^{\ell-1}(z) \ge \Delta(z) - c_f(\epsilon_\ell + \mathfrak{p}(\Delta(z)) + \mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)))$$

so

$$\begin{aligned}&c_a\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_a\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + (c_a+c_c-2c_ac_\Delta)\epsilon_\ell \\ &\quad\ge c_a(1-c_f)\mathfrak{p}(-\widehat{\Delta}_{\mathrm{safe}}(z)) + c_a(1-c_f)\mathfrak{p}(\Delta(z)) + (c_a+c_c-2c_ac_\Delta-c_ac_f)\epsilon_\ell.\end{aligned}$$

The result follows by combining these inequalities and as long as

$$c_a(1-c_f) \ge c_0, \qquad c_a+c_c-2c_ac_\Delta-c_ac_f \ge c_0. \tag{C.3}$$

□

Proof of Theorem 4. Theorem 4 follows directly from Lemma C.6 and Lemma C.5 since, by Lemma C.2, $\mathcal{E}_{\text{RAGE}^\epsilon}$ holds with probability at least $1-\delta$. □

D Safe Best-Arm Identification

D.1 Preliminaries

In general we want to consider multiple safety constraints, and let m denote the number of constraints. In such settings, we will denote $\Delta^i_{\text{safe}}(z)$ the safety gap for safety constraint i.

Define

$$\widetilde{\Delta}^\ell(z) := \theta_*^\top z - \min_{y \in \mathcal{Y}_\ell} \theta_*^\top y.$$

D.2 Algorithm and Main Result

Theorem 3 (Full version of Theorem 1)**.** *With probability at least $1-2\delta$, Algorithm 1 returns an arm $\widehat{z}$ such that*

$$\widehat{z}^\top \theta_* \geq (z_*)^\top \theta_* - \epsilon, \quad \Delta_{\text{safe}}(\widehat{z}) \geq -\epsilon \tag{D.1}$$

and terminates after collecting at most

$$C \cdot \sum_{\ell=1}^{\iota_\epsilon} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\left(\min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\epsilon_{\ell-1}}(z)) + \epsilon_\ell\right)^2}$$
$$+ C \log \frac{1}{\epsilon} \cdot \sum_{\ell=1}^{\iota_\epsilon} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \frac{\|z - z_*\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{8|\mathcal{Z}|^2 \log^4(1/\epsilon)}{\delta})}{(\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\epsilon_\ell}(z)) + \epsilon_\ell)^2} + C_0$$

samples for a universal constant C, $C_0 = \operatorname{poly}\log(\frac{1}{\epsilon}, |\mathcal{Z}|) \cdot \log \frac{1}{\delta}$.

D.3 Estimating the Safety Value

Lemma D.1. *Let $\mathcal{E}_{\text{safe}}$ denote the event that, for all ℓ, $z \in \mathcal{Z}$, $i \in [m]$:*

$$|z^\top(\widehat{\mu}^{i,\ell} - \mu^i_*)| \leq \sqrt{8\|z\|^2_{A(\lambda_\ell)^{-1}} \cdot \frac{\log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau_\ell}}.$$

Then $\mathbb{P}[\mathcal{E}_{\text{safe}}] \geq 1-\delta$.

Proof. This follows directly from Proposition 3 and a union bound, as in Lemma C.2. □

Lemma D.2. *On $\mathcal{E}_{\text{safe}}$, for all $z \in \mathcal{Z}$, $i \in [m]$, and all ℓ,*

$$|\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - \Delta^i_{\text{safe}}(z)| \leq 3c_d \left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))\right) + 3(c_d + c_e)\epsilon_\ell.$$

Proof. By construction, we have that

$$\max_{z \in \mathcal{Z}} -c_d \left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell\right) + \sqrt{\frac{\|z\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau_\ell}} \leq c_e\epsilon_\ell.$$

This implies that, for all $z \in \mathcal{Z}$,

$$\sqrt{\frac{\|z\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau_\ell}} \le \min_j c_d |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j c_d \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + c_d \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + (c_d + c_e)\epsilon_\ell.$$

On $\mathcal{E}_{\text{safe}}$, we have

$$\begin{aligned}|\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - \Delta^i_{\text{safe}}(z)| &\le \sqrt{8\frac{\|z\|^2_{A(\lambda_\ell)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau_\ell}} \\ &\le \min_j 3c_d |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j 3c_d \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + 3c_d \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + 3(c_d + c_e)\epsilon_\ell\end{aligned}$$

which proves the result. □

D.4 Tying Together Safety Estimation with Optimality Estimation

Definition D.1 (Optimality Good Event)**.** Let $\mathcal{E}^\ell_{\text{RAGE}^\epsilon}$ denote the success event of RAGE$^\epsilon$ when called at the ℓth epoch, and $\mathcal{E}_{\text{RAGE}^\epsilon} := \cup_\ell \mathcal{E}^\ell_{\text{RAGE}^\epsilon}$.

Lemma D.3. *On the event $\mathcal{E}_{\text{safe}} \cap \mathcal{E}_{\text{RAGE}^\epsilon}$, we have that:*

1. *For all $\ell \le \iota_\epsilon$, $y \in \mathcal{Y}_\ell$, and $i \in [m]$, $y^\top \mu_{*,i} \le \gamma$.*
2. *For all ℓ and $z \in \mathcal{Z}$, $\widetilde{\Delta}^{\ell-1}(z) \le \widetilde{\Delta}^\ell(z)$.*

Proof. By Lemma D.2, we have that

$$\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - 3c_d\left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))\right) - 3(c_d + c_e)\epsilon_\ell \le \Delta^i_{\text{safe}}(z).$$

Thus, if the inclusion condition of $\mathcal{Y}_\ell$ is met, it must be the case that $\Delta^i_{\text{safe}}(z) \ge 0$ for all i.

The second conclusion follows directly since $\mathcal{Y}_{\ell-1} \subseteq \mathcal{Y}_\ell$. □

Lemma D.4 (Key Estimation Error Bound)**.** *On the event $\mathcal{E}_{\text{safe}} \cap \mathcal{E}_{\text{RAGE}^\epsilon}$, for all $z \in \mathcal{Z}$, ℓ, and i, we have*

$$\begin{aligned}|\widehat{\Delta}^\ell(z) - \widetilde{\Delta}^\ell(z)| &\le c_3\left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z))\right) \\ |\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - \Delta^i_{\text{safe}}(z)| &\le c_4\left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z))\right).\end{aligned}$$

Proof. We prove this by induction. Assume that the above inequalities hold at epoch $\ell - 1$. On $\mathcal{E}_{\text{safe}} \cap \mathcal{E}_{\text{RAGE}^\epsilon}$, by Lemma C.5 and Lemma D.2, we have

$$\begin{aligned}|\widehat{\Delta}^\ell(z) - \widetilde{\Delta}^\ell(z)| &\le c_1(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z))) \\ |\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - \Delta^i_{\text{safe}}(z)| &\le c_2(\epsilon_\ell + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z))).\end{aligned}$$

By the inductive hypothesis, we can bound

$$\mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \le \mathfrak{p}\left(\widetilde{\Delta}^{\ell-1}(z) + c_3(\epsilon_{\ell-1} + \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)))\right)$$

$$\leq (1+c_3)\mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + c_3\epsilon_{\ell-1} + \max_j c_3\mathfrak{p}(-\Delta^j_{\text{safe}}(z))$$
$$\leq (1+c_3)\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + 2c_3\epsilon_{\ell} + \max_j c_3\mathfrak{p}(-\Delta^j_{\text{safe}}(z))$$

where the last inequality follows since, by Lemma D.3, $\widetilde{\Delta}^{\ell-1}(z) \leq \widetilde{\Delta}^{\ell}(z)$.

Furthermore, again applying the inductive hypothesis,

$$\begin{aligned}\widehat{\Delta}^{i,\ell-1}_{\text{safe}}(z) &\leq \Delta^i_{\text{safe}}(z) + c_4(\epsilon_{\ell-1} + \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + \min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)))\\ &\leq \Delta^i_{\text{safe}}(z) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \min_j c_4|\Delta^j_{\text{safe}}(z)| + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z))\\ &\leq \Delta^i_{\text{safe}}(z) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + c_4|\Delta^i_{\text{safe}}(z)| + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z)).\end{aligned}$$

Similarly,

$$\begin{aligned}\mathfrak{p}(-\widehat{\Delta}^{i,\ell-1}_{\text{safe}}(z)) &\leq \mathfrak{p}\left(-\Delta^i_{\text{safe}}(z) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \min_j c_4|\Delta^j_{\text{safe}}(z)| + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z))\right)\\ &\leq \mathfrak{p}\left(-\Delta^i_{\text{safe}}(z) + \min_j c_4|\Delta^j_{\text{safe}}(z)|\right) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z))\\ &\leq \mathfrak{p}\left(-\Delta^i_{\text{safe}}(z) + c_4|\Delta^i_{\text{safe}}(z)|\right) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z)).\end{aligned}$$

Note that if $\Delta^i_{\text{safe}}(z) \leq 0$, then

$$\mathfrak{p}(-\Delta^i_{\text{safe}}(z) + c_4|\Delta^i_{\text{safe}}(z)|) = \mathfrak{p}(-\Delta^i_{\text{safe}}(z) - c_4\Delta^i_{\text{safe}}(z)) = (1+c_4)\mathfrak{p}(-\Delta^i_{\text{safe}}(z))$$

and if $\Delta^i_{\text{safe}}(z) > 0$, then for $c_4 < 1$, $-\Delta^i_{\text{safe}}(z) + c_4|\Delta^i_{\text{safe}}(z)| \leq 0$, so

$$\mathfrak{p}(-\Delta^i_{\text{safe}}(z) + c_4|\Delta^i_{\text{safe}}(z)|) = 0 = (1+c_4)\mathfrak{p}(-\Delta^i_{\text{safe}}(z)).$$

Thus,

$$\mathfrak{p}(-\widehat{\Delta}^{i,\ell-1}_{\text{safe}}(z)) \leq (1+c_4)\mathfrak{p}(-\Delta^i_{\text{safe}}(z)) + 2c_4\epsilon_{\ell} + c_4\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \max_j c_4\mathfrak{p}(-\Delta^j_{\text{safe}}(z)).$$

Combining these inequalities, it follows that

$$\begin{aligned}|\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) - \Delta^i_{\text{safe}}(z)| &\leq c_2\left(\epsilon_{\ell} + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z))\right)\\ &\leq c_2(1+2c_3+4c_4)\epsilon_{\ell} + c_2(1+c_3+2c_4)\mathfrak{p}(\widetilde{\Delta}^{\ell}(z))\\ &\quad + c_2(1+c_3+3c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + c_2(1+c_4)\min_j |\Delta^j_{\text{safe}}(z)|\end{aligned}$$

and

$$\begin{aligned}|\widehat{\Delta}^{\ell}(z) - \widetilde{\Delta}^{\ell}(z)| &\leq c_1(\epsilon_{\ell} + \mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)))\\ &\leq c_1(1+2c_4)\epsilon_{\ell} + c_1(1+c_4)\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + c_1(1+2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)).\end{aligned}$$

This proves the inductive hypothesis, as long as

$$c_1(1+2c_4) \leq c_3, \quad c_2(1+2c_3+4c_4) \leq c_4. \tag{D.2}$$

For the base case, we need to show that

$$|\widehat{\Delta}^0(z) - \widetilde{\Delta}^0(z)| \le c_3(\epsilon_0 + \mathfrak{p}(\widetilde{\Delta}^0(z)) + \max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)))$$
$$|\widehat{\Delta}^0_{\mathrm{safe}}(z) - \Delta_{\mathrm{safe}}(z)| \le c_4(\epsilon_0 + \mathfrak{p}(\widetilde{\Delta}^0(z)) + \min_j |\Delta^j_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z))).$$

By construction, $\widehat{\Delta}^0(z) = \widehat{\Delta}^0_{\mathrm{safe}}(z) = 0$. Thus, it suffices to show $|\widetilde{\Delta}^0(z)| \le c_3\epsilon_0$ and $|\Delta_{\mathrm{safe}}(z)| \le c_4\epsilon_0$. However, both of these are true by our choice of ϵ_0. □

Lemma D.5. *On the event $\mathcal{E}_{\mathrm{safe}} \cap \mathcal{E}_{\mathrm{RAGE}^\epsilon}$, for all $z \in \mathcal{Z}$ and all ℓ, we will have*

$$\widetilde{\Delta}^\ell(z) \ge \Delta^{\epsilon_\ell}(z) \quad \textit{where} \quad \Delta^{\epsilon_\ell}(z) = \max_{y \in \mathcal{Z} \,:\, \epsilon_\ell \le \min_i \Delta^i_{\mathrm{safe}}(y)} y^\top\theta_* - z^\top\theta_*.$$

Proof. By definition, we will have $z \in \mathcal{Y}_\ell$ if

$$8c_d\left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))\right) + 8(c_d + c_e)\epsilon_\ell \le \widehat{\Delta}^{i,\ell}_{\mathrm{safe}}(z).$$

The following claim allows us to obtain a sufficient condition to guarantee $z \in \mathcal{Y}_\ell$.

Claim D.1. *On the event $\mathcal{E}_{\mathrm{safe}} \cap \mathcal{E}_{\mathrm{RAGE}^\epsilon}$,*

$$\begin{aligned}&\min_j |\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \\ &\quad \le 2(c_3 + 2c_4)\epsilon_\ell + (1 + c_3 + 2c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + (1 + 2c_4)\min_j |\Delta^j_{\mathrm{safe}}(z)| + (1 + c_3 + 2c_4)\max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)).\end{aligned}$$

Proof of Claim D.1. By Lemma D.3 and Lemma D.4, we can bound

$$\begin{aligned}\min_j |\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)| &\le (1 + c_4)\min_j |\Delta^j_{\mathrm{safe}}(z)| + 2c_4\epsilon_\ell + c_4\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + c_4\max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)) \\ \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)) &\le (1 + c_4)\max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)) + 2c_4\epsilon_\ell + c_4\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + c_4\min_j |\Delta^j_{\mathrm{safe}}(z)| \\ \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) &\le (1 + c_3)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + 2c_3\epsilon_\ell + c_3\max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)).\end{aligned}$$

The claim follows by summing these upper bounds. □

Thus, by Claim D.1, we can bound

$$\begin{aligned}&3c_d\left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\mathrm{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z))\right) + 3(c_d + c_e)\epsilon_\ell \\ &\quad \le 3(c_d + c_e + 2c_dc_3 + 4c_dc_4)\epsilon_\ell + 3c_d(1 + c_3 + 2c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) \\ &\quad + 3c_d(1 + 2c_4)\min_j |\Delta^j_{\mathrm{safe}}(z)| + 3c_d(1 + c_3 + 2c_4)\max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z)).\end{aligned}$$

Furthermore, by Lemma D.4,

$$\Delta^i_{\mathrm{safe}}(z) - c_4\left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \min_j |\Delta^j_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z))\right) \le \widehat{\Delta}^{i,\ell}_{\mathrm{safe}}(z)$$

It follows that a sufficient condition for $z \in \mathcal{Y}_\ell$ is

$$\begin{aligned}&(3c_d + 3c_e + 6c_dc_3 + 12c_dc_4 + c_4)\left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \min_j |\Delta^j_{\mathrm{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\mathrm{safe}}(z))\right) \\ &\quad \le \Delta^i_{\mathrm{safe}}(z), \quad \forall i \in [m].\end{aligned} \tag{D.3}$$

If $y_\ell = \arg\max_{y \in \mathcal{Z} \,:\, \epsilon_\ell \le \min_i \Delta^i_{\text{safe}}(y)} y^\top \theta_*$ is in $\mathcal{Y}_\ell$, then we are done. Assume then that $y_\ell \notin \mathcal{Y}_\ell$. By construction, since $\Delta^i_{\text{safe}}(y_\ell) > 0$ for all i, $\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) = 0$. Using that (D.3) is a sufficient condition for inclusion in $\mathcal{Y}_\ell$, this implies that

$$\exists i \in [m] \quad \text{s.t.} \quad (3c_d + 3c_e + 6c_d c_3 + 12 c_d c_4 + c_4) \left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(y_\ell)) + \min_j |\Delta^j_{\text{safe}}(y_\ell)| \right) > \Delta^i_{\text{safe}}(y_\ell).$$

which implies

$$\exists i \in [m] \quad \text{s.t.} \quad (3c_d + 3c_e + 6c_d c_3 + 12 c_d c_4 + c_4) \left(\epsilon_\ell + \mathfrak{p}(\widetilde{\Delta}^\ell(y_\ell)) + |\Delta^i_{\text{safe}}(y_\ell)| \right) > \Delta^i_{\text{safe}}(y_\ell). \tag{D.4}$$

By construction, though, $\Delta^i_{\text{safe}}(y_\ell) \ge \epsilon_\ell$. If we assume that

$$3c_d + 3c_e + 6c_d c_3 + 12 c_d c_4 + c_4 \le 1/4, \tag{D.5}$$

then (D.4) can only hold if $\mathfrak{p}(\widetilde{\Delta}^\ell(y_\ell)) > 0$. This implies that $\max_{y \in \mathcal{Y}_\ell} y^\top \theta_* > y_\ell^\top \theta_*$. Thus, in this case,

$$\widetilde{\Delta}^\ell(z) = \max_{y \in \mathcal{Y}_\ell} y^\top \theta_* - z^\top \theta_* > y_\ell^\top \theta_* - z^\top \theta_* = \Delta^{\epsilon_\ell}(z)$$

which proves the result. □

Lemma D.6. *On $\mathcal{E}_{\text{safe}} \cap \mathcal{E}_{\text{RAGE}^\epsilon}$, for all $z \in \mathcal{Y}_{\text{end}}$ we have*

$$\begin{aligned} \Delta^i_{\text{safe}}(z) &\ge -c_g \epsilon, \quad \forall i \in [m], \\ \widehat{\Delta}^{i,\iota_\epsilon}_{\text{safe}}(z) &\ge (3c_d + 3c_e - c_g)\epsilon, \quad \forall i \in [m]. \end{aligned}$$

Furthermore, $z_* \in \mathcal{Y}_{\text{end}}$.

Proof. Recall that

$$\begin{aligned} \mathcal{Y}_{\text{end}} = \{ z \in \mathcal{Z} \;:\; & 3c_d \left(\min_j |\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\iota_\epsilon}(z)) \right) \\ & + 3(c_d + c_e)\epsilon - c_g \epsilon \le \widehat{\Delta}^{i,\iota_\epsilon}_{\text{safe}}(z), \forall i \in [m] \} \end{aligned}$$

On $\mathcal{E}_{\text{safe}}$, we have

$$\widehat{\Delta}^{i,\iota_\epsilon}_{\text{safe}}(z) \le \Delta^i_{\text{safe}}(z) + 3c_d \left(\min_j |\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\iota_\epsilon}(z)) \right) + 3(c_d + c_e)\epsilon$$

so it follows that if $z \in \mathcal{Y}_{\text{end}}$, then

$$-c_g \epsilon \le \Delta^i_{\text{safe}}(x).$$

To see that $z_* \in \mathcal{Y}_{\text{end}}$, note that by definition of $\mathcal{Y}_{\text{end}}$, using a calculation analogous to (D.3), a sufficient condition for $z \in \mathcal{Y}_{\text{end}}$ is

$$\begin{aligned} & (3c_d + 3c_e + 6c_d c_3 + 12 c_d c_4 + c_4 - c_g)\epsilon + (3c_d + 3c_d c_3 + 6 c_d c_4 + c_4)\mathfrak{p}(\widetilde{\Delta}^{\iota_\epsilon}(z)) \\ & \quad + (3c_d + 6c_d c_4 + c_4) \min_j |\Delta^j_{\text{safe}}(z)| + (3c_d + 3c_d c_3 + 6c_d c_4 + c_4) \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) \\ & \quad \le \Delta^i_{\text{safe}}(z), \quad \forall i \in [m]. \end{aligned}$$

By definition of z_* and since, by Lemma D.3, all $z \in \mathcal{Y}_{\iota_\epsilon}$ are safe, we have $\Delta^{\epsilon_{\iota_\epsilon}}(z_*) \le 0$. Furthermore, by definition we also have $\Delta^j_{\text{safe}}(z_*) \ge 0$ for all j, so $\mathfrak{p}(-\Delta^j_{\text{safe}}(z_*)) = 0$. Thus, assuming that

$$3c_d + 3c_e + 6c_d c_3 + 12 c_d c_4 + c_4 - c_g \le 0 \tag{D.6}$$

a sufficient condition to guarantee $z_* \in \mathcal{Y}_{\text{end}}$ is that

$$(8c_d + 16 c_d c_4 + c_4) \min_j |\Delta^j_{\text{safe}}(z_*)| \le \Delta^i_{\text{safe}}(z_*), \quad \forall i \in [m].$$

However, as long as

$$3c_d + 6c_d c_4 + c_4 \le 1, \tag{D.7}$$

this is true, since by definition $\Delta^i_{\text{safe}}(z_*) \ge 0$. □

D.5 Algorithm Correctness and Sample Complexity

Lemma D.7 (Correctness). *On $\mathcal{E}_{\text{safe}} \cap \mathcal{E}_{\text{RAGE}^\epsilon}$, we will have that*

$$\widehat{z}^\top \theta_* \ge (z_*)^\top \theta_* - \frac{c_3(1+c_g)}{1-c_3}\epsilon, \quad \Delta^i_{\text{safe}}(\widehat{z}) \ge -c_g\epsilon, \forall i \in [m].$$

Proof. We choose $\widehat{z}$ to be any $z \in \mathcal{Y}_{\text{end}}$ such that $\widehat{\Delta}^{\text{end}}(z) = 0$. By Lemma D.6, we have that $\Delta^i_{\text{safe}}(\widehat{z}) \ge -c_g\epsilon$ for all $i \in [m]$. If $\widetilde{\Delta}^{\text{end}}(\widehat{z}) \le 0$, we are done, since by Lemma D.6, $z_* \in \mathcal{Y}_{\text{end}}$, so $\widehat{z}^\top\theta_* \ge (z_*)^\top\theta_*$. Assume that $\widetilde{\Delta}^{\text{end}}(\widehat{z}) > 0$. By Lemma D.4, we have that

$$\widetilde{\Delta}^{\text{end}}(\widehat{z}) \le c_3\epsilon + c_3 \mathfrak{p}(\widetilde{\Delta}^{\text{end}}(\widehat{z})) + c_3 \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(\widehat{z})).$$

By Lemma D.6, since $\widehat{z} \in \mathcal{Y}_{\text{end}}$, $\mathfrak{p}(-\Delta^j_{\text{safe}}(\widehat{z})) \le c_g\epsilon$ for all j, so we can bound

$$\widetilde{\Delta}^{\text{end}}(\widehat{z}) \le c_3(1+c_g)\epsilon + c_3\mathfrak{p}(\widetilde{\Delta}^{\text{end}}(\widehat{z})) = c_3(1+c_g)\epsilon + c_3\widetilde{\Delta}^{\text{end}}(\widehat{z}).$$

We can rearrange this as

$$\widetilde{\Delta}^{\text{end}}(\widehat{z}) \le \frac{c_3(1+c_g)}{1-c_3}\epsilon$$

which proves the result, since, by Lemma D.6, $\Delta^{\text{end}}(\widehat{z}) = \max_{y \in \mathcal{Y}_{\text{end}}} y^\top\theta_* - \widehat{z}^\top\theta_* \ge (z_*)^\top\theta_* - \widehat{z}^\top\theta_*$. □

Lemma D.8. *On $\mathcal{E}_{\text{RAGE}^\epsilon} \cap \mathcal{E}_{\text{safe}}$, the total complexity of Line 6 is bounded by*

$$C \cdot \sum_{\ell=1}^{\iota_\epsilon} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\left(\min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\Delta^{\epsilon_{\ell-1}}(z)) + \epsilon_\ell\right)^2} + 4\iota_\epsilon \log(\tfrac{4m|\mathcal{Z}|\iota_\epsilon^2}{\delta})$$

for an absolute constant C.

Proof. Applying the same argument as in Claim D.1 but in the opposite direction, we have

$$\begin{aligned}
&\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \\
&\ge -2(c_3+2c_4)\epsilon_\ell + (1-c_3-2c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + (1-2c_4)\min_j |\Delta^j_{\text{safe}}(z)| + (1-c_3-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)).
\end{aligned}$$

We assume that c_3, c_4, and c_0 are chosen such that

$$1 - 2c_3 - 4c_4 \geq c_0, \tag{D.8}$$

which allows us to bound:

$$\inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} -c_d \left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell \right) + \sqrt{\frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau}}$$

$$\leq \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} -c_d c_0 \left(\min_j |\Delta^j_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + \epsilon_\ell \right) + \sqrt{\frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau}}.$$

It follows that if, for all $z \in \mathcal{Z}$,

$$\tau \geq \frac{\|z\|^2_{A(\lambda)^{-1}}}{\left(c_d c_0 \min_j |\Delta^j_{\text{safe}}(z)| + c_d c_0 \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + c_d c_0 \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + (c_d c_0 + c_e)\epsilon_\ell \right)^2} \cdot \log \frac{4m|\mathcal{Z}|\ell^2}{\delta}$$

we will have that this is less than $c_e \epsilon_\ell$. Since we can take the best-case $\lambda \in \triangle_{\mathcal{X}}$, and since τ_ℓ is always within a factor of 2 of the optimal, it follows that

$$\tau_\ell \leq \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \frac{2\|z\|^2_{A(\lambda)^{-1}} \cdot \log \frac{4m|\mathcal{Z}|\ell^2}{\delta}}{\left(c_d c_0 \min_j |\Delta^j_{\text{safe}}(z)| + c_d c_0 \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + c_d c_0 \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) + (c_d c_0 + c_e)\epsilon_\ell \right)^2} + 4 \log \frac{4m|\mathcal{Z}|\ell^2}{\delta}$$

The result then follows by summing over epochs and lower bounding $\widetilde{\Delta}^{\ell-1}(z)$ by $\Delta^{\epsilon_{\ell-1}}(z)$ using Lemma D.5, and assuming that

$$c_d c_0 + c_e \geq c_0. \tag{D.9}$$

□

Proof of Theorem 3. By Lemma D.1 we have that $\mathcal{E}_{\text{safe}}$ holds with probability at least $1 - \delta$. By Lemma C.2, we have that $\mathcal{E}^\ell_{\text{RAGE}^\epsilon}$ holds with probability at least $1 - \delta/(4\ell^2)$. It follows then that $\mathcal{E}_{\text{safe}} \cup (\cup_\ell \mathcal{E}^\ell_{\text{RAGE}^\epsilon})$ holds with probability at least

$$1 - \delta - \sum_\ell \frac{\delta}{4\ell^2} \geq 1 - 2\delta.$$

Assume henceforth that $\mathcal{E}_{\text{safe}} \cup (\cup_\ell \mathcal{E}^\ell_{\text{RAGE}^\epsilon})$ holds. Equation (D.1) follows by Lemma D.7. The total number of samples collected on Line 6 can be bounded by Lemma D.8. It remains to bound the total number of samples used by RAGE$^\epsilon$.

By Lemma C.6, at epoch ℓ RAGE$^\epsilon$ will collect at most

$$C\lceil \log \frac{2}{c_f \epsilon_\ell} \rceil \cdot \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \frac{\|z - y^\ell_\star\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{8|\mathcal{Z}|^2 \log^4(1/\epsilon)}{\delta})}{(\max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \epsilon_\ell)^2} + 8\lceil \log \tfrac{2}{c_f \epsilon} \rceil \log(\frac{4|\mathcal{Z}|^2 \lceil \log \frac{2}{c_f \epsilon} \rceil^2}{\delta})$$

samples, where $y^\ell_\star = \arg\max_{y \in \mathcal{Y}_\ell} y^\top \theta_*$. Assume that $\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) > 0$, then we can upper bound $\min_j |\Delta^j_{\text{safe}}(z)| \leq \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z))$, and by Lemma D.4 we can lower bound

$$\max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) \geq (1 - 2c_4) \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) - c_4 \mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) - c_4 \epsilon_{\ell-1}.$$

Assume instead that $\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) = 0$. Then again by Lemma D.4:

$$\max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) \geq 0 = \max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) \geq (1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) - c_4\mathfrak{p}(\widetilde{\Delta}^{\ell-1}(z)) - c_4\epsilon_{\ell-1}.$$

By Lemma D.3, it follows that

$$\begin{aligned}&\max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + \epsilon_\ell \\ &\geq (1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + (1-c_4)\mathfrak{p}(\widetilde{\Delta}^{\ell}(z)) + (1-2c_4)\epsilon_\ell.\end{aligned}$$

By definition and Lemma D.3 and Lemma D.6 for all ℓ including $\ell = \text{end}$, we can bound $\mathfrak{p}(-\Delta^j_{\text{safe}}(y^\ell_\star)) \leq c_g\epsilon$. Furthermore, by definition $\mathfrak{p}(\widetilde{\Delta}^\ell(y^\ell_\star)) = 0$. Putting all of this together, we have:

$$\begin{aligned}&\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}} \frac{\|z-y^\ell_\star\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{8|\mathcal{Z}|^2\log^4(1/\epsilon)}{\delta})}{(\max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widetilde{\Delta}^\ell(z)) + \epsilon_\ell)^2} \\ &\leq \inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}} \frac{\|z-y^\ell_\star\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{8|\mathcal{Z}|^2\log^4(1/\epsilon)}{\delta})}{((1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + (1-c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + (1-2c_4)\epsilon_\ell)^2} \\ &\leq \inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}} \frac{2\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{8|\mathcal{Z}|^2\log^4(1/\epsilon)}{\delta})}{((1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + (1-c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + (1-2c_4)\epsilon_\ell)^2} \\ &\quad + \inf_{\lambda\in\triangle_{\mathcal{X}}} \frac{2\|z_*-y^\ell_\star\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{8|\mathcal{Z}|^2\log^4(1/\epsilon)}{\delta})}{((1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(y^\ell_\star)) + (1-c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(y^\ell_\star)) + (1-2c_4-c_g)\epsilon_\ell)^2} \\ &\leq \inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}} \frac{4\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{8|\mathcal{Z}|^2\log^4(1/\epsilon)}{\delta})}{((1-2c_4)\max_j \mathfrak{p}(-\Delta^j_{\text{safe}}(z)) + (1-c_4)\mathfrak{p}(\widetilde{\Delta}^\ell(z)) + (1-2c_4-c_g)\epsilon_\ell)^2}\end{aligned}$$

As long as

$$1-2c_4-c_g \geq c_0, \tag{D.10}$$

summing over the epochs and lower bounding $\widetilde{\Delta}^\ell(z)$ by $\Delta^{\epsilon_\ell}(z)$ via Lemma D.5 gives the result. Finally, the settings of the constants follows from Lemma C.1. □

D.6 Proofs of Corollaries to Theorem 1

Proof of Corollary 1. If $m = 1$, $\mu_{*,1} = 0$, and $\gamma = 1$, then we have $\Delta_{\text{safe}}(z) = 1$ for each z, and $\Delta^{\widetilde{\epsilon}}(z) = \Delta(z)$ for $\epsilon \leq 1$. The result follows directly from this and some algebra. □

Proof of Corollary 2. We can trivially upper bound the complexity given in Theorem 1 by

$$\begin{aligned}&C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{m|\mathcal{Z}|}{\delta})}{\epsilon^2} + C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{|\mathcal{Z}|}{\delta})}{\epsilon^2} + C_0 \\ &\leq C\cdot\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\frac{\|z\|^2_{A(\lambda)^{-1}}\cdot\log(\frac{m|\mathcal{Z}|}{\delta})}{\epsilon^2} + C_0.\end{aligned}$$

In the case when $\mathcal{X} = \mathcal{Z}$, we can bound $\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\mathcal{Z}}\|z\|^2_{A(\lambda)^{-1}} \leq d$ by Kiefer-Wolfowitz [Lattimore and Szepesvári, 2020], which proves the result. □

E Computationally Efficient Optimization

Throughout, we will let $\mathcal{R}(z; \xi_1, \ldots, \xi_n)$ denote some generic weighted risk estimate of the form

$$\mathcal{R}(z; \xi_1, \ldots, \xi_n) = \sum_{t=1}^{T} f_t(\xi_1, \ldots, \xi_n) \mathbb{I}\{z(u_t) \neq v_t\}$$

for some weights $f_t(\xi_1, \ldots, \xi_n)$ and observations (u_t, v_t). The exact setting of $\mathcal{R}$ will change from line to line—we simply use it as a stand-in for an objective that a cost-sensitive-classification oracle can efficiently minimize. We will also use $f(\xi_1, \ldots, \xi_n)$ to refer to some generic function (the particular form of which is not important).

Lemma E.1. *Consider some $z, \widetilde{z} \in \triangle_{\mathcal{H}}$. Denote*

$$\rho_\lambda(h, h') = \mathbb{E}_{U \sim \nu}\left[\frac{\mathbb{I}\{h(U) \neq h'(U)\}}{\lambda(U)/\nu(U)}\right] = \|h - h'\|^2_{A(\lambda)^{-1}}$$

and overload notation so that $z = \sum_{h \in \mathcal{H}} z_h h$ denotes the feature vector for the mixed classifier z. Then,

$$\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} \rho_\lambda(h, h') = \mathbb{E}_{U \sim \nu}\left[\frac{(z(U) - \widetilde{z}(U))^2}{\lambda(U)/\nu(U)}\right] = \|z - \widetilde{z}\|^2_{A(\lambda)^{-1}}.$$

Proof. Note that

$$\rho_\lambda(h, h') = \mathbb{E}_{U \sim \nu}\left[\frac{\mathbb{I}\{h(U) \neq h'(U)\}}{\lambda(U)/\nu(U)}\right] = \mathbb{E}_{U \sim \nu}\left[\frac{|h(U) - h'(U)|}{\lambda(U)/\nu(U)}\right] = \mathbb{E}_{U \sim \nu}\left[\frac{(h(U) - h'(U))^2}{\lambda(U)/\nu(U)}\right]$$

where the final equality holds because $|h(U) - h'(U)|$ is always either 0 or 1. Thus,

$$\begin{aligned}
\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} \rho_\lambda(h, h') &= \sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} \mathbb{E}_{U \sim \nu}\left[\frac{(h(U) - h'(U))^2}{\lambda(U)/\nu(U)}\right] \\
&= \mathbb{E}_{U \sim \nu}\left[\frac{\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} (h(U) - h'(U))^2}{\lambda(U)/\nu(U)}\right] \\
&= \mathbb{E}_{U \sim \nu}\left[\frac{\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} (h(U) + h'(U) - 2h(U)h'(U))}{\lambda(U)/\nu(U)}\right]. \qquad \text{(E.1)}
\end{aligned}$$

However,

$$\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} h(U) = \sum_{h \in \mathcal{H}} z_h h(U) = z(U), \quad \sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} h'(U) = \widetilde{z}(U)$$

and

$$\sum_{h,h' \in \mathcal{H}} z_h \widetilde{z}_{h'} h(U) h'(U) = (\sum_{h \in \mathcal{H}} z_h h(U))(\sum_{h' \in \mathcal{H}} \widetilde{z}_{h'} h'(U)) = z(U)\widetilde{z}(U).$$

Thus,

$$\text{(E.1)} = \mathbb{E}_{U \sim \nu}\left[\frac{(z(U) - \widetilde{z}(U))^2}{\lambda(U)/\nu(U)}\right]$$

which proves the first equality. To prove the second, recall that $[h]_u = \nu(u)h(u)$, so $[z]_u = \sum_{h\in\mathcal{H}} z_h[h]_u = \nu(u)z(u)$. It follows that,

$$\|z-\widetilde{z}\|^2_{A(\lambda)^{-1}} = \sum_u \frac{\nu(u)^2}{\lambda(u)}(z(u)-\widetilde{z}(u))^2 = \mathbb{E}_{U\sim\nu}\left[\frac{(z(U)-\widetilde{z}(U))^2}{\lambda(U)/\nu(U)}\right]$$

which proves the second equality. □

E.1 Computational Efficiency of RAGE$^\epsilon$

RAGE$^\epsilon$ requires solving the optimization

$$\inf_{\lambda\in\triangle_{\mathcal{X}}}\max_{z\in\triangle_{\mathcal{H}}}\min_{\alpha\in\mathcal{A}} -c_a(\mathfrak{p}(-\widehat{\Delta}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell) + \alpha\|z-\widehat{y}_{\ell-1}\|^2_{A(\lambda)^{-1}} + \frac{\log(2|\mathcal{Z}|^2|\mathcal{A}|\ell^2/\delta)}{\alpha\tau}. \tag{E.2}$$

Here we take τ to be fixed, and recall that

$$\widehat{y}_\ell \leftarrow \arg\min_{y\in\mathcal{Y}}\min_{\alpha\in\mathcal{A}} \widetilde{R}^\alpha_\ell(y) - \widetilde{R}^\alpha_\ell(\widehat{y}_{\ell-1}) + 2\alpha\|y-\widehat{y}_{\ell-1}\|^2_{A(\lambda_\ell)^{-1}} + \frac{2\log(2|\mathcal{Z}|^2|\mathcal{A}|\ell^2/\delta)}{\alpha\tau_\ell}$$
$$\widehat{\Delta}^\ell(y) \leftarrow \min_{\alpha\in\mathcal{A}} \widetilde{R}^\alpha_\ell(y) - \widetilde{R}^\alpha_\ell(\widehat{y}_\ell) + \alpha\|y-\widehat{y}_\ell\|^2_{A(\lambda_\ell)^{-1}} + \frac{\log(2|\mathcal{Z}|^2|\mathcal{A}|\ell^2/\delta)}{\alpha\tau_\ell}.$$

Furthermore, $\mathcal{Y}$ will be a set of the form

$$\bigcup_{k=1}^{\ell'}\mathcal{Y}_k = \bigcup_{k=1}^{\ell'}\left\{z\in\mathcal{Z} \;:\; c(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(z)) + \max_{j\in[n]}\mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(z)) + \min_{j\in[n]}|\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|) \le \widehat{\Delta}^{i,k}_{\text{safe}}(h), \forall i\in[n]\right\}$$

Recall also that

$$\|h-h'\|^2_{A(\lambda)^{-1}} = \mathbb{E}_{U\sim\nu}\left[\frac{\mathbb{I}\{h(U)\neq h'(U)\}}{(9\lambda(U)/10+1/10d)/\nu(U)}\right] = \sum_{U\in\mathcal{X}}\frac{\nu(U)^2}{9\lambda(U)/10+1/10d}\mathbb{I}\{h(U)\neq h'(U)\}$$

and

$$\widetilde{R}^\alpha_\ell(h) = \frac{1}{\tau_\ell}\sum_{t=1}^{\tau_\ell}\frac{1}{w_t+\alpha}\mathbb{I}\{h(u_t)\neq v_t\}.$$

For $z\in\triangle_{\mathcal{H}}$, we denote $\widetilde{R}^\alpha_\ell(z) = \sum_{h\in\mathcal{H}} z_h\widetilde{R}^\alpha_\ell(h)$ and $\mathcal{R}(z;\alpha) = \sum_{h\in\mathcal{H}} z_h\mathcal{R}(h;\alpha)$. Finally, we assume that $\widehat{\Delta}_{\text{safe}}(z) = \min_{\alpha\in\mathcal{A}}\mathcal{R}(z;\alpha) + f(\alpha)$.

E.1.1 Solving for $\widehat{y}_\ell$

Using Lemma E.1, we can write the optimization for $\widehat{y}_\ell$ as

$$\min_{k\in[\ell']}\min_{y\in\mathcal{Y}_k}\min_{\alpha\in\mathcal{A}} \frac{1}{\tau_\ell}\sum_{t=1}^{\tau_\ell}\frac{1}{w_t+\alpha}\sum_{h\in\mathcal{H}} y_h\mathbb{I}\{h(u_t)\neq v_t\} + \alpha\sum_{h,h'\in\mathcal{H}} y_h\widehat{y}_{\ell-1,h'}\sum_{U\in\mathcal{X}}\frac{\nu(U)^2}{9\lambda_\ell(U)/10+1/10d}\mathbb{I}\{h(U)\neq h'(U)\}$$
$$- \widetilde{R}^\alpha_\ell(\widehat{y}_{\ell-1}) + \frac{\log(2|\mathcal{Z}|^2|\mathcal{A}|\ell^2/\delta)}{\alpha\tau_\ell}$$

We can rewrite

$$\sum_{h,h' \in \mathcal{H}} y_h \widehat{y}_{\ell-1,h'} \sum_{U \in \mathcal{X}} \frac{\nu(U)^2}{9\lambda_\ell(U)/10 + 1/10d} \mathbb{I}\{h(U) \neq h'(U)\} = \sum_{h \in \mathcal{H}} y_h \sum_{i=1}^{\|\widehat{y}_{\ell-1}\|_0 |\mathcal{X}|} w_i \mathbb{I}\{h(u_i) \neq v_i\}$$

for some weights w_i. It follows that if $\|\widehat{y}_{\ell-1}\|_0$ is polynomial in problem parameters then the optimization for $\widehat{y}_\ell$ can be written as

$$\min_{k \in [\ell']} \min_{y \in \mathcal{Y}_k} \min_{\alpha \in \mathcal{A}} \mathcal{R}(y; \alpha) + f(\alpha)$$

for $\mathcal{R}(y;\alpha)$ a CSC loss over only polynomially many points (as well as linear in y and convex in α), and $f(\alpha)$ convex in α. Note also that, for any y, we can upper bound $\mathcal{R}(y;\alpha) \leq \mathcal{O}(\frac{1}{\alpha} + d\alpha)$. Here $\mathcal{Y}_k$ a set of the form

$$\left\{ z \in \triangle_{\mathcal{H}} \ : \ \sum_{h \in \mathcal{H}} z_h c\Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|\Big) \leq \sum_{h \in \mathcal{H}} z_h \widehat{\Delta}^{i,k}_{\text{safe}}(h), \forall i \in [n] \right\}$$

$\widehat{y}_\ell$ will be the element in $\mathcal{Y}_k$ minimizing the, for the k achieving the minimum. The dual of this problem has the form

$$\min_{k \in [\ell']} \min_{z \in \triangle_{\mathcal{H}}} \min_{\alpha \in \mathcal{A}} \max_{\mu_i \geq 0, i \in [n]} \mathcal{R}(z;\alpha) + f(\alpha)$$
$$+ \sum_{i=1}^n \mu_i \left(\sum_{h \in \mathcal{H}} z_h c\Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|\Big) - \sum_{h \in \mathcal{H}} z_h \widehat{\Delta}^{i,k}_{\text{safe}}(h) \right).$$

Note that we can swap the min over α and z without issue. Furthermore, for a fixed μ, the objective is linear in z, and for a fixed z, the objective is linear in μ. By the minimax theorem, we can then swap the min and max to obtain the equivalent optimization:

$$\min_{k \in [\ell']} \min_{\alpha \in \mathcal{A}} \max_{\mu_i \geq 0, i \in [n]} \min_{z \in \triangle_{\mathcal{H}}} \mathcal{R}(z;\alpha) + f(\alpha)$$
$$+ \sum_{i=1}^n \mu_i \left(\sum_{h \in \mathcal{H}} z_h c\Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|\Big) - \sum_{h \in \mathcal{H}} z_h \widehat{\Delta}^{i,k}_{\text{safe}}(h) \right).$$

We can simply enumerate over k and α, as there are a finite number of each of these constraints. For a fixed k and α, to solve the inner maxmin problem, we can apply the approach proposed in Agarwal et al. [2018]. In particular, we alternate between running the exponential gradient algorithm for the μ player, and computing the best-response for the z player. The update to the μ player is trivial, as the problem is simply linear in μ (in practice, as in Agarwal et al. [2018], we will also upper bound the domain of μ_i by some value B, to ensure this is finite).

Computing the best-response for the z player (with μ fixed) is slightly trickier. Ignoring all other parameters, which are all currently fixed, the minimization over z can be written as

$$\min_{z \in \triangle_{\mathcal{H}}} \sum_{h \in \mathcal{H}} z_h \sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\}$$
$$+ \sum_{i=1}^n \mu_i \left(\sum_{h \in \mathcal{H}} z_h c\Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|\Big) - \sum_{h \in \mathcal{H}} z_h \widehat{\Delta}^{i,k}_{\text{safe}}(h) \right).$$

$$= \min_{z \in \triangle_{\mathcal{H}}} \sum_{h \in \mathcal{H}} z_h \Bigg(\sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)| - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg).$$

Now note that $\max_{j \in [n]} \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(z)) = \sup_{\widetilde{\lambda} \in \triangle_n} \sum_{j \in [n]} \widetilde{\lambda}_j \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(z))$, and similarly for $\min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(z)|$. Using this, we can rewrite the above optimization as

$$\min_{z \in \triangle_{\mathcal{H}}} \max_{\widetilde{\lambda}^{1h} \in \triangle_n, h \in \mathcal{H}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \sum_{h \in \mathcal{H}} z_h \Bigg(\sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \mathfrak{p}(\widehat{\Delta}^{k-1}(h)) + \sum_{j \in [n]} \widetilde{\lambda}^{1h}_j \mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)) + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)| - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg).$$

We also have:

$$\mathfrak{p}(-\widehat{\Delta}^{j,k-1}_{\text{safe}}(z)) = \max_{\beta \in [0,1]} -\beta \widehat{\Delta}^{j,k-1}_{\text{safe}}(z), \quad |\widehat{\Delta}^{j,k-1}_{\text{safe}}(z)| = \max_{\beta \in [-1,1]} \beta \widehat{\Delta}^{j,k-1}_{\text{safe}}(z).$$

So we can further simplify the above to:

$$\min_{z \in \triangle_{\mathcal{H}}} \max_{\widetilde{\lambda}^{1h} \in \triangle_n, h \in \mathcal{H}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \max_{\beta_1^h, \beta_2^{hj} \in [0,1], \beta_3^{hj} \in [-1,1], h \in \mathcal{H}} \sum_{h \in \mathcal{H}} z_h \Bigg(\sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta_1^h \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^{1h}_j \beta_2^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta_3^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg).$$

Note that the objective is linear in β and $\widetilde{\lambda}^2$, and both have continuous, compact, convex constraint sets, so we can swap the min and max to get that the above is equivalent to

$$\min_{z \in \triangle_{\mathcal{H}}} \max_{\widetilde{\lambda}^{1h} \in \triangle_n, h \in \mathcal{H}} \max_{\beta_1^h, \beta_2^{hj} \in [0,1], \beta_3^{hj} \in [-1,1], h \in \mathcal{H}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \sum_{h \in \mathcal{H}} z_h \Bigg(\sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta_1^h \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^{1h}_j \beta_2^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta_3^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg).$$

We can write this in the form

$$\min_{z \in \triangle_{\mathcal{H}}} \max_{\widetilde{\lambda}^1, \beta} g(z; \widetilde{\lambda}^1, \beta) \tag{E.3}$$

for

$$g(z; \widetilde{\lambda}^1, \beta) := \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \sum_{h \in \mathcal{H}} z_h \Bigg(\sum_t a_t \mathbb{I}\{h(u_t) \neq o_t\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta_1^h \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^{1h}_j \beta_2^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta_3^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg).$$

To solve this, we will apply a version of Frank-Wolfe that handles adversarial losses to the outer player (see Section 4.2 of Hazan and Kale [2012]), and will play best response for the inner player.

From the perspective of the outer player, at iteration t of the algorithm given in Hazan and Kale [2012], they must optimize the function

$$f_t(z) = g(z; \widetilde{\lambda}^1_t, \beta_t) = \sum_{h \in \mathcal{H}} z_h c_h(\widetilde{\lambda}^1_t, \beta_t)$$

for some $c_h(\widetilde{\lambda}^1_t, \beta_t)$. Note that this is $L = \max_h |c_h(\widetilde{\lambda}^1_t, \beta_t)|$ Lipschitz in the ℓ_1-norm, and that we can bound this L for all t by something like $\mathcal{O}(\frac{1}{\alpha} + d\alpha + n)$. The algorithm introduced in Section 4.2 of Hazan and Kale [2012] computes the standard FW update

$$\widetilde{z}_t = \arg\min_{z \in \triangle_{\mathcal{H}}} \nabla F_t(z_t)^\top z, \quad z_{t+1} = (1 - t^{-1/4}) z_t + t^{-1/4} \widetilde{z}_t$$

for

$$F_t(z) = \frac{1}{t} \sum_{\tau=1}^{t} \nabla f_\tau(z_\tau)^\top z + \sigma_t \|z - z_1\|_2^2$$

for $\sigma_t = (L/D)t^{-1/4}$ for $D = \max_{z_1, z_2 \in \triangle_{\mathcal{H}}} \|z_1 - z_2\|_1$ (note that in that work, the function seems to be Lipschitz in the ℓ_2 norm while here we use ℓ_1—this does not seem to change their result at all). It is shown in Hazan and Kale [2012] that running this procedure we obtain the bound, for any $z \in \triangle_{\mathcal{H}}$,

$$\sum_{t=1}^{T} (f_t(z_t) - f_t(z)) \leq 57 L D T^{3/4}.$$

It follows that if we are able to compute $\widetilde{z}_t$ efficiently, and if the max player plays best response (and the best response can be computed efficiently), using analysis similar to that in Agarwal et al. [2018], we can show that an approximate solution to (E.3) will be found in a polynomial number of iterations.

Computing the Best Response for $\widetilde{\lambda}^1, \beta$. For the inner player, they must solve

$$\max_{\widetilde{\lambda}^1, \beta} g(z_t; \widetilde{\lambda}^1, \beta).$$

Assume that $\|\widetilde{z}_t\|_0 \leq m$ for each t, and that $\|z_1\|_0 = 1$. Then z_t will be $(mt+1)$-sparse, so the sum in $g(z_t; \widetilde{\lambda}^1, \beta)$ will contain at most $(mt+1)$ values. Note that the optimization over β^h and $\widetilde{\lambda}^{1h}$ is completely independent, so to compute the best-response, we need to solve the following problem at most $(mt+1)$ times:

$$\max_{\widetilde{\lambda}^{1h} \in \triangle_n} \max_{\beta_1^h, \beta_2^{hj} \in [0,1], \beta_3^{hj} \in [-1,1]} \min_{\widetilde{\lambda}^{2h} \in \triangle_n} \sum_{i \in [n]} c_i \Big(\beta_1^h \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}_j^{1h} \beta_2^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) + \sum_{j \in [n]} \widetilde{\lambda}_j^{2h} \beta_3^{hj} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) \Big).$$

The optimization over the first two terms is trivial and can be solved by enumerating. The third term now is a maxmin problem, however, this can also be solved trivially as it is equivalent to $\min_{j \in [n]} |\widehat{\Delta}^{j,k-1}_{\text{safe}}(h)|$. Note that each of these gap terms is themself the solution to an optimization over $\alpha \in \mathcal{A}$, but that can be solved easily for each (since there are at most polynomial of them), so they can be regarded as constants.

Thus, we conclude that the best response for $\widetilde{\lambda}^1, \beta$ can be computed efficiently, assuming that m is polynomial in problem parameters. Note that the values of β^h and $\widetilde{\lambda}^{1h}$ do not matter for $h \notin \text{support}(z_t)$ do not matter to compute the best response, so we can set them to the same value for all $h \notin \text{support}(z_t)$.

Computing $\widetilde{z}_t$. It remains to show that we can efficiently find a near-optimal $\widetilde{z}_t$ such that $\|\widetilde{z}_t\|_0 \leq m$. The optimization for $\widetilde{z}_t$ will have the form

$$\widetilde{z}_t = \arg\min_{z \in \triangle_{\mathcal{H}}} \sum_{\tau=1}^{t} \nabla f_\tau(z_\tau)^\top z + 2\sigma_t (z_t - z_1)^\top z$$

for

$$\begin{aligned}
[\nabla f_\tau(z_\tau)]_h &= c_h(\widetilde{\lambda}^1_\tau, \beta_\tau) \\
&= \min_{\widetilde{\lambda}^{2h} \in \triangle_n} \sum_j a_j \mathbb{I}\{h(u_j) \neq o_j\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta^h_{1\tau} \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^{1h}_{j\tau} \beta^{hj}_{2\tau} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) \\
&\quad + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta^{hj}_{3\tau} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big).
\end{aligned}$$

Let $C_t \subseteq \mathcal{H}$ denote the classifiers supported on z_t and assume that z_1 is only supported on a single classifier h_0. Note from our discussion on computing the best-response for the $\widetilde{\lambda}^1$ and β player, we have that β^h and $\widetilde{\lambda}^{1h}$ are identical for all $h \notin C_t$. We can therefore rewrite the above objective as (dropping the τ subscript and denoting, e.g. $\beta^h_1 = \sum_{\tau=1}^t \beta^h_{1\tau}$):

$$\begin{aligned}
&\min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \sum_{h \in \mathcal{H} \setminus C_t} z_h \Bigg(\sum_j a_j \mathbb{I}\{h(u_j) \neq o_j\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta_1 \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^1_j \beta^j_2 \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) \\
&\quad + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta^j_3 \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) \Bigg) \\
&\quad + \sum_{h \in C_t} \Bigg(\sum_j a_j \mathbb{I}\{h(u_j) \neq o_j\} + \sum_{i \in [n]} c_i \Big(\epsilon_k + \beta^h_{1\tau} \widehat{\Delta}^{k-1}(h) - \sum_{j \in [n]} \widetilde{\lambda}^{1h}_{j\tau} \beta^{hj}_{2\tau} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) \\
&\quad + \sum_{j \in [n]} \widetilde{\lambda}^{2h}_j \beta^{hj}_{3\tau} \widehat{\Delta}^{j,k-1}_{\text{safe}}(h) - \widehat{\Delta}^{i,k}_{\text{safe}}(h) \Big) + 2\sigma_t z_t \Bigg) - 2\sigma_t z_{h_0}.
\end{aligned}$$

We will focus first on the sum over $\mathcal{H} \setminus C_t$. Note that $\widehat{\Delta}^{k-1}(h)$ and $\widehat{\Delta}^{j,k}_{\text{safe}}(h)$ are both of the form

$$\min_{\alpha \in \mathcal{A}} \sum_t \frac{1}{w_t + \alpha} \mathbb{I}\{h(u_t) \neq o_t\} + \alpha \sum_t \widetilde{w}_t \mathbb{I}\{h(u_t) \neq o_t\} + \frac{c}{\alpha}.$$

Given this, we can rewrite the minimization over the first term as (where the $\widetilde{\alpha}$ correspond to the gaps that have negative coefficients, which is where the max comes from):

$$\min_{z \in \triangle_{\mathcal{H}}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \min_{\alpha^h \in \mathcal{A}^k, h \in \mathcal{H}} \max_{\widetilde{\alpha}^h \in \mathcal{A}^k, h \in \mathcal{H}} \sum_{h \in \mathcal{H} \setminus C_t} z_h \Big(\mathcal{R}(h; \alpha^h, \widetilde{\alpha}^h, \widetilde{\lambda}^{2h}) + f(\alpha^h, \widetilde{\lambda}^{2h}) + g(\widetilde{\alpha}^h, \widetilde{\lambda}^{2h}) \Big)$$

for $\mathcal{R}$ convex in α, and concave in $\widetilde{\alpha}$, f convex in α, and g concave in $\widetilde{\alpha}$, and all functions are linear in $\widetilde{\lambda}^2$. Normally $\mathcal{A}$ is a discrete set, but if we let $\widetilde{\mathcal{A}}$ be a continuous relaxation of it, we can rewrite the above as

$$\min_{z \in \triangle_{\mathcal{H}}} \max_{\widetilde{\alpha}^h \in \mathcal{A}^k, h \in \mathcal{H}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \min_{\alpha^h \in \mathcal{A}^k, h \in \mathcal{H}} \sum_{h \in \mathcal{H} \setminus C_t} z_h \Big(\mathcal{R}(h; \alpha^h, \widetilde{\alpha}^h, \widetilde{\lambda}^{2h}) + f(\alpha^h, \widetilde{\lambda}^{2h}) + g(\widetilde{\alpha}^h, \widetilde{\lambda}^{2h}) \Big).$$

To solve this we can again apply the FW algorithm of Hazan and Kale [2012] with the max player playing best-response. As before, as long as $\mathfrak{z}_t$ (where $\mathfrak{z}_t$ denotes the update for this inner optimization) is sparse, we can efficiently compute the best-response for the $\widetilde{\alpha}$ player, since we only need to compute it for $h \in \mathfrak{z}_t$. The FW-style update will then have the form

$$\min_{z \in \triangle_{\mathcal{H}}} \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \min_{\alpha^h \in \mathcal{A}^k, h \in \mathcal{H}} \sum_{h \in \mathcal{H} \setminus C_t} z_h \left(\mathcal{R}(h; \alpha^h, \widetilde{\alpha}_t^h, \widetilde{\lambda}^{2h}) + f(\alpha^h, \widetilde{\lambda}^{2h}) + g(\widetilde{\alpha}_t^h, \widetilde{\lambda}^{2h}) \right)$$
$$= \min_{\widetilde{\lambda}^{2h} \in \triangle_n, h \in \mathcal{H}} \min_{\alpha^h \in \mathcal{A}^k, h \in \mathcal{H}} \min_{h \in \mathcal{H} \setminus C_t} \mathcal{R}(h; \alpha^h, \widetilde{\alpha}_t^h, \widetilde{\lambda}^{2h}) + f(\alpha^h, \widetilde{\lambda}^{2h}) + g(\widetilde{\alpha}_t^h, \widetilde{\lambda}^{2h})$$

where the equality follows since we can always swap min, and since there will always be an optimal solution supported on a single h. We can solve the inner min using a CSC oracle that is able to optimize over a set $\mathcal{H} \setminus C_t$, and by enumerating $\widetilde{\lambda}^2$ and α (since we can always find an optimal solution supported on a single h, we can set $\widetilde{\lambda}^{2h}, \alpha^h$ identical for all h and will arrive at the same minimum).

This will converge in polynomially many steps, and will produce some $\mathfrak{z}_{t'}$ which is m-sparse (for m polynomial in parameters). It follows that $\mathfrak{z}_{t'}$ is the near-optimal value for $\widetilde{z}_t$ supported on $\mathcal{H} \setminus C_t$. To pick a final value for $\widetilde{z}_t$, we can simply enumerate over the (polynomially many) $h \in C_t$, compute their loss values, and then pick the minimum out of those and the value achieved by $\mathfrak{z}_{t'}$. This procedure will always return some $\widetilde{z}_t$ supported on at most polynomially many h, so m can be chosen suitably to make the best-response of the max player efficient.

Putting all of this together, we can efficiently solve for $\widehat{y}_\ell$.

E.1.2 Solving for λ_ℓ

We turn now to solving the optimization (E.2). Using arguments similar to what we have already shown, we have that

$$\text{(E.2)} = \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{z \in \mathcal{Z}} \min_{\alpha \in \mathcal{A}, \alpha_2, \ldots, \alpha_p \in \mathcal{A}} \max_{\beta_1, \ldots, \beta_m \in \mathcal{B}} \mathcal{R}(z; \alpha, \alpha_2, \ldots, \alpha_p, \beta_1, \ldots, \beta_m)$$
$$+ 2\alpha \sum_{U \in \mathcal{X}} \frac{\nu(U)^2}{9\lambda(U)/10 + 1/10d} \mathbb{I}\{z(U) \neq \widehat{z}_{\ell-1}(U)\} + f(\alpha, \alpha_2, \ldots, \alpha_p, \beta_1, \ldots, \beta_m).$$

As before, we can simply enumerate over all possible choices of α and β. For a fixed setting of α and β, to solving the inf over λ, we can apply Mirror Descent. In this case we choose the mirror map to be the negative entropy, which is strongly convex with respect to the ℓ_1 norm.

Given this, to solve this in a computationally efficient manner, all we need is that the objective is convex (which it is) and Lipschitz with respect to the ℓ_1 norm. Let $g(\lambda)$ denote the objective of the above optimization. By the Mean Value Theorem,

$$|g(\lambda) - g(\widetilde{\lambda})| = \nabla g((1-c)\lambda + c\widetilde{\lambda})^\top (\lambda - \widetilde{\lambda})$$

for some $c \in [0, 1]$. So, for any $\lambda, \widetilde{\lambda} \in \triangle_{\mathcal{X}}$, we can bound

$$|g(\lambda) - g(\widetilde{\lambda})| \leq \left(\sup_{\lambda' \in \triangle_{\mathcal{X}}} \|\nabla g(\lambda')\|_\infty \right) \cdot \|\lambda - \widetilde{\lambda}\|_1.$$

We have,

$$\frac{d}{dt} \sum_{U \in \mathcal{X}} \frac{\nu(U)^2}{9\lambda(U)/10 + 1/10d + 9t\lambda_0(U)/10} \mathbb{I}\{z(U) \neq \widehat{z}_{\ell-1}(U)\}|_{t=0}$$

$$= \sum_{U \in \mathcal{X}} \frac{-\lambda_0(U)\nu(U)^2}{(9\lambda(U)/10 + 1/10d)^2} \mathbb{I}\{z(U) \neq \widehat{z}_{\ell-1}(U)\}.$$

It follows that

$$\sup_{\lambda' \in \triangle_{\mathcal{X}}} \|\nabla g(\lambda')\|_\infty \leq 100d^2$$

so we can apply Mirror Descent to optimize the above with computational complexity scaling only polynomially in problem parameters.

E.2 Computational Efficiency of BESIDE

The primary computational cost of BESIDE is incurred by calling RAGE$^\epsilon$, and solving the optimization on Line 6 of Algorithm 1. We have already shown that RAGE$^\epsilon$ can be run in a computationally efficient manner. The optimization on Line 6 has a form very similar to the optimization we solve in RAGE$^\epsilon$, so the same argument and solution approach (applying Mirror Descent) allows us to compute the optimal distribution, λ_ℓ, here as well.

F Experimental details and additional results

F.1 Experimental details

All code was written in Python and run on a Intel Xeon 6226R CPU with 64 cores.

Algorithm 6 is the precise implementation of BESIDE using elimination. It largely resemble to Algorithm 1, with the difference that it explicitly eliminates arms.

F.2 Additional results

We evaluate Algorithm 2 and the passive baseline on two other datasets. Recall that the passive baseline selects points uniformly at randoms from the pool of examples $\mathcal{X}$ and then retrains the model using the same Constrained Empirical Risk Minimization oracle (`CERM`).

Half circle dataset. We consider a two-dimensional half circle dataset, visualized on Figure 7. We report in Figures 8 and 9 the precision and (respectively) the recall obtained when varying the number of labels given to each method. The confidence intervals are obtained over 25 repetitions. We observe that Algorithm 2 allows us to provide a classifier satisfying a given recall or precision in far fewer queries. This is in line with the results of Jain and Jamieson [2020] on One Dimensional Thresholds, where the sample complexity of the active strategy is $O(log(n))$ while the sample complexity of the passive strategy is at least of order n.

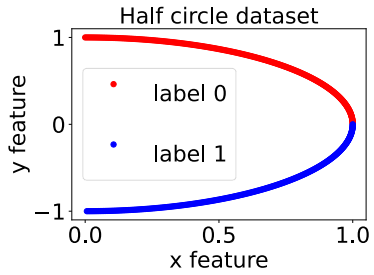

Figure 7: Half circle dataset.

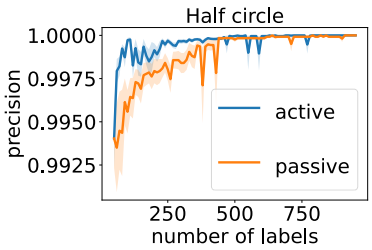

Figure 8: Precision

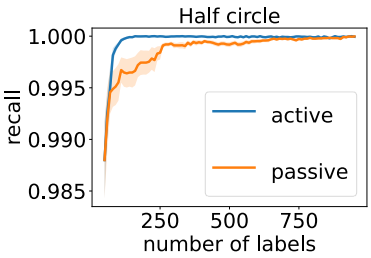

Figure 9: Recall

Algorithm 5 **Be**st **S**afe Arm **Id**entification (BESIDE, defined with generic constants)

input: tolerance ϵ, confidence δ

$\iota_\epsilon \leftarrow \lceil \log(\frac{2}{\min\{c_3,c_4\}\cdot\epsilon}) \rceil$, $\widehat{\Delta}^0_{\text{safe}}(z) \leftarrow 0, \widehat{\Delta}^0(z) \leftarrow 0$ for all $z \in \mathcal{Z}$

for $\ell = 1, 2, \ldots, \iota_\epsilon$ **do**

  $\epsilon_\ell \leftarrow \frac{2}{\min\{c_3,c_4\}} \cdot 2^{-\ell}$

// Solve experiment to reduce uncertainty on safety constraints

  Let τ_ℓ be the minimal value of $\tau = 2^j \geq 4 \log \frac{4m|\mathcal{Z}|\ell^2}{\delta}$ such that the objective to the following is no greater than $c_e\epsilon_\ell$, and λ_ℓ the corresponding optimal distribution

$$\inf_{\lambda \in \triangle_\mathcal{X}} \max_{z \in \mathcal{Z}} -c_d \left(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) + \epsilon_\ell \right) + \sqrt{\frac{\|z\|^2_{A(\lambda)^{-1}} \cdot \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}{\tau}}$$

  Sample $x_t \sim \lambda_\ell$, collect τ_ℓ observations $\{(x_t, r_t, s_{t,1}, \ldots, s_{t,m})\}_{t=1}^{\tau_\ell}$

  $\{\widehat{\mu}^{i,\ell}\}_{i=1}^m \leftarrow \mathsf{RIPS}(\{(x_t, s_{t,i})\}_{t=1}^{\tau_\ell}, \mathcal{Z}, \frac{\delta}{2m\ell^2})$ // Estimate safety constraints

  $\widehat{\Delta}^{i,\ell}_{\text{safe}}(z) \leftarrow \gamma - z^\top \widehat{\mu}^{i,\ell} + \|z\|_{A(\lambda_\ell)^{-1}} \sqrt{\tau_\ell^{-1} \log(\frac{4m|\mathcal{Z}|\ell^2}{\delta})}$ // Safety gap estimates

// Form set of arms guaranteed to be safe

$$\mathcal{Y}_\ell \leftarrow \Big\{ z \in \mathcal{Z} \ : 8c_d \Big(\min_j |\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell-1}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\ell-1}(z)) \Big) + 8(c_d + c_e)\epsilon_\ell \leq \widehat{\Delta}^{i,\ell}_{\text{safe}}(z), \forall i \in [n] \Big\} \cup \mathcal{Y}_{\ell-1}$$

// Refine estimates of optimality gaps

  $\{\widehat{\Delta}^\ell(z)\}_{z\in\mathcal{Z}} \leftarrow \text{RAGE}^\epsilon\Big(\mathcal{Z}, \mathcal{Y}_\ell, \epsilon_\ell, \frac{\delta}{4\ell^2}, \{\widehat{\Delta}_{\text{safe}}(z) \leftarrow \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\ell}_{\text{safe}}(z))\}_{z\in\mathcal{Z}}\Big)$

// Form set of arms guaranteed to be at most ϵ-unsafe

$$\mathcal{Y}_{\text{end}} \leftarrow \Big\{ z \in \mathcal{Z} \ : 8c_d \Big(\min_j |\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)| + \max_j \mathfrak{p}(-\widehat{\Delta}^{j,\iota_\epsilon}_{\text{safe}}(z)) + \mathfrak{p}(\widehat{\Delta}^{\iota_\epsilon}(z)) \Big) + 8(c_d + c_e)\epsilon - c_g\epsilon \leq \widehat{\Delta}^{i,\iota_\epsilon}_{\text{safe}}(z), \forall i \in [n] \Big\}$$

// Find ϵ-good arm out of ϵ-safe arms

  $\{\widehat{\Delta}^{\text{end}}(z)\}_{z\in\mathcal{Y}_{\text{end}}} \leftarrow \text{RAGE}^\epsilon(\mathcal{Y}_{\text{end}}, \mathcal{Y}_{\text{end}}, \epsilon, \delta)$

return $\widehat{z} = \arg\min_{z\in\mathcal{Y}_{\text{end}}} \widehat{\Delta}^{\text{end}}(z)$

Algorithm 6 Best Safe Arm Identification with Elimination

input: tolerance ϵ, confidence δ
$\iota_\epsilon \leftarrow \lceil \log(\frac{1}{\epsilon}) \rceil$, $\mathcal{Z}^0_{\text{active}} \leftarrow \mathcal{Z}$, $\mathcal{Z}^0_{\text{safe}} \leftarrow \emptyset$
for $\ell = 1, 2, \ldots, \iota_\epsilon$ **do**
    $\epsilon_\ell \leftarrow 2^{-\ell}$
    Compute allocation $\mathcal{XY}_{\text{safe}}$ on $\mathcal{Z}^{\ell-1}_{\text{active}}$ and sample from it $\tau_\ell = \mathcal{O}(\mathcal{XY}_{\text{safe}}(\mathcal{Z}^{\ell-1}_{\text{active}})/\epsilon_\ell^2)$ times
    $\widehat{\mu}^\ell \leftarrow \mathsf{RIPS}(\{(x_t, s_{t,i})\}_{t=1}^{\tau_\ell}, \mathcal{Z}, \frac{\delta}{2\ell^2})$
    Set $\widehat{\Delta}^\ell_{\text{safe}}(z) \leftarrow \gamma - z^\top \widehat{\mu}^\ell$ for all $z \in \mathcal{Z}^{\ell-1}_{\text{active}}$ and

$$\widetilde{Z}^\ell_{\text{active}} = \{z \in \widetilde{Z}^{\ell-1}_{\text{active}} \ : \ \widehat{\Delta}^\ell_{\text{safe}}(z) \in [-\epsilon_\ell, 2\epsilon_\ell]\} \quad \widetilde{Z}^\ell_{\text{safe}} = \{z \in \widetilde{Z}^{\ell-1}_{\text{active}} \ : \ \widehat{\Delta}^\ell_{\text{safe}}(z) \geq 2\epsilon_\ell]\}$$

    $\mathcal{Z}^\ell_{\text{active}}, \mathcal{Z}^\ell_{\text{safe}} \leftarrow \textsc{Rage-elim}^\epsilon\Big(\widetilde{Z}^\ell_{\text{active}} \cup \widetilde{Z}^\ell_{\text{safe}} \cup \mathcal{Z}^{\ell-1}_{\text{safe}}, \widetilde{Z}^\ell_{\text{safe}} \cup \mathcal{Z}^{\ell-1}_{\text{safe}}, \epsilon_\ell\Big)$
$\mathcal{Z}_{\text{final}}, \emptyset \leftarrow \textsc{Rage-elim}^\epsilon\Big(\mathcal{Z}^\ell_{\text{active}} \cup \mathcal{Z}^\ell_{\text{safe}}, \mathcal{Z}^\ell_{\text{active}} \cup \mathcal{Z}^\ell_{\text{safe}}, \epsilon_\ell\Big)$
return Any arm in $\mathcal{Z}_{\text{final}}$.

Algorithm 7 $\textsc{Rage-elim}^\epsilon$

input: active set $\mathcal{Z}$, optimal set $\mathcal{Y}$, tolerance ϵ
$\iota_\epsilon \leftarrow \lceil \log(\frac{1}{\epsilon}) \rceil$, $\mathcal{Z}^0 \leftarrow \mathcal{Z}$, $\mathcal{Y}^0 \leftarrow \mathcal{Y}$
for $\ell = 1, 2, \ldots, \iota_\epsilon$ **do**
    $\epsilon_\ell \leftarrow 2^{-\ell}$
    Compute allocation $\mathcal{XY}_{\text{diff}}$ on $(\mathcal{Z}^{\ell-1} \cup \mathcal{Y}^{\ell-1}, \mathcal{Y}^{\ell-1})$ and sample from it $\tau_\ell = \mathcal{O}(\mathcal{Z}^{\ell-1} \cup \mathcal{Y}^{\ell-1}, \mathcal{Y}^{\ell-1})/\epsilon_\ell^2)$ times
    $\widehat{\theta}^\ell \leftarrow \mathsf{RIPS}(\{(x_t, s_{t,i})\}_{t=1}^{\tau_\ell}, \mathcal{Z}, \frac{\delta}{2\ell^2})$
    Set $\widehat{\Delta}^\ell(z) \leftarrow \max_{y \in \mathcal{Y}^{\ell-1}} y^\top \widehat{\theta}^\ell - z^\top \widehat{\theta}^\ell$ for all $z \in \mathcal{Z} \cup \mathcal{Y}$ and

$$\mathcal{Z}^\ell = \{z \in \mathcal{Z}^{\ell-1} \ : \ \widehat{\Delta}^\ell(z) \leq \epsilon_\ell\} \quad \mathcal{Y}^\ell = \{y \in \mathcal{Y}^{\ell-1} \ : \ \widehat{\Delta}^\ell(y) \leq \epsilon_\ell]\}$$

return $\mathcal{Z}^\ell, \mathcal{Y}^\ell$