Thresholded Linear Bandits

by

Trang Thu Nguyen
B.Sc., University of Victoria, 2021

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Thresholded Linear Bandits

by

Trang Thu Nguyen
B.Sc., University of Victoria, 2021

Supervisory Committee

Dr. N. Mehta, Supervisor
(Computer Science)

Dr. Y. Lu, Departmental Member
(Computer Science)

ABSTRACT

Thresholded linear bandits is a novel bandit problem that lies in the intersection of several important multiarmed bandit (MAB) variants, including active learning, structured bandits, and learning halfspaces. To achieve sublinear regret in the presence of exponentially many arms, one method is to exploit the structure of the reward function. However, the presence of an unknown threshold component makes previously known algorithms for structured bandits unsuitable. Moreover, the threshold introduces a discontinuity to the reward function, making the problem significantly more difficult. In this thesis, we study the union of axis-parallel halfspace variant of the thresholded linear bandits problem. We suggest an algorithm that achieves sublinear regret and provide theoretical guarantees on the performance of the algorithm.

Contents

ACKNOWLEDGEMENTS

The work presented here would not have been possible without the guidance of Dr. Nishant Mehta, whose mentorship has been instrumental in shaping my journey as a researcher. Thank you for challenging me and believing in my potential. Thank you to Dr. Yun Lu, for taking the time to improve my thesis. I want thank my friends and labmates for the countless hours of discussions and laughs. Many chairs were pushed, countless cups of espresso were consumed, and numerous whiteboard markers sacrificed in the making of this thesis.

A special appreciation goes to my loved ones, for always urging me to move forward and to never stop learning.

Chapter 1

Introduction

Multi-armed bandits (MAB) is a learning paradigm used to solve sequential decision making problems, also known as online learning problems, where data arrives sequentially, and the learner (or algorithm) must make decisions iteratively. A poor decision can be costly, so the learner must develop a reliable policy for decision making. A common metaphor for this problem is the slot machine (colloquially known as the one-armed bandit), where the player must pull a lever to receive a reward. Each machine has a probability of success, and the player must decide which machine to pull to maximize their cumulative reward. Typically, the player does not know the probability of success for each machine, and it is not feasible to test all possible machines enough times to estimate these probabilities. Sequential decision making problems are frequently encountered in practical settings, making MAB algorithms highly applicable to a wide range of applications, from recommender systems (Mary et al., 2015) to clinical trials (Durand et al., 2018) and dynamic pricing (Misra et al., 2019). Moreover, many problems of both practical and theoretical importance involve choosing the best configuation or a set of parameters to achieve a specific goal.

The MAB framework can be viewed as a game between the learner and the environment played over T rounds. At each round $t \in \{1, ..., T\}$, the learner chooses an arm (also known as an action) A_t from a set $\mathcal{A}$, and the environment returns a stochastic reward $X_t \sim \mathcal{P}_{A_t}$ where $\mathcal{P}_{A_t}$ is not known to the learner. Unlike in the full-information setting, the learner receives information associated with only the pulled arm at the end of each round. In the stochastic setting, the learner's performance can be evaluated by its pseudo-regret, which quantifies the difference between the expected cumulative reward of the best constant strategy and the expected cumulative reward of the learner's strategy. Naturally, one goal is to minimize the cumulative pseudo-regret. If the reward distribution of every action were known, the optimal strategy would be to choose the action with

the highest mean reward. However, perfect knowledge is unrealistic and often unavailable, so the learner must choose from its current assessment of the actions.

Energy demand optimization This thesis introduces a novel problem setting called Thresholded Linear Bandits, which intersects several important MAB variants, including learning halfspaces and structured bandits. To motivate the reader, consider an energy management system where the goal is to optimize the use of renewable energy sources (e.g., solar panels, wind turbines, etc.) to meet certain energy demands. Here, the set of energy sources, represented as $\boldsymbol{A} \in [0,1]^d$, can be adjusted by tuning each of the d continuous features, such as the output levels of each source. Each source has a fixed cost per unit of energy produced, represented by the cost vector $\boldsymbol{v} \in \mathbb{R}^d_+$. The system's performance is evaluated based on its ability to meet energy demands during peak hours while minimizing the cost of energy production.

Additional demand satisfaction that can be achieved when more resources are deployed to handle peak demands, improving overall system efficiency. Demand satisfaction is captured as follows:

- p_0: Represents the fraction of demand met by the baseline energy production.
- $\Delta = p_1 - p_0$: Represents the potential increase in demand satisfaction during peak hours. The energy output $\langle \boldsymbol{\theta}^*, \boldsymbol{A} \rangle$ needs to be at least τ to prevent grid instability, for an unknown parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$.
- $1 - p_1$: Represents the fraction of demand that the system cannot meet due to limitations in grid infrastructure or production capacity.

During peak hours, the system may need to draw on stored energy or incorporate more expensive sources of energy to meet demands. If the electric grid cannot maintain a stable balance between generation and consumption, it becomes unstable, which can lead to power outages or damage to sensitive equipment. Although demand can stochastic from the customer's perspective– changing throughout the day, as noted in Jain and Gujar (2020)– it can be viewed from an energy management perspective. Specifically, demand can be viewed as the demand on a grid, representing the overall consumption on the grid. Having a fixed threshold helps companies better plan for financial costs and manage grid stability. The assumption is that, on average, the energy demand during peak hours remains relatively constant. In this scenario, efficiency is achieved as soon as the energy production is within the positive halfspace $\{\boldsymbol{A} : \langle \boldsymbol{\theta}^*, \boldsymbol{A} \rangle \geq \tau\}$. More specifically, the demand satisfaction jumps from p_0 to p_1.

The key challenge lies in finding the most cost-effective configuration. What is the strategy if the energy provider does not know the exact values of τ or Δ? Although the energy provider could maximize the energy production, this comes with the undesirable trade-off of high costs. When considering the influx of appliance and electronic uses during peak hours, one management strategy is to produce energy in discete units (such as megawatt-hours), and then later consumed in continuous quantities. Thus, modeling the demand using a discrete demand function is appropriate. Moreover, the jump in demand introduces discontinuity in the function, which introduces unique challenges for arm selection and regret minimization. To address this, we introduce a new framework called thresholded linear bandits. This variant of MAB incorporates a reward structure that is piecewise constant, with a jump discontinuity at the threshold (see Figure 1.1 for an example in a simple one-dimensional setting).

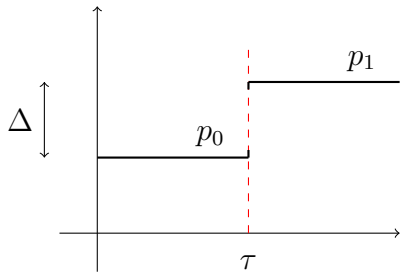

Figure 1.1: A one-dimensional example of the reward function. The discontinuity in the reward function is a key characteristic of thresholded linear bandits.

To informally illustrate the thresholded linear bandit framework, consider the one-dimensional case of the above energy optimization problem, where $A \in [0, 1]$ and $\tau \in [0, 1]$ partitions the search space into two regions: $\{A : A \geq \tau\}$ and $\{A : A < \tau\}$. Each arm's stochastic reward depends on which region the arm falls into. Since selecting an arm incurs a cost, modeled as a linear cost function ($v \cdot A$), the goal is identify the arm that optimally balances reward and cost. As discussed in Chapter 3, the optimal arm is either arm 0 (which incurs no cost) or an arm on the threshold hyperplane (in this example, arm τ).

This thesis focuses on a generalization of this problem in d dimensional space, where a union of axis-parallel halfspaces partitions the search space into a positive and a negative region. This structure creates a positive region that is non-convex, making it non-trivial to determine which MAB strategies are applicable. Previous work by Mehta et al. (2023) introduced both a one dimensional case and a multidimensional case with a single hyperplane in the thresholded linear bandit setting. This thesis extends on this by considering a case of a union of axis-parallel halfspaces. While the

halfspace structure suggests a connection to learning halfspaces, the presence of a cost makes naive exploration inefficient. In Chapter 3, we show that learning the parameters of the halfspaces is not necessary to find the optimal arm.

Alternatively, the thresholded linear bandit problem can be viewed as a structured bandit problem, where decisions leverage structural relationships between arms to improve the overall sampling strategy. However, unlike typical structured bandit settings, the decision problem in the thresholded linear bandit is focused on identifying the threshold where the trade-off between reward and cost is optimal. We propose an algorithm that leverages the structural properties of the thresholded linear bandit to efficiently identify an optimal arm while also determining whether the arm lies in the positive region.

Our contributions are as follows:

- We propose an algorithm called AxisParallelLeftist that extends the single-hyperplane results in Mehta et al. (2023) to handle the union of axis-parallel halfspaces problem variant, where multiple hyperplanes must be considered simultaneously.
- In Section 3.2.1 show that the fundamental relationship between Δ/v and τ characterizes the optimal arm structure, and we prove that AxisParallelLeftist leverages this property to determine whether to explore positive arms or commit to the minimum-cost negative arm.
- We provide high probability correctness guarantees for all components of our algorithm: AxisParallelLeftist, DimensionReduction, and ParallelNoisyBinarySearch.
- We establish sublinear regret bounds for AxisParallelLeftist. We provide an instance-dependent regret that grows logarithmically with T and is inversely proportional to Δ, as well as a worst-case regret $O(\log(T)\sqrt{dT})$ where d is the dimensionality.

The rest of this thesis is organized as follows. In Chapter 2, we briefly discuss some related works. We formalize the general problem setting of thresholded linear bandits and the union of axis-parallel halfspaces problem variant in Chapter 3. In Chapter 4, we present our algorithm and provide theoretical guarantees. This is followed by the regret analysis in Chapter 5, and we conclude with a discussion in Chapter 6.

Chapter 2

Related Work

In this section, we review relevant work in the field, focusing on algorithms designed to minimize regret.

In the classical bandit problem (Lai and Robbins, 1985), a learner repeatedly selects from a set of arms and receives rewards drawn according to an unknown distribution associate with the arm. The learner's goal is to maximize the total rewards over time by balancing exploration and exploitation. Popular strategies such as Upper Confidence Bound (UCB, (Auer et al., 2002)) and Thompson Sampling (Thompson, 1933; Agrawal and Goyal, 2012) provide effective approaches for balancing this trade-off. UCB maintains an upper confidence bound for each arm's empirical mean. Initially, the algorithm samples all arms to construct a confidence bound based on the empirical estimates of the mean rewards. Since an arm's mean reward is bounded by this confidence bound with high probability, the algorithm subsequently selects the arm with the highest upper bound. As more arms are explored, the empirical estimates of the means converge to their true values, thereby limiting exploration of suboptimal arms. In contrast, Thompson sampling employs a Bayesian approach, maintaining a posterior distribution over the mean reward of each arm. The algorithm begins with a prior distribution for each arm's reward and selects an arm by sampling from its posterior distribution. Both approaches encourage exploration in the early stages and exploitation once sufficient information is gathered. However, when the number of arms exceeds the number of rounds, naive implementations of these methods become suboptimal, as they require each arm to be pulled at least once. Furthermore, classical UCB and Thompson sampling algorithms are incapable of exploiting structural information, as shown by Lattimore and Szepesvari (2017). In UCB, optimism fails when the information gained about other actions is more relevant to the regret. Similarly, in Thompson sampling, suboptimal actions have an extremely small probability of being

chosen, but the structural information is not exploited. This highlights the need for more refined, data-dependent exploration strategies.

To achieve sublinear regret in the presence of exponentially or infinitely many arms, one approach is to assume a structure on the reward function. In the structured bandit setting, rewards from one arm can reveal information about the expected rewards of another, which a learner can exploit for faster convergence. In linear bandits, the expected reward of each arm is a linear function of an unknown parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^d$ that defines the linear relationship between the arms and their expected rewards. Worked by Abbasi-Yadkori et al. (2011) and Rusmevichientong and Tsitsiklis (2010) leverage this linear structure to construct a confidence ellipsoid around $\boldsymbol{\theta}^*$, achieving sublinear regret bounds of $O(d\sqrt{T}\log(T))$. However, this approach is not directly applicable to thresholded linear bandits, where the reward function is discontinuous. Furthermore, we argue that learning the parameters of the reward function is not necessary for identifying the optimal arm. Instead, our approach leverages the known cost function to guide exploration. The generalized linear bandits (Filippi et al., 2010; Jun et al., 2017) extends the structured bandit framework, where the expected reward of each arm is a function of a parameter vector $\boldsymbol{\theta}$ through a known link function. While this framework can model more complex relationships between the dimensions of the arms and their expected reward, it relies on the smoothness of the link function and thus cannot be immediately apply to thresholded linear bandits.

Lipschitz bandits are particularly useful in continuous-armed bandit problems, assuming that the mean reward function is Lipschitz continuous. This condition ensures that the reward function's rate of change is bounded, allowing the learner to generalize knowledge from explored arms to nearby unexplored ones. Kleinberg et al. (2008) leverages this property to reduce the need for exhaustive exploration. However, these Lipschitz-based approaches are not directly applicable to the thresholded linear bandits setting, where continuity is not guaranteed.

Under a unimodality assumption, Yu and Mannor (2011) deployed a method that achieves a regret of $O(\sqrt{T}\log T)$. The unimodality assumption assumes that there exists a unique optimal arm $A^* \in \mathbb{R}$ where the expected reward increases monotonically as the learner approaches A^* and decreases moving away from it in any direction. Unlike Lipschitz bandits that rely on smoothness, this approach uses a weaker Lipschitz condition. Specifically, it assumes a fixed minimum constant $D \in [0, 1]$ between the rewards of any two neighboring arms. This weaker assumption allows for efficient detection of changes in the reward without relying on the smoothness of the reward function. In contrast, thresholded linear bandits lacks this property and presents a more challenging problem. Proximity to other arms may not provide additional information about changes in the reward. Specifically, arms within the same region share the same mean reward, making the problem

more challenging. Furthermore, the thresholded linear bandits setting introduces additional complexity by requiring the learner to make decisions based on a reward threshold, necessitating new algorithmic designs.

We briefly mention a seemingly related problem, dynamic pricing (Cesa-Bianchi et al., 2019). While p_0 and p_1 in the thresholded linear bandits settings can be characterized as unknown demand components, the notion of cost is different. Dynamic pricing optimizes the price of a product over time to maximize revenue, with the inherent cost of exploration being the potential loss in revenue from suboptimal pricing. In contrast, in thresholded linear bandits the cost of exploration arises from the expense of producing (or selecting) an arm. Naive exploration of costly arms may lead to costs exceeding the potential reward gained from pulling an arm above the threshold.

Chapter 3

Thresholded Linear Bandits

We now formalize the general problem setting of thresholded linear bandit and the union of axis-parallel halfspaces problem variant. We also briefly mention the union of general halfspaces, the intersection of axis-parallel halfspaces, and the intersection of general halfspaces problem variants at the end of this chapter.

3.1 General Problem Setting

Let $\mathcal{A} = [0,1]^d$ be a continuous set of arms. In each round $t \in [T]$, the learner chooses an arm $\boldsymbol{A_t} \in \mathcal{A}$ and receives a stochastic revenue X_{A_t}. Each decision incurs a cost $c_t = \langle \boldsymbol{v}, \boldsymbol{A_t} \rangle$ according to a known cost vector $\boldsymbol{v} \in (0,\infty)^d$. We assume that for each arm $\boldsymbol{A}$, its stochastic revenue in each round is independent and identically distributed according to a Bernoulli distribution with an unknown success probability p_0 or p_1, such that $0 \leq p_0 \leq p_1 \leq 1$, and mean μ_A. We also assume that the behavior of the *expected revenue* function μ is specified by some *unknown* monotone increasing function $f : \mathbb{R}^d \to \{0,1\}$ given by,

$$\mu(\boldsymbol{A}) = p_0 + (p_1 - p_0) \cdot f(\boldsymbol{A}). \tag{3.1}$$

The key idea here is that $\mu(\boldsymbol{A})$ is either p_0 or p_1 depending on the output of f. Furthermore, we say arm $\boldsymbol{A}$ is *positive* if $f(\boldsymbol{A}) = 1$ and *negative* if $f(\boldsymbol{A}) = 0$. The following defines f formally.

Definition 1. A function f is a monotone increasing function if for all $\boldsymbol{A}, \boldsymbol{A'} \in \mathcal{A}$ we have $f(\boldsymbol{A}) \leq f(\boldsymbol{A'})$ when $A_i \leq A'_i$ for all $i \in [d]$.

Observe that pulling a positive arm guarantees an immediate expected revenue equal to or greater than the revenue of a negative arm. The difference in expected revenue is the probability gap, which, to foreshadow its importance, we define as

$$\Delta := p_1 - p_0.$$

Finally, we account for the cost and define the *expected reward* as

$$\mu_c(\boldsymbol{A_t}) = \mu(\boldsymbol{A_t}) - c_t = p_0 + \Delta \cdot f(\boldsymbol{A_t}) - \langle \boldsymbol{v}, \boldsymbol{A_t} \rangle \tag{3.2}$$

and an optimal arm as

$$\boldsymbol{A^*} = \underset{\boldsymbol{A} \in \mathcal{A}}{\operatorname{argmax}}\, \mu_c(\boldsymbol{A}).$$

Our performance measure is the cumulative pseudo-regret (hereafter referred to simply as regret), which is

$$R_T = T\mu_c(\boldsymbol{A^*}) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_c(\boldsymbol{A_t})\right] \tag{3.3}$$

and the learner's objective is to minimize the regret. In the following sections, we define the function f for each problem instance and describe what an optimal arm would look like.

3.2 Union of Axis-Parallel Halfspaces

Before we present the union of axis-parallel halfspaces, consider the case of the union of general halfspaces. Given a collection of d unknown halfspaces, the boundary of each halfspace is a hyperplane H_i for $i \in [d]$ and is given by the equation

$$H_i := \{\boldsymbol{A} \in \mathbb{R}^d : \langle \boldsymbol{\theta_i}, \boldsymbol{A} \rangle = \tau_i\}, \tag{3.4}$$

where, for all $i \in [d]$, $\boldsymbol{\theta_i} \in \mathbb{R}^d_+ \setminus \{\boldsymbol{0}\}$ is an *unknown* nonnegative normal vector and $\tau_i > 0$ is an *unknown* threshold. Note that for all i, the sign of $\langle \boldsymbol{\theta_i}, \boldsymbol{A} \rangle - \tau_i$ is unchanged when multiplied by any positive constant, so without loss of generality, we assume that $||\boldsymbol{\theta_i}||_2 = 1$. Next, the function f is defined as

$$f(\boldsymbol{A}) = \mathbb{1}[\exists i \in [d] : \langle \boldsymbol{\theta_i}, \boldsymbol{A} \rangle \geq \tau_i] \tag{3.5}$$

by definition of a union of halfspaces. Furthermore, substituting the above function (3.5) into the expected reward (3.2) gives us

$$\mu_c(\boldsymbol{A}) = p_0 + \Delta \cdot \mathbb{1}[\exists i \in [d] : \langle \boldsymbol{\theta_i}, \boldsymbol{A} \rangle \geq \tau_i] - \langle \boldsymbol{v}, \boldsymbol{A} \rangle \,. \tag{3.6}$$

To intuit what an optimal arm looks like, first observe that the function (3.5) gives us a piecewise constant expected revenue function over two regions, a positive region and a negative region. Then, since the expected revenue is constant in each space, the optimal arm must be a minimum-cost arm in one of the two spaces. If the optimal arm is negative, it is easy to see from the expected reward function (3.6) that the minimum-cost arm is arm $\mathbf{0}$. If the optimal arm is positive, the minimum-cost arm must belong to at least one separating hyperplane $H_1, ..., H_d$. We prove this by way of contradiction. Let the optimal arm $\boldsymbol{A}^*$ be a positive arm in the interior of the positive space. Since, for all $i \in [d]$, we have $\langle \boldsymbol{v}, \boldsymbol{\theta_i} \rangle > 0$ then, if we move $\boldsymbol{A}^*$ infinitesimally in the direction of $-\boldsymbol{\theta_i}$, we will simultaneously decrease the cost and the resulting arm will still be positive. Therefore, $\boldsymbol{A}^*$ is not the minimum cost arm, and we have a contradiction. Thus, positive arms inside $[0, 1]^d$ cannot be the minimum-cost arm unless they belong to at least one hyperplane. Let $\mathcal{H} = \bigcup_{i \in [d]} H_i$ be the union of the axis-parallel hyperplanes. We can characterize the set of potentially optimal arms as follows.

Proposition 2. *The optimal arm $\boldsymbol{A}^*$ belongs to the set $(\mathcal{H} \cap [0, 1]^d) \cup \{\mathbf{0}\}$. In particular, $\boldsymbol{A}^* \in \mathcal{H}$ if $\min_{A \in \mathcal{H} \cap [0,1]^d} \langle \boldsymbol{v}, \boldsymbol{A} \rangle \leq \Delta$ and arm $\mathbf{0}$ is optimal otherwise.*

3.2.1 One-dimensional Case

Next, we briefly consider the one-dimensional case to illustrate a key property that we extend to the union of axis-parallel halfspaces setting. Let $d = 1$ and $v \in \mathbb{R}^+$. Since $||\boldsymbol{\theta}||_2 = 1$, it implicitly follows that $\boldsymbol{\theta} = 1$. Then, for this case, we have

$$\mu_c(A) = p_0 + \Delta \cdot \mathbb{1}[A \geq \tau] - v \cdot A, \qquad H = \{A \in \mathbb{R} : A = \tau\}, \quad \text{and} \qquad \mathcal{H} = \{\tau\}.$$

Thus, by Proposition 2, the set of potential optimal arms is $\{0, \tau\}$.

Corollary 3. *The optimal arm A^* belongs to the set $\{0, \tau\}$. Arm τ is optimal if $\tau \leq \frac{\Delta}{v}$ and arm 0 is optimal if $\tau \geq \frac{\Delta}{v}$.*

This gives us a key property; the identity of the optimal arm depends on Δ/v and τ. In essence, Δ/v represents a return on investment, while τ represents the minimum-cost investment. If $\tau >$

Δ/v, the investment cost exceeds the return on investment, and arm 0 is optimal. Conversely, if the return exceeds the cost, arm τ (in this case) is a worthwhile investment. In fact, during exploration, our algorithm leverages this property and uses a lower bound approximation on Δ to determine the most favorable course of action.

3.2.2 Union of Axis-Parallel Halfspaces

From Proposition 2, we know that the optimal arm must either belong to the union of hyperplanes $\mathcal{H}$ or is arm $\mathbf{0}$. Should the former be true, then $\boldsymbol{A^*}$ must be a 1-sparse vector, and we only need to find the threshold of one hyperplane. The difficulty is knowing which one. Our approach reduces the multidimensional case into simpler, one-dimensional subproblems and reduces cost by limiting exploration to only the coordinate axes. To illustrate this, consider an arbitrary dimension $i \in [d]$, and let H_i be its i^{th} axis-parallel hyperplane. By definition of axis-parallel hyperplane, we have $\boldsymbol{\theta_i} = \boldsymbol{e_i}$, where $\boldsymbol{e_i}$ is the i^{th} standard basis vector. It is easy to see that the arm on the hyperplane is arm $\boldsymbol{e_i} \cdot \tau_i$. For clarity, we will denote $\boldsymbol{e_i} \cdot \tau_i$ as $\boldsymbol{\tau^i}$. Since this holds for all dimensions, the set of arms belonging to $\mathcal{H}$ is the set $\{\boldsymbol{\tau^1}, \boldsymbol{\tau^2}, ..., \boldsymbol{\tau^d}\}$, which are all arms on the coordinate axes. Thus, we reduce the cost of exploring suboptimal positive arms by restricting the search space to the coordinate axes.

Next, let a dimension i^* be defined as

$$i^* = \underset{i \in [d]}{\operatorname{argmin}} \langle v, \boldsymbol{\tau^i} \rangle \qquad \text{such that} \qquad \tau_i \leq \Delta / v_i \quad \text{and} \quad \tau_i \leq 1. \tag{3.7}$$

Then, by Corollary 3, we have that arm $\boldsymbol{\tau^{i^*}}$ is optimal if $\tau_{i^*} \leq \frac{\Delta}{v_{i^*}}$ and the next corollary follows.

Corollary 4. *The optimal arm $\boldsymbol{A^*}$ belongs to the set $\{\mathbf{0}, \boldsymbol{\tau^{i^*}}\}$. Arm $\boldsymbol{\tau^{i^*}}$ is optimal if $\tau_{i^*} \leq \frac{\Delta}{v_{i^*}}$ and arm $\mathbf{0}$ is optimal if $\tau_{i^*} \geq \frac{\Delta}{v_{i^*}}$.*

3.3 Extensions

Here we describe some additional thresholded linear bandit problem variants and highlight the differences from the union of axis-parallel halfspaces variant.

Union of general halfspaces. In this setting, the parameter vectors $\boldsymbol{\theta_i}$ are no longer restricted to the coordinate axes, meaning the minimum cost positive arm may have nonzero components

along multiple dimensions. This makes the search space more complex than the axis parallel case, where the problem could be reduced to a one-dimensional search along the coordinate axes and may require an entirely different algorithm all together.

Intersection of general halfspaces. The key difference in this setting is that the region of positive arms is convex due to the nature of intersections. Unlike in the union case, an arm must satisfy multiple conditions simultaneously to be positive. Here, the function f is defined as

$$f(\boldsymbol{A}) = \mathbb{1}[\forall i \in [k] : \langle \boldsymbol{\theta_i}, \boldsymbol{A} \rangle \geq \tau_i],$$

where $k > 0$ is the number of halfspaces. The structural property of convexity can be leveraged for optimization, allowing for uses of methods such as projected gradient descent or convex hull estimation. However, if $k > 0$ is unknown, identifying the positive region is non-trivial.

Intersection of axis-parallel halfspaces. While restricting the halfspaces to be axis-parallel simplifies the problem, it still presents a challenge due to unknown number of halfspaces. The intersection structure here might allow for a similar greedy strategy as the multidimensional case in Mehta et al. (2023). In particular, we can use a similar geometrically decreasing exploration strategy at first, then switch to a gradient-based approach once the boundary is identified. However, we leave a detailed exploration of this method for future work.

Chapter 4

Axis-Parallel Leftist

In this section, we present the main algorithm of this thesis and provide the algorithmic guarantees. We begin by highlighting two significant challenges in the thresholded linear bandit setting to motivate our approach. The first challenge is efficiently learning Δ with limited feedback. When comparing two arms, the learner only observes the difference in their rewards, which can be noisy. The second challenge is the cost of exploration. Since each arm incurs a cost, naive exploration can lead to excessive costs, especially when the number of arms is large. Furthermore, a simple binary search with a fixed number of comparisons may be inefficient since Δ is unknown, and the number of samples required to differentiate a positive arm from a negative arm depends on Δ.

4.1 Notation and Overview

We propose AxisParallelLeftist (APL, Algorithm 1), an epoch-based algorithm that operates in three stages. The algorithmic decisions made in each stage depend on an estimate of Δ, where $0 \leq \Delta \leq 1$, which is computed by pulling a right arm and a left arm $n > 0$ number of times.

In the first stage, AxisParallelLeftist determines whether to search for the minimum-cost positive arm $\boldsymbol{\tau}^{i^*}$ or commit to arm $\mathbf{0}$. The algorithm's behaviour is governed by several variables that change geometrically with each epoch r, namely the number of pulls $n_r = O\left(\frac{\log(1/\delta)}{\varepsilon_r^2}\right)$ where δ is the failure probability, the precision $\varepsilon_r = 2^{-r-3}$, and arm scalar

$$a_{r,i} = \begin{cases} 1 & \text{if } v_i < 8\varepsilon_r \\ 2^{\lfloor \log_2(1/v_i) \rfloor - r} & \text{if } v_i \geq 8\varepsilon_r \end{cases}$$

where v_i is the i^{th} component of the cost vector v. At each epoch $r \geq 0$, AxisParallelLeftist makes n_r pulls to a right arm $\boldsymbol{A_{r,i}} = a_{r,i} \cdot e_i$ and left-arm $\mathbf{0}$ for each dimension $i \in [d]$, and computes an estimate of Δ

$$\hat{\Delta}_{r,i} = \hat{p}_1 - \hat{p}_0, \tag{4.1}$$

where $\hat{p}_1$ and $\hat{p}_0$ are the empirical means of arm $\boldsymbol{A_{r,i}}$ and arm $\mathbf{0}$, respectively.

To control the regret, we set a threshold on the number of suboptimal arm draws, ensuring the algorithm stops exploring if this limit is reached. This strategy ensures a minimax regret of $O(\log(T)\sqrt{dT})$ as proved in Chapter 5. Since pulling arms with costs exceeding 1 can lead to high cumulative regret, AxisParallelLeftist focuses on pulling arms with costs at most Δ. Specifically, for dimensions where $v_i > 1$, we scale the initial arm scalar as

$$a_{0,i} = \frac{1}{v_i}, \qquad \text{for all } i \in [d] \text{ where } v_i > 1. \tag{4.2}$$

This ensures that the cost of each pull to arm $\boldsymbol{A_{r,i}}$ is at most 1. While this approach may forgo rewards of Δ in these dimensions, it prevents the risk of incurring regret greater than Δ.

The first stage is divided into two phases: Phase 1 (*non-halving*) and Phase 2 (*halving*). In Phase 1, where the cost for dimension i, v_i, satisfies $v_i < 8\varepsilon_r$, the scalar $a_{r,i}$ remains fixed. Phase 2 begins when $v_i \geq 8\varepsilon_r$, at which point the algorithm halves $a_{r,i}$ after each epoch to adapt to the increased number of pulls needed to approximate Δ. More specifically, the number of samples required to differentiate a positive arm from a negative arm is proportional to $1/\Delta^2$, according to Hoeffding's inequality. By halving $a_{r,i}$, the cumulative cost of pulling arm $\boldsymbol{A_{r,i}}$ is limited to approximately $1/\Delta$, ensuring that exploration cost is proportional to the potential revenue gain. This process continues until either a lower bound for Δ is detected or, if Δ is too small, the algorithm commits to arm $\mathbf{0}$ for the remaining rounds.

In the second stage, AxisParallelLeftist invokes the algorithms DimensionReduction (DR, Algorithm 2) and ParallelNoisyBinarySearch (PNBS, Algorithm 3) to refine the search for the optimal arm $\boldsymbol{A}^*$. If a lower bound on Δ is detected, this information determines the number of pulls required to approximate arm $\boldsymbol{A}^*$ with high probability in the second stage. A *knowledge set* $S_{\text{dim}} := \{1, ..., d\}$, which will be the set of possible dimensions containing an optimal arm, is updated by DimensionReduction and passed to ParallelNoisyBinarySearch. DimensionReduction eliminates dimensions from the knowledge set when its right arm is identified as negative with high probability. This process is crucial, as subsequent statistical guarantees rely on the positivity of the right arms.

The next step is to identify a positive arm that is close in cost to the optimal arm to ensure algorithmic correctness and near-optimality. To achieve this, ParallelNoisyBinarySearch uses a standard Noisy Binary Search (NBS) on each dimension in S_{dim} to obtain a set of potentially optimal arms and returns a positive arm $\hat{\boldsymbol{\tau}}$ that is approximately $\varepsilon_{\text{pnbs}}$-close in cost to the optimal arm $\boldsymbol{\tau}^{i^*}$. The final stage runs UCB on arms $\hat{\boldsymbol{\tau}}$ and $\mathbf{0}$ until round T, achieving low regret against these two arms.

4.2 High Probability Guarantees

In this section, we analyze the algorithmic guarantees of the main algorithm and its subroutine algorithms. In particular, we first show that for any $i \in [d]$, if $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$, then with high probability, $\tau_i \in [0, a_{r,i}]$ for all epochs r. We show that DimensionReduction returns a set of dimensions containing i^*. Following that, we show that ParallelNoisyBinarySearch returns a positive arm $\hat{\boldsymbol{\tau}}$ whose cost is within some $\varepsilon_{\text{pnbs}}$ of arm $\boldsymbol{\tau}^{i^*}$.

4.2.1 Correctness analysis for AxisParallelLeftist

Algorithm 1: AxisParallelLeftist

Epoch $r \leftarrow 0$, $\varepsilon_0 \leftarrow \frac{1}{8}$, $n_0 \leftarrow \frac{\log \frac{2}{\delta}}{2\varepsilon_0^2}$
for $i \in [d]$ **do**
  if $v_i > 1$ **then**
    $a_{r,i} \leftarrow 1/v_i$
  else
    $a_{r,i} \leftarrow 1$
  $\hat{\Delta}_{r,i} \leftarrow T$
  $\boldsymbol{A_{r,i}} \leftarrow a_{r,i} \cdot \boldsymbol{e_i}$
`// Stage 1`
while $\varepsilon_r \geq \log(T)\sqrt{\frac{d}{T}}$ **do**
  for $i \in [d]$ **do**
    Make n_r pulls of arm **0** to get empirical mean $\hat{p}_0$
    Make n_r pulls of arm $\boldsymbol{A_{r,i}}$ to get empirical mean $\hat{p}_1$
    $\hat{\Delta}_{r,i} \leftarrow \hat{p}_1 - \hat{p}_0$
    if $\hat{\Delta}_{r,i} - \varepsilon_r \geq \varepsilon_r$ **then**
      `// Stage 2`
      $a_{r,1}, \ldots, a_{r,d}, S_{\text{dim}} \leftarrow$ DimensionReduction$(a_{r,1}, \ldots, a_{r,d}, \varepsilon_r)$
      $\hat{\boldsymbol{\tau}} \leftarrow$ Parallel NBS$(\varepsilon_r, a_{r,1}, \ldots, a_{r,d}, S_{\text{dim}})$
      `// Stage 3`
      Run UCB on $\{\mathbf{0}, \hat{\boldsymbol{\tau}}\}$ until time T
    else
      if $\langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \geq 8\varepsilon_r$ **then**
        `// Phase 2`
        $a_{r+1,i} \leftarrow a_{r,i}/2$
      else
        `// Phase 1`
        $a_{r+1,i} \leftarrow a_{r,i}$
  $\varepsilon_{r+1} \leftarrow \varepsilon_r/2$
  $n_{r+1} \leftarrow 4n_r$
  $r \leftarrow r + 1$
Commit to arm **0**

We introduce some epoch-related notation used throughout the analysis. For all $i \in [d]$, let epoch $r_{\mathrm{i},i}$ be the last epoch of Phase 1 for dimension i. More formally,

$$r_{\mathrm{i},i} := \operatorname*{argmin}_{r \geq 0} \{v_i \geq 8\varepsilon_r\}. \tag{4.3}$$

Let ρ be the stopping epoch of AxisParallelLeftist, where ρ is a random variable. It is either the epoch at which DimensionReduction is called or, if the former is never called, the largest possible epoch

$$r_{\max} := \operatorname*{argmax}_{r \geq 0} \left\{ \varepsilon_r \geq \log(T)\sqrt{\frac{d}{T}} \right\}. \tag{4.4}$$

Recall that if arm $\boldsymbol{\tau}^{i^*}$ is optimal and a lower bound on Δ is not detectable, the DimensionReduction is not called, and AxisParallelLeftist eventually commits to pulling arm $\mathbf{0}$ for the remainder of the rounds. Similarly, if arm $\mathbf{0}$ is optimal, then identifying a lower bound on Δ is not necessary, and AxisParallelLeftist converges to an arm with a cost of order $O(\log(T)\sqrt{d/T})$ before committing to pulling arm $\mathbf{0}$. In Chapter 5, we show that committing to arm $\mathbf{0}$ in either case incurs a regret of order $O(\sqrt{dT})$.

When arm $\boldsymbol{\tau}^{i^*}$ is optimal, we can show that with high probability, ρ is no greater than 3 epochs after the critical epoch r_Δ, defined as

$$r_\Delta := \operatorname*{argmax}_{r \geq 0} \{\varepsilon_r > \Delta\}. \tag{4.5}$$

In the following, we show that if a lower bound on Δ is detected, AxisParallelLeftist's stopping epoch is no greater than three epochs after r_Δ. We will need the following lemma for later.

Lemma 5. *If r_Δ exists, then for all epochs $r_\Delta + 1 + j$ where $j \geq 0$,*

$$\frac{\Delta}{2^j} \geq \varepsilon_{r_\Delta + 1 + j}$$

Proof. From the definition of r_Δ (4.5), at epoch $r_\Delta + 1$ we have $\Delta \geq \varepsilon_{r_\Delta+1}$. Since ε_r is halved after every epoch, the claim follows. □

We need to ensure that the lower confidence bound $\hat{\Delta}_{r,i} - \varepsilon_r$ used by AxisParallelLeftist is not too large. For this, we introduce an event that occurs when a positive arm is pulled. Let $\mathcal{E}_{\text{apl}}$ be the event that for any $i \in [d]$ and for all epochs $r \leq \rho$, if arm $\boldsymbol{A_{r,i}}$ is positive, then

$$|\hat{\Delta}_{r,i} - \Delta| \leq \varepsilon_r. \tag{4.6}$$

Lemma 6. *Let $r = r_\Delta + 3$. For any $i \in [d]$, if $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$, then on event $\mathcal{E}_{\text{apl}}$, the lower confidence bound of Δ is lower bounded as*

$$\frac{\Delta}{2} \leq \hat{\Delta}_{r,i} - \varepsilon_r.$$

Proof. Fix i and let $r = r_\Delta + 3$. If $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$, we want to know if the following inequality holds

$$\frac{\Delta}{2} \leq \hat{\Delta}_{r,i} - \varepsilon_r. \tag{4.7}$$

It is true that (4.7) is equivalent to

$$\frac{\Delta}{2} + 2\varepsilon_r \leq \hat{\Delta}_{r,i} + \varepsilon_r. \tag{4.8}$$

We will show that the right-hand side is lower bounded by Δ and the left-hand side is upper bounded by Δ, after which the above inequality follows.

Consider the right-hand side of the above inequality (4.8). For $\Delta \leq \hat{\Delta}_{r,i} + \varepsilon_r$ to hold, arm $\boldsymbol{A}^*$ must provide useful information. Specifically, when $a_{r,i} \geq \tau_i$, arm $\boldsymbol{A}^*$ is positive and contributes reliable information for approximating Δ. First, we know from the definition of r_Δ (4.5) that $\Delta < \varepsilon_{r_\Delta}$, and if $\tau_i \leq \frac{\Delta}{v_i}$ then

$$\tau_i \leq \frac{\Delta}{v_i} < \frac{\varepsilon_{r_\Delta}}{v_i}. \tag{4.9}$$

Next, recall that in Phase 1 we have

$$8\varepsilon_r > v_i \cdot a_{r,i}$$

Suppose that $r \leq r_{\text{i},i}$, then the critical epoch happens before the end of Phase 1. There are two possible cases.

- Case 1: $v_i \leq 1$ and $a_{r,i} = 1$. If $\tau_i \leq 1$ then we have $\tau_i \leq 1 = a_{r,i}$.
- Case 2: $v_i > 1$ and $a_{r,i} = 1/v_i \geq \Delta/v_i$ since $\Delta \in [0, 1]$. Thus we have $\tau_i \leq \frac{\Delta}{v_i} \leq a_{r,i}$.

Now, suppose that $r > r_{\text{i},i}$. We find that

$$v_i \cdot a_{r,i} \geq 8\varepsilon_r = \varepsilon_{r_\Delta},$$

where the first inequality uses the fact that $r > r_{\mathrm{i},i}$. Combining this with (4.9), we get

$$\tau_i < \frac{\varepsilon_{r_\Delta}}{v_i} \leq \frac{v_i \cdot a_{r,i}}{v_i}.$$

Hence, $\tau_i \leq a_{r,i}$, arm $\boldsymbol{A}^*$ is positive, and provides useful information. Therefore, if $\mathcal{E}_{\text{apl}}$ occurs, then $\hat{\Delta}_{r,i} + \varepsilon_r \geq \Delta$.

For the left-hand side of (4.8), we have $\varepsilon_r = \varepsilon_{r_\Delta+3}$, which, by Lemma 5, is at most $\frac{\Delta}{4}$, giving us

$$\frac{\Delta}{2} + 2\varepsilon_r \leq \frac{\Delta}{2} + \frac{\Delta}{2} = \Delta. \tag{4.10}$$

Combining the left- and right-hand sides, we get

$$\frac{\Delta}{2} + 2\varepsilon_r \leq \Delta \leq \hat{\Delta}_{r,i} + \varepsilon_r,$$

and so (4.8) holds, which implies that (4.7) holds. □

Next, we prove that AxisParallelLeftist will stop no later than epoch $r_\Delta + 3$.

Lemma 7. *Suppose that there exists some $i \in [d]$ such that $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$. If $\mathcal{E}_{\text{apl}}$ occurs, AxisParallelLeftist will stop no later than epoch $r_\Delta + 3$.*

Proof. It suffices to show that if AxisParallelLeftist reaches epoch $r_\Delta + 3$, then the algorithm stops. Fix i. Let $r = r_\Delta + 3$ and assume $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$. Suppose AxisParallelLeftist reaches epoch r but does not stop. That must mean $\hat{\Delta}_{r,i} - \varepsilon_r < \varepsilon_r$ occurred. Assuming event $\mathcal{E}_{\text{apl}}$ occurred, then we must have

$$\frac{\Delta}{2} \leq \hat{\Delta}_{r,i} - \varepsilon_r < \varepsilon_r,$$

where the first inequality is from Lemma 6. By Lemma 5, $\varepsilon_{r_\Delta+3} < \frac{\Delta}{4}$, and so

$$\frac{\Delta}{2} \leq \hat{\Delta}_{r,i} - \varepsilon_r < \frac{\Delta}{4},$$

which is a contradiction. □

Finally, we show that for any $i \in [d]$ where $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$, if event $\mathcal{E}_{\text{apl}}$ occurs then arm $\boldsymbol{A}^*$ is positive for all epochs $r \leq \rho$.

Corollary 8. *For any $i \in [d]$, if $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$, and $\mathcal{E}_{\text{apl}}$ occurs, for any epoch r, either $\tau_i \leq a_{r,i}$ or AxisParallelLeftist has stopped before this epoch.*

Proof. Fix i. Assume $\tau_i \leq \frac{\Delta}{v_i}$, $\tau_i \leq 1$ and that event $\mathcal{E}_{\text{apl}}$ occurred. We begin by considering the scenario in which r_Δ does not exist. Then, for all $r \geq 0$, we have $\Delta < \varepsilon_r$. Moreover, when $\tau_i \leq \frac{\Delta}{v_i}$, it follows that $\tau_i \to 0$ when $T \to \infty$. This implies that $\tau_i \leq a_{r,i}$, which, from the positivity of $a_{r,i}$, holds for all epochs $r \leq \rho$. Now consider the scenario in which r_Δ exists. From Lemma 7, we know that AxisParallelLeftist will stop no later than epoch $r_\Delta + 3$. Since $a_{r,i}$ is non-increasing, it suffices to show that $\tau_i \leq a_{r_\Delta+3,i}$. Suppose that $r_\Delta + 3 \leq r_{\text{i},i}$. Then, we have that $\tau_i \leq 1 = a_{r_\Delta+3,i}$ if $v_i \leq 1$ or $\tau_i \leq \Delta/v_i \leq a_{r_\Delta+3,i}$ if $v_i > 1$. Next, suppose that $r_\Delta + 3 > r_{\text{i},i}$. It follows that $8\varepsilon_{r_\Delta+3} \leq v_i \cdot a_{r_\Delta+3,i}$. Thus, by definition of r_Δ, we have $\Delta < \varepsilon_{r_\Delta}$ and can upper bound τ_i as

$$\begin{aligned}
\tau_i \leq \frac{\Delta}{v_i} &< \frac{\varepsilon_{r_\Delta}}{v_i} \\
&= 8 \cdot \frac{\varepsilon_{r_\Delta+3}}{v_i} \\
&\leq \frac{v_i \cdot a_{r_\Delta+3,i}}{v_i} \\
&= a_{r_\Delta+3,i},
\end{aligned}$$

which shows that $\tau_i < a_{r_\Delta+3,i}$. □

Lemma 9. *Take $\delta = 1/T^2$. The event $\mathcal{E}_{\text{apl}}$ holds with probability at least $1 - 1/T$.*

Proof. This is a direct proof using Hoeffding's inequality. Let $i \in [d]$ be arbitrary. For any epoch $r \leq \rho$ in which arm $\boldsymbol{A}^*$ is positive, Hoeffding's inequality gives that

$$\begin{aligned}
\Pr\left(|\hat{\Delta}_{r,i} - \Delta| > \varepsilon_r\right) &\leq 2e^{-2n_r\varepsilon_r^2} \\
&= \delta. \qquad (4.11)
\end{aligned}$$

Next, we upper bound the number of epochs for each dimension by T/d (and overcount) then using

a double union bound over all such epochs and all dimensions we get

$$\begin{aligned}\Pr\left(\exists i \in [d], \exists r \leq \rho : |\hat{\Delta}_{r,i} - \Delta| > \varepsilon_r\right) &< \Pr\left(\bigcup_{i=1}^{d}\bigcup_{r=1}^{T/d}\left\{|\hat{\Delta}_{r,i} - \Delta| > \varepsilon_r\right\}\right) \\ &\leq \sum_{i=1}^{d}\sum_{r=1}^{T/d}\Pr\left(|\hat{\Delta}_{r,i} - \Delta| > \varepsilon_r\right) \\ &\leq T\delta,\end{aligned}$$

where the final inequality is computed using (4.11). Thus, for $\delta = 1/T^2$, we can conclude that the event $\mathcal{E}_{\text{apl}}$ holds with probability at least $1 - 1/T$. □

While no algorithmic guarantees are made when arm **0** is optimal, we show in Chapter 5 that of AxisParallelLeftist is at most $O(\log(T)\sqrt{dT})$ in the worst case. For the subsequent sections, we establish the algorithmic guarantees under that assumption that arm $\boldsymbol{\tau}^{i^*}$ is optimal.

4.2.2 Correctness analysis for DimensionReduction

Algorithm 2: DimensionReduction

Input: scalars $a_{r,1}, ..., a_{r,d}$, lower bound $\varepsilon_r \leq \Delta$

$N \leftarrow 16 \log\left(\frac{2}{\delta}\right) / \varepsilon_r^2$

$\varepsilon' \leftarrow \varepsilon_r / 4$

$S_{\text{dim}} \leftarrow \{1, ..., d\}$

Make N pulls to arm $\mathbf{0}$ to get empirical mean $\hat{p}_0$

`// First elimination round`

for $i \in [d]$ **do**

  $\boldsymbol{R_i} \leftarrow a_{r,i} \cdot \boldsymbol{e_i}$

  Make N pulls to arm $\boldsymbol{R_i}$ to get empirical mean $\hat{p}_{\text{right},i}$

  $\bar{\Delta}_i \leftarrow \hat{p}_{\text{right},i} - \hat{p}_0$

  if $\bar{\Delta}_i - \varepsilon' < \varepsilon'$ **then**

    Remove i from S_{dim}

`// Scaling round`

$k \leftarrow \text{argmin}_{i \in S_{\text{dim}}}(v_i \cdot a_{r,i})$

for $i \in S_{dim}$ **do**

  $a_{r,i} \leftarrow a_{r,k} \cdot \frac{v_k}{v_i}$

`// Second elimination round`

for $i \in S_{dim}$ **do**

  $\boldsymbol{R'_i} \leftarrow a_{r,i} \cdot \boldsymbol{e_i}$

  Make N pulls to arm $\boldsymbol{R'_i}$ to get empirical mean $\hat{p}_{\text{right,i}}$

  $\bar{\Delta}_i \leftarrow \hat{p}_{\text{right},i} - \hat{p}_0$

  if $\bar{\Delta}_i - \varepsilon' < \varepsilon'$ **then**

    Remove i from S_{dim}

return $a_{r,i}, ..., a_{r,d}, S_{dim}$

When DimensionReduction is invoked, two parameters are passed in: a set of scalars $a_{r,1}, ..., a_{r,d}$, used to initialize the right arms $\boldsymbol{R_i} = a_{r,i} \cdot \boldsymbol{e_i}$ for each dimension, and ε_r, the lower confidence bound for Δ. The algorithm consists of three stages: a first elimination round, a scaling stage, and a second elimination round.

To analyze the guarantees of the first elimination round, we first establish a general property of any negative arm pulled for all epochs $r \leq \rho$. This will allow us to show that, with high probability, in the first elimination round, the event $\mathcal{E}_{\text{dr-1}}$ occurs, where:

1. If arm $\boldsymbol{R_i}$ is negative, then DimensionReduction eliminates dimension i.
2. If arm $\boldsymbol{R_i}$ is positive and $\tau_i \leq \Delta/v_i$, then DimensionReduction does not eliminate dimension i.

Lemma 10. *Assume event $\mathcal{E}_{\text{apl}}$ happened. For all $i \in [d]$ and for all epochs $r \leq \rho$, if arm $\boldsymbol{A^*}$ is negative, then $\tau_i > 1$ or $\tau_i > \Delta/v_i$.*

Proof. Fix an arbitrary dimension $i \in [d]$ and assume that event $\mathcal{E}_{\text{apl}}$ happened. Suppose $\tau_i \leq \Delta/v_i$ and $\tau_i \leq 1$. Corollary 8 implies that for any i satisfying $\tau_i \leq \Delta/v_i$ and $\tau_i \leq 1$, in all epochs $r \leq \rho$, arm $\boldsymbol{A^*}$ is positive. Therefore, arm $\boldsymbol{A^*}$ must be positive. If it is negative, then at least one of the premises must be wrong, in which case we either have $\tau_i > 1$ or $\tau_i > \Delta/v_i$. □

Next, we lower bound epoch ρ when DimensionReduction is called.

Lemma 11. *On event $\mathcal{E}_{\text{apl}}$, if the early stopping condition $\hat{\Delta}_{r,i} - \varepsilon_r \geq \varepsilon_r$ is satisfied, then $r_\Delta < \rho$.*

Proof. For all $r \leq \rho$, on event $\mathcal{E}_{\text{apl}}$, we have $\Delta \geq \hat{\Delta}_{r,i} - \varepsilon_r$ with probability at least $1 - \delta$. The early stopping condition $\hat{\Delta}_{r,i} - \varepsilon_r \geq \varepsilon_r$ further implies that $\Delta \geq \varepsilon_r$. This follows directly by combining the inequalities, which gives us

$$\Delta \geq \hat{\Delta}_{\rho,i} - \varepsilon_\rho \geq \varepsilon_\rho.$$

Thus, it must hold that $\Delta \geq \varepsilon_\rho$. Moreover, this is equivalent to

$$2^{-(\rho+3)} \leq \Delta,$$

which can be rewritten as

$$\rho \geq \lceil \log_2 \frac{1}{\Delta} \rceil - 3.$$

To complete the proof, it suffices to show that ρ occurs after the largest epoch in which $\varepsilon_r > \Delta$, which, by definition, is epoch r_Δ. We can rewrite the inequality $\varepsilon_{r_\Delta} > \Delta$ as

$$2^{-(r_\Delta+3)} > \Delta.$$

The above implies that

$$r_\Delta < \lceil \log_2 \frac{1}{\Delta} \rceil - 3.$$

Therefore, $\rho \geq \lceil \log_2 \frac{1}{\Delta} \rceil - 3 > r_\Delta$ and we conclude our proof. □

Finally, we show that $\mathcal{E}_{\text{dr-1}}$ holds with high probability.

Lemma 12. *Take $\delta = 1/T^2$. On event $\mathcal{E}_{\text{apl}}$, the event $\mathcal{E}_{\text{dr-1}}$ holds with probability at least $1 - 1/T$.*

Proof. Take $\delta = 1/T^2$ and let $i \in S_{\text{dim}}$ be arbitrary. Assume $\mathcal{E}_{\text{apl}}$ happened. Since the early stopping condition was satisfied, Lemma 7 and Lemma 11 imply that $r_\Delta + 1 \leq \rho \leq r_\Delta + 3$.

Let us first suppose arm $\boldsymbol{R_i}$ is negative. Lemma 10 implies that the dimension cannot contain $\boldsymbol{A^*}$. Moreover, since arm $\boldsymbol{R_i}$ and arm $\mathbf{0}$ are negative, their mean difference is zero. Hoeffding's inequality implies that $\bar{\Delta}_i - \varepsilon' < 0 < \varepsilon'$. Therefore, the first If statement evaluates to true with high probability and the algorithm eliminates dimension i.

Now, suppose arm $\boldsymbol{R_i}$ is positive and $\tau_i \leq \Delta / v_i$. For $N = 16 \log(2/\delta)/\varepsilon_r^2$, Hoeffding's inequality tells us that

$$|\bar{\Delta}_i - \Delta| \leq \varepsilon'$$

with probability at least $1 - \delta$ (see appendix A.1 for proof). Next, we leverage the ideas from Lemma 6 to lower bound $\bar{\Delta}_i - \varepsilon'$ and retrace the main steps for completeness. We want to know if $\frac{\Delta}{2} \leq \bar{\Delta}_i - \varepsilon'$ holds. It is true that the inequality is equivalent to

$$\frac{\Delta}{2} + 2\varepsilon' \leq \bar{\Delta}_i + \varepsilon'.$$

First, consider the left-hand side of the inequality. By Lemma 5 we have that $\varepsilon_\rho \leq \Delta$ for $\rho \geq r_\Delta + 1$, giving us

$$\frac{\Delta}{2} + 2\varepsilon' \leq \frac{\Delta}{2} + \frac{\Delta}{2} = \Delta.$$

On the right-hand side, we have $\Delta \leq \bar{\Delta}_i + \varepsilon'$, which follows from Hoeffding's inequality since arm $\boldsymbol{R_i}$ is positive. Combining the left-hand side and right-hand side gives us

$$\frac{\Delta}{2} + 2\varepsilon' \leq \Delta \leq \bar{\Delta}_i + \varepsilon',$$

and thus $\frac{\Delta}{2} \leq \bar{\Delta}_i - \varepsilon'$ holds. Moreover, since

$$\frac{\Delta}{2} \geq \frac{\varepsilon_\rho}{2} = 2\varepsilon' > \varepsilon',$$

then we have $\bar{\Delta}_i - \varepsilon' \geq \Delta/2 \geq \varepsilon'$. DimensionReduction's first **if** statement evaluates to false with high probability and the dimension is not eliminated.

Finally, applying a union bound over all d dimensions, the probability that DimensionReduction removes all dimensions with negative right-arms in the first elimination round is $1 - d\delta = 1 - d/T^2 \geq 1 - 1/T$. □

After the first elimination round, DimensionReduction scales the arms' costs relative to the lowest-cost dimension in the set S_{dim} so that for all $i \in S_{\text{dim}}$, the resulting scaled right arms become $\boldsymbol{R}'_{\boldsymbol{i}} = a_{r,i} \cdot \boldsymbol{e_i}$. Recall that $k = \text{argmin}_{i \in S_{\text{dim}}}(v_i \cdot a_{r,i})$, as defined by the algorithm. We note that, since all arms $\boldsymbol{R_i}$ for $i \in S_{\text{dim}}$ are positive after the first elimination round and $a_{r,k}$ remains unchanged, then arm $\boldsymbol{R}'_{\boldsymbol{k}}$ is positive. Furthermore, the cost of each scaled arm becomes $\langle \boldsymbol{v}, \boldsymbol{R}'_{\boldsymbol{i}} \rangle = a_{r,k} \cdot v_k$ which satisfies the following important property required for the subsequent elimination round.

Proposition 13. *For all $i \in S_{dim}$ such that $i \neq i^*$, if $\boldsymbol{R}'_{\boldsymbol{i}}$ is negative, then $\langle v, \boldsymbol{\tau^i} \rangle > \langle v, \boldsymbol{\tau^{i^*}} \rangle$.*

The proof for this is straightforward. Assume that arm $\boldsymbol{R}'_{\boldsymbol{j}}$ is positive. For all $i \neq j \in S_{\text{dim}}$, if arm $\boldsymbol{R}'_{\boldsymbol{i}}$ is negative, then

$$\langle \boldsymbol{v}, \boldsymbol{\tau^i} \rangle > \langle \boldsymbol{v}, \boldsymbol{R}'_{\boldsymbol{j}} \rangle \geq \langle \boldsymbol{v}, \boldsymbol{\tau^j} \rangle,$$

and the cost of arm $\boldsymbol{\tau^i}$ is strictly greater than the cost of arm $\boldsymbol{R}'_{\boldsymbol{j}}$.

Some right arms may become negative after scaling which poses a problem for ParallelNoisy-BinarySearch. To address this, we run a second elimination round to ensure that all arms $\boldsymbol{R}'_{\boldsymbol{i}}$ for $i \in S_{\text{dim}}$ are positive. We introduce one more event for DimensionReduction. Let event $\mathcal{E}_{\text{dr-2}}$ be the event that DimensionReduction returns a set of dimensions containing i^*.

Lemma 14. *Take $\delta = 1/T^2$. On event $\mathcal{E}_{\text{apl}}$ and event $\mathcal{E}_{\text{dr-1}}$, the event $\mathcal{E}_{\text{dr-2}}$ holds with probability at least $1 - 1/T$.*

Proof. The proof is nearly identical to Lemma 12, so we will be brief. Take $\delta = 1/T^2$. Let $i \neq k \in S_{\text{dim}}$ be arbitrary, where $k = \text{argmin}_{i \in S_{\text{dim}}}(v_i \cdot a_{r,i})$. We consider two cases based on whether arm $\boldsymbol{R}'_{\boldsymbol{i}}$ is negative or positive.

Suppose arm $\boldsymbol{R}'_{\boldsymbol{i}}$ is positive. Lemma 7 and Lemma 11 imply that $r_\Delta + 1 \leq \rho \leq r_\Delta + 3$. By Hoeffding's inequality, with probability at least $1 - \delta$, we have that $|\bar{\Delta}_i - \Delta| \leq \varepsilon'$. Then, at epoch

$r = \rho$, Lemma 5 ensures that

$$\bar{\Delta}_i - \varepsilon' \geq \Delta/2 \geq \varepsilon'.$$

Thus, the second If statement evaluates to false with high probability, and dimension i is not eliminated.

Now, suppose arm $\boldsymbol{R}'_{\boldsymbol{i}}$ is negative. Then $\bar{\Delta}_i - \varepsilon' < 0 < \varepsilon'$. The second If statement will evaluate to true, and DimensionReduction removes dimension i from S_{dim}. Moreover, Lemma 13 guarantees that $i \neq i^*$ in this case. Finally, applying a union bound over all d dimensions yields a probability of at least $1 - d\delta \geq 1 - 1/T$. □

4.2.3 Correctness analysis for ParallelNoisyBinarySearch

Algorithm 3: Parallel Noisy Binary Search

Input: Lower bound $\varepsilon_r \leq \Delta$, scalars $a_{r,i}$ for $i \in S_{\text{dim}}$, set of dimensions S_{dim}
$N \leftarrow 4\log(\frac{1}{\delta})/\ \varepsilon_r^2$
$\boldsymbol{R_i} \leftarrow a_{r,i} \cdot \boldsymbol{e_i}$ for all $i \in S_{\text{dim}}$
$\boldsymbol{L_i} \leftarrow \boldsymbol{0}$ for all $i \in S_{\text{dim}}$
Parallel for each $i \in S_{\text{dim}}$
  Make N pulls to arm $\boldsymbol{R_i}$ to get empirical mean $\hat{p}_{\text{right},i}$
  while $\langle \boldsymbol{v}, \boldsymbol{R_i} \rangle - \langle \boldsymbol{v}, \boldsymbol{L_i} \rangle \geq \varepsilon_{\text{pnbs}}$ **do**
    $\boldsymbol{M_i} \leftarrow \frac{\boldsymbol{R_i} + \boldsymbol{L_i}}{2}$
    Make N pulls to arm $\boldsymbol{M_i}$ to get empirical mean $\hat{p}_{\text{mid,i}}$
    if $\hat{p}_{\text{right},i} - \hat{p}_{\text{mid,i}} \geq \frac{\varepsilon_r}{2}$ **then**
      $\boldsymbol{L_i} \leftarrow \boldsymbol{M_i}$
    else
      $\hat{p}_{\text{right},i} \leftarrow \hat{p}_{\text{mid,i}}$
      $\boldsymbol{R_i} \leftarrow \boldsymbol{M_i}$
$\hat{i} = \text{argmin}_{i \in S_{\text{dim}}} \langle \boldsymbol{v}, \boldsymbol{R_i} \rangle$
$\hat{\boldsymbol{\tau}} \leftarrow \boldsymbol{R}_{\hat{\boldsymbol{i}}}$
return $\hat{\boldsymbol{\tau}}$

ParallelNoisyBinarySearch follows a straightforward strategy. Each right arm, $R_i = a_{\text{right},i} \cdot \boldsymbol{e_i}$, is initialized using the scalars passed in after DimensionReduction, while each left arm, $L_i = a_{\text{left},i} \cdot \boldsymbol{e_i}$, is initially set to arm $\boldsymbol{0}$. For each dimension $i \in S_{\text{dim}}$, the algorithm runs NoisyBinarySearch

to identify a positive arm $\hat{\boldsymbol{\tau}}$ whose cost is $\varepsilon_{\text{pnbs}}$-close to arm $\boldsymbol{\tau}^{i^*}$. At each step, if the middle arm $\boldsymbol{M_i} = a_{\text{mid},i} \cdot \boldsymbol{e_i}$ is positive, then the algorithm updates the right arm to be $\boldsymbol{M_i}$; otherwise, the left arm is updated to be $\boldsymbol{M_i}$.

Lemma 15. *Each NBS subroutine in ParallelNoisyBinarySearch runs for at most* $\lfloor \log_2 \frac{2}{\varepsilon_{\text{pnbs}}} \rfloor$ *epochs.*

Proof. For each dimension $i \in S_{\text{dim}}$, the cost difference between the right arm $\boldsymbol{R_i}$ and the left arm $\boldsymbol{L_i}$ is bounded by

$$\langle \boldsymbol{v}, \boldsymbol{R_i} - \boldsymbol{L_i} \rangle = v_i(a_{\text{right},i} - a_{\text{left},i}) \leq v_i \cdot a_{\text{right},i} \leq 1.$$

This follows from the initial scaling in AxisParallelLeftist, ensuring the cost of pulling any arm is at most 1.

At epoch q, halving the distance interval gives

$$\langle \boldsymbol{v}, \boldsymbol{R_i} - \boldsymbol{L_i} \rangle \leq v_i \cdot a_{\text{right},i} \cdot 2^{-q} \leq 2^{-q}.$$

ParallelNoisyBinarySearch stops when $\langle \boldsymbol{v}, \boldsymbol{R_i} - \boldsymbol{L_i} \rangle < \varepsilon_{\text{pnbs}}$. The smallest q where $2^{-q} < \varepsilon_{\text{pnbs}}$ is epoch $\lfloor \log_2 \frac{1}{\varepsilon_{\text{pnbs}}} \rfloor + 1$, limiting the number of epochs to at most $\lfloor \log_2 \frac{2}{\varepsilon_{\text{pnbs}}} \rfloor$. □

Next, we bound the probability that ParallelNoisyBinarySearch makes a mistake. For all dimensions $i \in S_{\text{dim}}$, there is a possibility of mislabeling arm $\boldsymbol{M_i}$ each time the **if** statement is reached. If a mistake is not made, then arm $\boldsymbol{R_i}$ is positive, and arm $\boldsymbol{L_i}$ is negative for the subsequent epoch, which is what we desire.

Lemma 16. *For any dimension* $i \in S_{dim}$*, consider an arbitrary epoch* q *in ParallelNoisyBinarySearch. Suppose that in epoch* q*, arm* $\boldsymbol{R_i}$ *is positive. On events* $\mathcal{E}_{\text{apl}}$*,* $\mathcal{E}_{\text{dr-1}}$*, and* $\mathcal{E}_{\text{dr-2}}$*, the probability that the algorithm misclassifies the label of arm* $\boldsymbol{M_i}$ *is at most* δ*.*

Proof. Fix a dimension $i \in S_{\text{dim}}$ and let epoch q be arbitrary. Assume that ParallelNoisyBinarySearch did not make any mistakes in the epochs prior to q. Then on events $\mathcal{E}_{\text{apl}}$, $\mathcal{E}_{\text{dr-1}}$, and $\mathcal{E}_{\text{dr-2}}$, arm $\boldsymbol{R_i}$ is positive and is correctly labeled. Moreover, in each iteration, arms $\boldsymbol{R_i}$ and $\boldsymbol{M_i}$ are fixed conditional on the samples drawn in the previous iterations.

When labeling $\boldsymbol{M_i}$, one of two mistakes can occur. Let $\bar{X}_{q,i} = \hat{p}_{\text{right,i}} - \hat{p}_{\text{mid,i}}$ be the difference in the estimated probabilities. The first mistake occurs when the algorithm thinks the labels are different $\bar{X}_{q,i} \geq \varepsilon_r/2$, but the true expected difference is $\mathbf{E}[\bar{X}_{q,i}] = 0$. The second mistake occurs when it thinks the labels are the same $\bar{X}_{q,i} < \varepsilon_r/2$, but the true expected difference is $\mathbf{E}[\bar{X}_{q,i}] = \Delta$.

We bound the probability of making such mistakes using Hoeffding's inequality, as follows. First, the probability of falsely thinking that the labels are different is bounded by

$$\begin{aligned}\Pr\left(\bar{X}_{q,i} \geq \frac{\varepsilon_r}{2} \,|\, \mathbf{E}[\bar{X}_{q,i}] = 0\right) &= \Pr\left(\bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] \geq \frac{\varepsilon_r}{2} \,|\, \mathbf{E}[\bar{X}] = 0\right) \\ &\leq e^{\frac{-2(2N)(\varepsilon_r/2)^2}{2^2}} \\ &= e^{-N\varepsilon_r^2/4}.\end{aligned}$$

Similarly, we show the probability of failing to detect that the labels are different is also exponentially small. The probability is bounded by

$$\begin{aligned}\Pr\left(\bar{X}_{q,i} < \frac{\varepsilon_r}{2} \,|\, \mathbf{E}[\bar{X}_{q,i}] = \Delta\right) &= \Pr(\bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] < \frac{\varepsilon_r}{2} - \Delta \,|\, \mathbf{E}[\bar{X}_{q,i}] = \Delta) \\ &\leq \Pr\left(\bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] < \frac{\varepsilon_r}{2} - \varepsilon_r \,|\, \mathbf{E}[\bar{X}_{q,i}] = \Delta\right) \\ &= \Pr\left(\bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] < -\frac{\varepsilon_r}{2} \,|\, \mathbf{E}[\bar{X}_{q,i}] = \Delta\right) \\ &\leq e^{\frac{-2(2N)(-\varepsilon_r/2)^2}{2^2}} \\ &\leq e^{-N\varepsilon_r^2/4},\end{aligned}$$

where the first inequality uses the fact that $\varepsilon_r \leq \hat{\Delta}_{r,i} - \varepsilon_r \leq \Delta$ under event $\mathcal{E}_{\text{apl}}$.

Since the two bounds are equal and the mistakes are mutually exclusive, the probability of making a mistake is at most $e^{-N\varepsilon_r^2/4} = \delta$. □

Next, we show that the output arm is positive and satisfies a condition on the cost difference.

Lemma 17. *Arm $\hat{\boldsymbol{\tau}}$ is positive and satisfies*

$$\left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i^*} \right\rangle \leq \varepsilon_{\text{pnbs}}.$$

Proof. Recall that $\hat{\boldsymbol{\tau}}$ is defined as

$$\hat{\boldsymbol{\tau}} = R_{\hat{i}}, \qquad \text{where} \qquad \hat{i} = \operatorname*{argmin}_{i \in S_{\text{dim}}} \{v_i \cdot a_{\text{right},i}\}.$$

The lemma follows by establishing that arm $\hat{\boldsymbol{\tau}}$ is positive and that its cost is not much greater than that of arm $\boldsymbol{\tau}^{i^*}$. First, following Lemma 16, we have that ParallelNoisyBinarySearch correctly labels the arms with high probability. After the final epoch, ParallelNoisyBinarySearch returns a right arm $\hat{\boldsymbol{\tau}}$. Let us assume that for all epochs q and all $i \in S_{\text{dim}}$, ParallelNoisyBinarySearch

makes no mistakes. Then, arm $\boldsymbol{R_i}$ is positive, and arm $\boldsymbol{L_i}$ is negative, and therefore arm $\hat{\boldsymbol{\tau}}$ must be positive.

Next, we show that arm $\hat{\boldsymbol{\tau}}$ satisfies the condition $\left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i^*} \right\rangle \leq \varepsilon_{\text{pnbs}}$. Trivially, if $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}^{i^*}$, then the cost difference is $0 < \varepsilon_{\text{pnbs}}$ and the condition is satisfied.

Now, suppose that $\hat{\boldsymbol{\tau}} \neq \boldsymbol{\tau}^{i^*}$. Recall that after the final epoch of ParallelNoisyBinarySearch, we have

$$\langle \boldsymbol{v}, \boldsymbol{R_i} - \boldsymbol{L_i} \rangle < \varepsilon_{\text{pnbs}}$$

for all $i \in S_{\text{dim}}$. By definition of $\hat{i}$, it must be that $\langle \boldsymbol{v}, \boldsymbol{R_{i*}} \rangle \geq \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle$, and we can bound the cost difference as

$$\left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i^*} \right\rangle \leq \left\langle \boldsymbol{v}, \boldsymbol{R}_{i^*} - \boldsymbol{\tau}^{i^*} \right\rangle.$$

Since $\boldsymbol{R}_{i^*}$ and $\boldsymbol{L_{i*}}$ bound the region containing $\boldsymbol{\tau}^{i^*}$, we get

$$\begin{aligned} \left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i^*} \right\rangle &\leq \langle \boldsymbol{v}, \boldsymbol{R}_{i^*} - \boldsymbol{L_{i*}} \rangle \\ &< \varepsilon_{\text{pnbs}} \end{aligned}$$

Thus, the cost difference is at most $\varepsilon_{\text{pnbs}}$, as desired. □

Now, we introduce another event. Let $\mathcal{E}_{\text{pnbs}}$ be the event that ParallelNoisyBinarySearch returns a positive arm $\hat{\boldsymbol{\tau}}$ satisfying $\left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i*} \right\rangle \leq \varepsilon_{\text{pnbs}}$.

Lemma 18. *Take $\delta = 1/T^2$, and assume that the number of epochs required for each dimension is upper bounded as*

$$\lfloor \log_2(2/\varepsilon_{\text{pnbs}}) \rfloor \leq T/d.$$

Then, on event $\mathcal{E}_{\text{apl}}$, event $\mathcal{E}_{\text{dr-1}}$, and event $\mathcal{E}_{\text{dr-2}}$, the event $\mathcal{E}_{\text{pnbs}}$ holds with probability at least $1 - 1/T$.

Proof. Suppose that ParallelNoisyBinarySearch made no mistakes in any iteration. Then, following Lemma 17, arm $\hat{\boldsymbol{\tau}}$ is positive and satisfies $\left\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{i^*} \right\rangle \leq \varepsilon_{\text{pnbs}}$. Next, take $\delta = 1/T^2$. From Lemma 16, we have that the probability of a mistake is at most δ. For clarity, let $Q = \lfloor \log_2(2/\varepsilon_{\text{pnbs}}) \rfloor$. For all dimensions $i \in S_{\text{dim}}$ and all epochs $q \leq Q$, let $\bar{X}_{q,i} = \hat{p}_{\text{right},i} - \hat{p}_{\text{left},i}$. We

apply a union bound to get that the probability of ParallelNoisyBinarySearch making a mistake is

$$\begin{aligned}\Pr\left(\exists i \in S_{\text{dim}}, \exists q \leq Q : \bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] \geq \frac{\varepsilon_r}{2}\right) &\leq \sum_{i \in S_{\text{dim}}} \sum_{q=1}^{Q} \Pr\left(\bar{X}_{q,i} - \mathbf{E}[\bar{X}_{q,i}] \geq \frac{\varepsilon_r}{2}\right) \\ &= |S_{\text{dim}}| \cdot Q \cdot \delta.\end{aligned}$$

Upper bounding S_{dim} by d and Q by T/d gives us $T \cdot \delta$. We have that event $\mathcal{E}_{\text{pnbs}}$ holds with probability at least $1 - T\delta = 1 - 1/T$. □

Chapter 5

Regret Analysis

For the reader's convenience, we state the explicit values of epochs $r_{\mathrm{i},i}$ (4.3), r_Δ (4.5), and $r_{\max}$ (4.4) below:

- $r_{\mathrm{i},i} = \lceil \log_2 \frac{1}{v_i} \rceil$, for all $i \in [d]$
- $r_\Delta = \lceil \log_2 \frac{1}{\Delta} \rceil - 4$
- $r_{\max} = \lfloor \log_2 \frac{\sqrt{T/d}}{\log T} \rfloor - 3$

We also define a variable that will be used later: Let $v_{\min} = \min_{i \in [d]} v_i$ be the minimum value of v_i for $i \in [d]$. For the purpose of our regret analysis, we assume that $v_{\min} \geq \log(T)\sqrt{d/T}$, and leave the exploration of the implications of smaller values of v_i for future work.

We now give bounds on the regret of AxisParallelLeftist when applied to the setting of a union of axis-parallel halfspaces, as outlined in Section 3.2. We assume that if arm $\boldsymbol{\tau}^{i^*}$ is optimal, then there exists a dimension $i \in [d]$ such that $\tau_i \leq 1$ and $\tau_i \leq \Delta / v_i$. The following theorems provide bounds on the total regret of AxisParallelLeftist, which is the sum of the regret from each of the algorithms' pulls. In the subsequent sections, we will present the regret bounds for each component of the algorithm. The proofs for the following theorems are given in Section 5.5.

Theorem 19. *Take $\delta = 1/T^2$ and $\varepsilon_{\mathrm{pnbs}} = 1/T$, and suppose that arm $\boldsymbol{\tau}^{i^*}$ is optimal. Then, the regret of AxisParallelLeftist is bounded as*

$$R_T = O\left(\frac{d \cdot \log^2 T}{\Delta} + \min\left\{ \frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle|}, \sqrt{T \log T} \right\} \right)$$

AxisParallelLeftist enjoys a logarithmic regret in T when arm $\boldsymbol{\tau}^{i^*}$ is optimal and Δ is sufficiently large. However, when Δ is very small or close to $\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle$, the second term induces a worst-case regret of $O\left(\sqrt{T \log T}\right)$. The next theorem is useful to handle cases when Δ is small. AxisParallelLeftist does not have a logarithmic regret bound, but still achieves sublinear regret, which matches that of Theorem 19's worst-case bound.

Theorem 20. *Take $\delta = 1/T^2$ and suppose arm $\boldsymbol{\tau}^{i^*}$ is optimal. If $\Delta = \Omega\left(\log(T)\sqrt{\frac{d}{T}}\right)$, then the regret of AxisParallelLeftist is bounded as*

$$R_T = O\left(\log(T)\sqrt{dT} + \min\left\{\frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle|}, \sqrt{T \log T}\right\}\right).$$

Furthermore, if $\Delta \leq \frac{1}{2}\log(T)\sqrt{\frac{d}{T}}$, then the bound improves to

$$R_T = O\left(\log(T)\sqrt{dT}\right).$$

The following theorem captures the case where arm $\mathbf{0}$ is optimal and matches the worst-case bound from Theorem 20.

Theorem 21. *Take $\delta = 1/T^2$, $\varepsilon_{\text{pnbs}} = 1/T$, and suppose that arm $\mathbf{0}$ is optimal. Then, the regret of AxisParallelLeftist is bounded as*

$$R_T = O\left(\log(T)\sqrt{dT}\right).$$

5.1 Regret analysis for AxisParallelLeftist

In this section, we present the proof for the following AxisParallelLeftist lemmas.

Lemma 22. *Take $\delta = 1/T^2$. If arm $\boldsymbol{\tau}^{i^*}$ is optimal, then on event $\mathcal{E}_{\text{apl}}$, the regret contribution from any arm $\boldsymbol{A}_{r,i}$ for a single dimension $i \in [d]$ is at most of order*

$$\min\left\{\frac{\log T}{\Delta}, \sqrt{T/d}\right\}.$$

Lemma 23. *Take $\delta = 1/T^2$. If arm $\mathbf{0}$ is optimal, the regret of AxisParallelLeftist is at most of order $\sqrt{dT}$.*

Before delving into the proofs for the lemmas, we first analyze the instantaneous regret incurred for each comparison in AxisParallelLeftist. Recall that during each epoch and for each dimension i, the algorithm pulls arm $\mathbf{0}$ and arm $\boldsymbol{A_{r,i}}$. Take any epoch $r \leq \rho$:

- Case 1: If arm $\boldsymbol{\tau^{i^*}}$ is optimal, the instantaneous regret from pulling arm $\mathbf{0}$ and a positive arm $\boldsymbol{A_{r,i}}$ is bounded by

$$\begin{aligned}\left(\mu_c(\boldsymbol{\tau^{i^*}}) - \mu_c(\mathbf{0})\right) + \left(\mu_c(\boldsymbol{\tau^{i^*}}) - \mu_c(\boldsymbol{A_{r,i}})\right) &= \left(p_1 - \left\langle \boldsymbol{v}, \boldsymbol{\tau^{i^*}} \right\rangle - p_0\right) \\ &\quad + \left(p_1 - \left\langle \boldsymbol{v}, \boldsymbol{\tau^{i^*}} \right\rangle - (p_1 - \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle)\right) \\ &= \Delta - 2\left\langle \boldsymbol{v}, \boldsymbol{\tau^{i^*}} \right\rangle + \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \\ &\leq \Delta + \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \end{aligned} \tag{5.1}$$

- Case 2: If arm $\mathbf{0}$ is optimal, pulling it incurs no regret. The instantaneous regret from pulling a positive arm $\boldsymbol{A_{r,i}}$, for dimension $i \in [d]$ is bounded by

$$\begin{aligned}\mu_c(\mathbf{0}) - \mu_c(\boldsymbol{A_{r,i}}) &= p_0 - (p_1 - \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle) \\ &\leq \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \end{aligned} \tag{5.2}$$

For the subsequent analyses, we mention that the regret bounds are of the same order regardless of the scale of the cost vector $\boldsymbol{v}$. For brevity, we present the analysis for $v_i \leq 1$ here and leave the details for $v_i > 1$ to Appendix A.2.

Now, we present the lemmas required to prove the above lemmas.

5.1.1 Problem dependent bounds

Since the cost component can vary across dimensions, here we focus our analysis on the regret incurred by pulls in a single dimension.

Arm $\boldsymbol{\tau^{i^*}}$ is optimal

We derive a general regret bound for an arbitrary dimension $i \in [d]$ when arm $\boldsymbol{\tau^{i^*}}$ is optimal. Substituting the bound from (5.1) into the regret expression (3.3), the regret incurred by AxisPar-

allelLeftist for any dimension $i \in [d]$ can be bounded as

$$R_{\text{APL},i} = \underbrace{\sum_{r=0}^{\min\{r_{\text{i},i},\rho\}} n_r \left(\Delta + \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle\right)}_{\text{Phase 1}} + \underbrace{\mathbb{1}[r_{\text{i},i} < \rho] \sum_{r=r_{\text{i},i}+1}^{\rho} n_r \left(\Delta + \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle\right)}_{\text{Phase 2}}$$
$$\leq \sum_{r=0}^{\min\{r_{\text{i},i},\rho\}} n_r \left(\Delta + v_i\right) + \mathbb{1}[r_{\text{i},i} < \rho] \sum_{r=r_{\text{i},i}+1}^{\rho} n_r \left(\Delta + v_i \cdot 2^{r_{\text{i},i}-r}\right),$$

where, for $v_i \leq 1$, scalar $a_{r,i} = 1$ in Phase 1 and $a_{r,i} = 2^{-(r-r_{\text{i},i})}$ in Phase 2. Further simplification using the explicit forms of $r_{\text{i},i}$ and n_r, yields the following bound

$$\begin{aligned} R_{\text{APL},i} &\leq \log\left(\frac{2}{\delta}\right) 2^5 \left(\sum_{r=0}^{\min\{r_{\text{i},i},\rho\}} 2^{2r} \left(\Delta + v_i\right) + \mathbb{1}[r_{\text{i},i} < \rho] \sum_{r=0}^{\rho} \left(2^{2r}\Delta + 2^{r+1}\right) \right) \\ &< \log\left(\frac{2}{\delta}\right) 2^5 \left(\left(\frac{4}{3}(\Delta + v_i) \min\{2^{2r_{\text{i},i}}, 2^{2\rho}\}\right) + \mathbb{1}[r_{\text{i},i} < \rho] \left(\frac{4}{3}\Delta 2^{2\rho} + 2^{\rho+2}\right) \right) \\ &= O\left(\log\left(\frac{1}{\delta}\right) (\Delta + v_i) \min\{2^{2r_{\text{i},i}}, 2^{2\rho}\} + \mathbb{1}[r_{\text{i},i} < \rho] \log\left(\frac{1}{\delta}\right) \left(\Delta 2^{2\rho} + 2^{\rho}\right) \right). \end{aligned} \tag{5.3}$$

Problem instance: $\Delta \leq v_i/16$ **for some** $i \in [d]$**.**

Lemma 24. *Let $\delta = 1/T^2$. For any $i \in [d]$, if $\Delta \leq v_i/16$ and arm $\boldsymbol{\tau^{i^*}}$ is optimal, then on event $\mathcal{E}_{\text{apl}}$, the regret of AxisParallelLeftist in dimension i is of order at most $\frac{\log(T)}{\Delta}$.*

Proof. Fix $i \in [d]$ and take $\delta = 1/T^2$. Assume arm $\boldsymbol{\tau^{i^*}}$ is optimal. The optimality of τ^{i^*} implies that there exists a dimension i such that $\tau_i \leq \frac{\Delta}{v_i}$ and $\tau_i \leq 1$. Thus, on event $\mathcal{E}_{\text{apl}}$, Lemma 7 implies that $\rho \leq r_\Delta + 3$. From (5.3), we bound the regret as follows:

$$\begin{aligned} R_{\text{APL},i} &= O\left(\log\left(\frac{1}{\delta}\right) (\Delta + v_i) \min\{2^{2r_{\text{i},i}}, 2^{2\rho}\} + \mathbb{1}[r_{\text{i},i} < \rho] \log\left(\frac{1}{\delta}\right) \left(\Delta \cdot 2^{2\rho} + 2^{\rho}\right) \right) \\ &= O\left(\log\left(\frac{1}{\delta}\right) \left((\Delta + v_i) 2^{2r_{\text{i},i}} + \Delta \cdot 2^{2\rho} + 2^{\rho}\right) \right) \\ &= O\left(\log\left(\frac{1}{\delta}\right) \left((\Delta + v_i)\frac{1}{v_i^2} + \Delta \cdot \frac{1}{\Delta^2} + \frac{1}{\Delta}\right) \right) \\ &= O\left(\frac{\log(T)}{\Delta} \right), \end{aligned} \tag{5.4}$$

where the last inequality follows from $\Delta \leq v_i/16$. □

Problem instance: $\Delta > v_i/16$ **for some** $i \in [d]$**.**

Lemma 25. *Let $\delta = 1/T^2$. For any $i \in [d]$, if $\Delta > v_i/16$ and arm $\boldsymbol{\tau}^{i^*}$ is optimal, then on event $\mathcal{E}_{\text{apl}}$, the regret incurred by AxisParallelLeftist in dimension i is of order at most $\frac{\log(T)}{\Delta}$.*

Proof. Since the steps within the analysis are similar to the previous problem instance, we will be brief. Fix $i \in [d]$ and take $\delta = 1/T^2$. Assume arm $\boldsymbol{\tau}^{i^*}$ is optimal. On event $\mathcal{E}_{\text{apl}}$, we still have from Lemma 7 that $\rho \leq r_\Delta + 3$. Moreover, $\Delta > v_i/16$ implies that $r_{\text{i},i} > r_\Delta$.

Suppose $r_{\text{i},i} < r_\Delta + 3$ and Phase 2 happens. It suffices to consider the earliest epoch in which the algorithm can complete Phase 1, that is $r_{\text{i},i} + 1 \leq r_\Delta + 1$. Unpacking the explicit values, it follows that $\Delta < v_i/8$. Hence $v_i/16 < \Delta < v_i/8$ and Δ is of the same order as v_i. The regret bound (5.3) in this scenario is at most

$$\begin{aligned} R_{\text{APL},i} &= O\left(\log(T)\Delta \min\{2^{2r_{\text{i},i}}, 2^{2\rho}\} + \log(T)\left(\Delta \cdot 2^{2\rho} + 2^{\rho}\right)\right) \\ &= O\left(\log(T)\Delta\left(\frac{1}{v_i^2}\right) + \log(T)\left(\frac{1}{\Delta}\right)\right). \end{aligned} \tag{5.5}$$

Since $\Delta = \Theta(v_i)$, this implies that

$$\frac{1}{v_i^2} = O\left(\frac{1}{\Delta^2}\right).$$

Thus, the regret bound (5.5) simplifies to

$$R_{\text{APL},i} = O\left(\log(T)\left(\frac{1}{\Delta}\right)\right). \tag{5.6}$$

Next, suppose $r_{\text{i},i} \geq r_\Delta + 3$. Then Phase 2 does not happen and we can drop the second term from the regret bound (5.3), to get

$$\begin{aligned} R_{\text{APL},i} &= O\left(\log(T)(\Delta + v_i)2^{2r_\Delta}\right) \\ &= O\left(\log(T)\left((\Delta + v_i)\frac{1}{\Delta^2}\right)\right) \\ &= O\left(\log(T)\left(\frac{1 + \frac{v_i}{\Delta}}{\Delta}\right)\right). \end{aligned}$$

Furthermore, $\Delta > v_i/16$ implies that $1 + v_i/\Delta$ is $O(1)$, and the above regret bound can be simplified

to

$$R_{\text{APL},i} = O\left(\frac{\log(T)}{\Delta}\right).$$

The regret is of the same order in both scenarios, which concludes the proof. □

Problem instance: $\Delta < 16\log(T)\sqrt{d/T}$

Lemma 26. *Take $\delta = 1/T^2$. For any $i \in [d]$, if $\Delta < 16\log(T)\sqrt{\frac{d}{T}}$ and arm $\boldsymbol{\tau}^{i^*}$ is optimal, then on event $\mathcal{E}_{\text{apl}}$, the regret of AxisParallelLeftist in dimension i is of order at most $\sqrt{T/d}$.*

Proof. Let $\delta = 1/T^2$. Suppose $\Delta < 16\log(T)\sqrt{\frac{d}{T}}$ and AxisParallelLeftist exhausts all epochs before detecting a lower bound on Δ. Recall that the maximum number of epochs is r_{max}. Using the regret bound (5.3) we obtain:

$$R_{\text{APL}} = O\left(\log(T)(\Delta + v_i)\min\{2^{2r_{\text{i},i}}, 2^{2r_{\text{max}}}\} + \mathbb{1}[r_{\text{i},i} < \rho]\log(T)\left(\Delta 2^{2r_{\text{max}}} + 2^{r_{\text{max}}}\right)\right).$$

Using the explicit form of r_{max}, we get

$$\begin{aligned}
R_{\text{APL}} &= O\left(\log(T)\left((\Delta + v_i)\frac{1}{v_i^2}\right) + \log(T)\left(\Delta\frac{T/d}{\log^2(T)} + \frac{\sqrt{T/d}}{\log(T)}\right)\right) \\
&= O\left(\log(T)\left(\Delta\frac{T/d}{\log^2(T)} + \frac{\sqrt{T/d}}{\log(T)}\right) + \log(T)\left(\Delta\frac{T/d}{\log^2(T)} + \frac{\sqrt{T/d}}{\log(T)}\right)\right) \\
&= O\left(\log(T)\left(\Delta\frac{T/d}{\log^2(T)} + \frac{\sqrt{T/d}}{\log(T)}\right)\right),
\end{aligned}$$

where the second inequality uses our assumption that $v_i \geq v_{\text{min}} \geq \log(T)\sqrt{d/T}$. It is easy to see that the regret will be of the same order regardless of the phase the algorithm is in. Finally, for $\Delta = O\left(\log(T)\sqrt{\frac{d}{T}}\right)$, we obtain a regret of order

$$O\left(\sqrt{T/d}\right)$$

which concludes our proof. □

Now, combining Lemma 24, Lemma 25, and Lemma 26, we can prove Lemma 22.

Proof of Lemma 22. Assume arm $\boldsymbol{\tau}^{i^*}$ is optimal and event $\mathcal{E}_{\text{apl}}$ happened. For any $i \in [d]$, consider the following cases:

Case 1: If $\Delta \leq v_i/16$, then by Lemma 24, the regret is bounded by

$$O\left(\frac{\log(T)}{\Delta}\right).$$

Moreover, if $\Delta \geq 16\log(T)\sqrt{\frac{d}{T}}$, then the bound is at most $\sqrt{T/d}$.

Case 2: If $\Delta > v_i/16$, then by Lemma 25 the regret is bounded by

$$\begin{aligned} O\left(\frac{\log(T)}{\Delta}\right) &= O\left(\frac{\log(T)}{v_i}\right) \\ &= O\left(\sqrt{\frac{T}{d}}\right), \end{aligned}$$

where the first equality uses $\Delta > v_i/16$, and the second equality uses $v_{\min} \geq \log(T)\sqrt{\frac{d}{T}}$.

Case 3: If $\Delta < 16\log(T)\sqrt{\frac{d}{T}}$, then Lemma 26 implies the worst-case bound

$$O\left(\sqrt{\frac{T}{d}}\right).$$

Therefore, the contribution to the regret from any arm $\boldsymbol{A_{r,i}}$ for a single dimension is of order at most

$$\min\left\{\frac{\log T}{\Delta}, \sqrt{d/T}\right\}.$$

□

5.1.2 Problem independent bounds

In this section, we analyze the regret bound over all dimensions.

Arm 0 is optimal

Finally, we consider the problem instances where arm **0** is optimal. The regret (3.3) can be expressed using the instantaneous bound (5.2) for the arm **0** optimal case to give us

$$R_{\text{APL}} = \sum_{i=1}^{d} \left(\underbrace{\sum_{r=0}^{\min\{r_{\text{i},i},\rho\}} n_r \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle}_{\text{Phase 1}} + \underbrace{\mathbb{1}[r_{\text{i},i} < \rho] \sum_{r=r_{\text{i},i}+1}^{\rho} n_r \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle}_{\text{Phase 2}} \right). \tag{5.7}$$

efore presenting the proof for Lemma 23, we note that when arm **0** is optimal and event $\mathcal{E}_{\text{apl}}$ occurs, there are no guarantees that all arms $\boldsymbol{A_{r,i}}$ (which are meant to be positive) will remain positive throughout epochs $r \leq \rho$. In fact, since the cost of arm $\boldsymbol{\tau^{i^*}}$ is greater than the reward gap Δ, the algorithm may continue to run during epochs where some arms $\boldsymbol{A_{r,i}}$ become negative. Moreover, any estimate $\hat{\Delta}_{r,i}$ obtained from such negative arms is uninformative and does not trigger a call to DimensionReduction. Consequently, the regret may be bounded using the largest epoch $\rho = r_{\max}$

Proof of Lemma 23. Take $\delta = 1/T^2$ and assume arm **0** is optimal. For any i such that $v_i \leq 1$, the cost $\langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \leq 8\varepsilon_r$ in Phase 1 and $a_{r,i} = 2^{-(r-r_{\text{i},i})}$ in Phase 2. From the regret expression (5.7), we have that

$$\begin{aligned} R_{\text{APL}} &= \sum_{i=1}^{d} \left(\sum_{r=0}^{\min\{r_{\text{i},i},r_{\max}\}} n_r \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle + \mathbb{1}[r_{\text{i},i} < r_{\max}] \sum_{r=r_{\text{i},i}+1}^{r_{\max}} n_r \langle \boldsymbol{v}, \boldsymbol{A_{r,i}} \rangle \right) \\ &< \sum_{i=1}^{d} \left(\sum_{r=0}^{\min\{r_{\text{i},i},r_{\max}\}} n_r \cdot 8\varepsilon_r + \mathbb{1}[r_{\text{i},i} < r_{\max}] \sum_{r=r_{\text{i},i}+1}^{r_{\max}} n_r \cdot v_i \cdot 2^{r_{\text{i},i}-r} \right). \end{aligned}$$

Using the explicit form of $r_{\mathrm{i},i}$, a direct analysis yields a regret of order

$$\begin{aligned}
R_{\mathrm{APL}} &= \sum_{i=1}^{d}\left(\sum_{r=0}^{\min\{r_{\mathrm{i},i},r_{\max}\}} n_r\cdot 8\varepsilon_r + \mathbb{1}[r_{\mathrm{i},i}<r_{\max}]\sum_{r=r_{\mathrm{i},i}+1}^{r_{\max}} n_r\cdot v_i\cdot 2^{\lceil\log_2\frac{1}{v_i}\rceil}\cdot 2^{-r}\right)\\
&\le \sum_{i=1}^{d}\left(2^5\log\left(\frac{2}{\delta}\right)\sum_{r=0}^{\min\{r_{\mathrm{i},i},r_{\max}\}} 2^r + \mathbb{1}[r_{\mathrm{i},i}<r_{\max}]2^6\log\left(\frac{2}{\delta}\right)\sum_{r=r_{\mathrm{i},i}+1}^{r_{\max}} 2^r\right)\\
&< \sum_{i=1}^{d}\left(2^5\log\left(\frac{2}{\delta}\right)\sum_{r=0}^{\min\{r_{\mathrm{i},i},r_{\max}\}} 2^r + \mathbb{1}[r_{\mathrm{i},i}<r_{\max}]2^6\log\left(\frac{2}{\delta}\right)\sum_{r=0}^{r_{\max}} 2^r\right)\\
&= O\left(\sum_{i=1}^{d}\left(\log\left(\frac{1}{\delta}\right)\min\{2^{r_{\mathrm{i},i}},2^{r_{\max}}\}+\log\left(\frac{1}{\delta}\right)2^{r_{\max}}\right)\right).
\end{aligned}$$

Next, fix i and let us consider the terms in the summation. Using the explicit values of $r_{\mathrm{i},i}$ and $r_{\max}$, we get that

$$\begin{aligned}
O\left(\log\left(\frac{1}{\delta}\right)\min\{2^{r_{\mathrm{i},i}},2^{r_{\max}}\}+\log\left(\frac{1}{\delta}\right)2^{r_{\max}}\right) &= O\left(\log\left(\frac{1}{\delta}\right)2^{r_{\mathrm{i},i}}+\log\left(\frac{1}{\delta}\right)2^{r_{\max}}\right)\\
&= O\left(\log\left(\frac{1}{\delta}\right)\frac{1}{v_i}+\log\left(\frac{1}{\delta}\right)\frac{\sqrt{T/d}}{\log(T)}\right)\\
&= O\left(\log\left(\frac{1}{\delta}\right)\frac{\sqrt{T/d}}{\log(T)}\right), \qquad (5.8)
\end{aligned}$$

where the last equality uses our assumption that $v_{\min}\ge\log(T)\sqrt{\frac{d}{T}}$.

Therefore, when arm **0** is optimal, AxisParallelLeftist incurs a regret at most of order

$$O\left(\sum_{i=1}^{d}\left(\log\left(\frac{1}{\delta}\right)\frac{\sqrt{T/d}}{\log(T)}\right)\right)=O\left(\sqrt{dT}\right).$$

□

In the following sections 5.2 and 5.3, we consider the case when arm $\boldsymbol{\tau}^{i^*}$ is optimal. The analysis for when arm **0** is optimal is presented in Section 5.5.

5.2 Regret analysis for DimensionReduction

DimensionReduction uses ε_ρ as a lower bound to determine the number of pulls N needed to distinguish positive from negative arms with high probability. We analyze the regret incurred from these pulls during both elimination rounds. When no dimensions are eliminated (worst case), the algorithm incurs regret of order $O(\frac{d \log T}{\Delta})$.

Lemma 27. *Take $\delta = 1/T^2$ and assume arm $\boldsymbol{\tau}^{i^*}$ is optimal. On event $\mathcal{E}_{\text{apl}}$, the regret of DimensionReduction is at most*

$$O\left(\frac{d \log T}{\Delta}\right)$$

Proof. Assume arm $\boldsymbol{\tau}^{i^*}$ is optimal. Let $i \in S_{\text{dim}}$ be arbitrary and assume event $\mathcal{E}_{\text{apl}}$ happened. Note that r in DimensionReduction is the stopping epoch ρ and from Lemma 7 and Lemma 11, we have $\rho \asymp r_\Delta \asymp \log_2 \frac{1}{\Delta}$.

Pulls to arm 0. Recall that pulling arm **0** incurs at most Δ regret. To obtain the empirical mean $\hat{p}_0$, the algorithm incurs at most $N \cdot \Delta = O(\log(\frac{1}{\delta})(\frac{1}{\Delta}))$ regret.

Pulls to arm $\boldsymbol{R_i}$ in elimination round 1. The regret of arm $\boldsymbol{R_i}$ when $r_{\text{i},i} \geq \rho$, which implies that ρ is in Phase 1, is of order

$$\begin{aligned} N \cdot \langle \boldsymbol{v}, \boldsymbol{R_i} \rangle &\leq \frac{\log(\frac{2}{\delta})}{\varepsilon_\rho^2} \cdot 8\varepsilon_\rho \\ &= O\left(\log\left(\frac{1}{\delta}\right) \cdot 2^\rho\right) \\ &= O\left(\log\left(\frac{1}{\delta}\right) \frac{1}{\Delta}\right). \end{aligned}$$

The regret of arm $\boldsymbol{R_i}$ when $r_{\text{i},i} < \rho$, which implies that ρ is in Phase 2, is of order

$$\begin{aligned} N \cdot \langle \boldsymbol{v}, \boldsymbol{R_i} \rangle &\leq \frac{\log(\frac{2}{\delta})}{\varepsilon_\rho^2} \cdot v_i \cdot 2^{r_{\text{i},i} - \rho} \\ &= O\left(\log\left(\frac{1}{\delta}\right) \cdot 2^\rho\right) \\ &= O\left(\log\left(\frac{1}{\delta}\right) \frac{1}{\Delta}\right). \end{aligned}$$

Pulls to arm $\boldsymbol{R_i'}$ in second elimination round. Recall that $\langle \boldsymbol{v}, \boldsymbol{R_i'} \rangle = v_k \cdot a_{\rho,k}$, where $k = \text{argmin}_{i \in S_{\text{dim}}} v_i \cdot a_{\rho,i}$. Again, a direct analysis of N pulls to arm $\boldsymbol{R_i'}$ gives us a regret of order

$$N \cdot \langle \boldsymbol{v}, \boldsymbol{R_i'} \rangle = N \cdot v_k \cdot a_{\rho,k}.$$

The analysis for each phase is identical to the previous elimination round giving us a regret of order $O\left(\log\left(\frac{1}{\delta}\right)\frac{1}{\Delta}\right)$.

Finally, we upper bound the regret for both elimination rounds by summing over d dimensions, in the worst case. Then, taking $\delta = 1/T^2$, we get that the regret of DimensionReduction is at most

$$O\left(\underbrace{\left(\log\left(\frac{1}{\delta}\right)\frac{1}{\Delta}\right)}_{\text{arm L}} + \underbrace{\left(\sum_{i=1}^{d}\log\left(\frac{1}{\delta}\right)\frac{1}{\Delta}\right)}_{\text{1st elimination}} + \underbrace{\left(\sum_{i=1}^{d}\log\left(\frac{1}{\delta}\right)\frac{1}{\Delta}\right)}_{\text{2nd elimination}}\right) \tag{5.9}$$

$$= O\left(\frac{d\log T}{\Delta}\right) \tag{5.10}$$

□

5.3 Regret analysis for ParallelNoisyBinarySearch

To control the regret of ParallelNoisyBinarySearch, we bound the values of the scalars $a_{\rho,i}$. By design, AxisParallelLeftist ensures that

- in Phase 1: $a_{\rho,i} < \frac{8\varepsilon_\rho}{v_i}$
- in Phase 2: $a_{\rho,i} \leq \frac{16\varepsilon_\rho}{v_i}$

Furthermore, on event $\mathcal{E}_{\text{apl}}$, we have $\varepsilon_r \leq \Delta$ by Lemma 11. Consequently, $a_{\rho,i} \leq \frac{16\Delta}{v_i}$ for all dimensions i in S_{dim}.

Lemma 28. *On event $\mathcal{E}_{\text{apl}}$, if arm $\boldsymbol{\tau^{i^*}}$ is optimal, then the instantaneous regret for any arm pulled in ParallelNoisyBinarySearch is at most* 17Δ.

Proof. We analyze the instantaneous regret of pulling arm $\boldsymbol{M_i}$. For any epoch q, and any $i \in S_{\text{dim}}$, we can use $a_{\rho,i}$ as an upper bound for $a_{\text{mid},i}$. Consider the following cases:

- Case (1): $a_{\text{mid},i} \geq \tau_i$ and arm $\boldsymbol{\tau}^{i^*}$ is optimal.

$$\left(p_1 - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle\right) - \left(p_1 - \left\langle \boldsymbol{v}, \boldsymbol{M_i} \right\rangle\right) < v_i \cdot a_{\text{mid},i} < v_i \cdot a_{r,i} \leq 16\Delta$$

- Case (2): $a_{\text{mid},i} < \tau_i$ and arm $\boldsymbol{\tau}^{i^*}$ is optimal.

$$\left(p_1 - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle\right) - \left(p_0 - \left\langle \boldsymbol{v}, \boldsymbol{M_i} \right\rangle\right) < \Delta + v_i \cdot a_{\text{mid},i} \leq \Delta + 16\Delta$$

□

The next lemma bounds the regret for ParallelNoisyBinarySearch.

Lemma 29. *Take $\delta = 1/T^2$ and $\varepsilon_{\text{pnbs}} = 1/T$, and assume arm $\boldsymbol{\tau}^{i^*}$ is optimal. On the event $\mathcal{E}_{\text{apl}} \cap \mathcal{E}_{\text{dr-1}} \cap \mathcal{E}_{\text{dr-2}}$, the regret of ParallelNoisyBinarySearch is at most*

$$O\left(\frac{d \log^2 T}{\Delta}\right).$$

Proof. Recall that r in ParallelNoisyBinarySearch is the stopping epoch ρ. From Lemma 7 and Lemma 11, we have $\rho \asymp r_\Delta \asymp \log_2 \frac{1}{\Delta}$. Next, following Lemma 28, the regret of $N = O(\log(1/\delta) \cdot 2^{2\rho})$ pulls to an arm by ParallelNoisyBinarySearch is bounded by at most $17\Delta N$. Then, using Lemma 15 to bound the number of iterations as $\log_2(2/\varepsilon_{\text{pnbs}})$ for each dimension and bounding the number of dimensions by d, in the worst-case, we find that the regret contribution from ParallelNoisyBinarySearch is

$$\begin{aligned}
\sum_{i \in S_{\text{dim}}} \sum_{q=1}^{\left\lfloor \log_2 \frac{2}{\varepsilon_{\text{pnbs}}} \right\rfloor} 17\Delta N &\leq d \cdot \log_2\left(\frac{2}{\varepsilon_{\text{pnbs}}}\right) \cdot 17\Delta \left(\frac{4 \log(\frac{1}{\delta})}{\varepsilon_\rho^2}\right) \\
&\leq d \cdot \log_2\left(\frac{2}{\varepsilon_{\text{pnbs}}}\right) \cdot 17\Delta \left(2^{10} \log\left(\frac{1}{\delta}\right) \cdot 2^{2\lceil \log_2 \frac{1}{\Delta} \rceil}\right) \\
&= O\left(d \cdot \log_2\left(\frac{1}{\varepsilon_{\text{pnbs}}}\right) \cdot \log\left(\frac{1}{\delta}\right) \cdot \frac{1}{\Delta}\right)
\end{aligned}$$

Taking $\delta = 1/T^2$ and $\varepsilon_{\text{pnbs}} = 1/T$ yields

$$O\left(\frac{d \cdot \log^2 T}{\Delta}\right).$$

□

5.4 Regret Analysis for UCB

Here, we bound the regret of UCB run on arm $\mathbf{0}$ and arm $\hat{\boldsymbol{\tau}}$.

Lemma 30. *On the event $\mathcal{E}_{\text{apl}} \cap \mathcal{E}_{\text{dr-1}} \cap \mathcal{E}_{\text{dr-2}} \cap \mathcal{E}_{\text{pnbs}}$, the regret contribution from UCB is at most*

$$O\left(\min\left\{\frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle|}, \sqrt{T \log T}\right\}\right).$$

Proof. We begin by recalling the regret bound for UCB. From Auer et al. (2002), the instance-dependent cumulative regret of UCB after T rounds for a 2-armed bandit is of order

$$O\left(\frac{\log T}{\Delta_i}\right)$$

where $\Delta_i := \mu^* - \mu_i$ is the reward gap between the optimal mean reward and the mean reward μ_i for the suboptimal arm i. However, this bound can become loose when Δ_i is asymptotically small, meaning that as $T \to \infty$ the gap $\Delta_i \to 0$. In this case, UCB also enjoys an instance-independent regret bound of order

$$O\left(\sqrt{T \log T}\right).$$

Combining these two bounds, we obtain the overall regret bound for UCB:

$$O\left(\min\left\{\frac{\log T}{\Delta_i}, \sqrt{T \log T}\right\}\right). \tag{5.11}$$

We now analyze the UCB regret with respect to the set of arms $\{\mathbf{0}, \hat{\boldsymbol{\tau}}\}$, where arm $\hat{\boldsymbol{\tau}}$ is the arm returned by ParallelNoisyBinarySearch. For the set of arms $\{\mathbf{0}, \hat{\boldsymbol{\tau}}\}$, the gap Δ_i is given by

$$\Delta_i = |\mu_c(\hat{\boldsymbol{\tau}}) - \mu_c(\mathbf{0})|.$$

Under the high probabilty event $\mathcal{E}_{\text{pnbs}}$, arm $\hat{\boldsymbol{\tau}}$ is positive and satisfies

$$\left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle \leq \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle \leq \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle + \varepsilon_{\text{pnbs}}, \tag{5.12}$$

where $\varepsilon_{\text{pnbs}} = 1/T$. Then, using our notion of mean expected revenue $\mu_c(\mathbf{0}) = p_0$ for arm $\mathbf{0}$ and $\mu_c(\hat{\boldsymbol{\tau}}) = p_1 - \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle$ for arm $\hat{\boldsymbol{\tau}}$, we have that Δ_i is

$$\begin{aligned} |\mu_c(\hat{\boldsymbol{\tau}}) - \mu_c(\mathbf{0})| &= |p_1 - \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle - p_0| \\ &= |\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle + \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle - \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle \,|. \end{aligned}$$

Next, consider the right hand side of the above equation. The triangle inequality yields a lower bound of

$$|\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle + \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle - \langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle \,| \geq |\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle| - |\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle|.$$

Since by (5.12), we have

$$\langle \boldsymbol{v}, \hat{\boldsymbol{\tau}} \rangle - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle \leq \varepsilon_{\text{pnbs}} = 1/T,$$

it follows that

$$|\mu_c(\hat{\boldsymbol{\tau}}) - \mu_c(\mathbf{0})| \geq |\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle| - 1/T.$$

Observe that if $|\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle| < 1/T$ then the regret is negative, which is not possible, and the regret bound (5.11) is given by

$$O\left(\min\left\{\frac{\log T}{\max\{0, |\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle| - 1/T\}}, \sqrt{T \log(T)}\right\}\right).$$

Finally, if $|\Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \right\rangle| \approx 1/T$, then $\sqrt{T \log T}$ dominates. Therefore, the regret can be simplified to

$$O\left(\min\left\{\frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*} \rangle|}, \sqrt{T \log T}\right\}\right),$$

and we get our bound. □

5.5 Cumulative Regret

Proof of Theorem 19. From Lemma 9, we have that event $\mathcal{E}_{\text{apl}}$ happens with probability at least $1 - 1/T$. Then, let us consider the case where event $\mathcal{E}_{\text{apl}}$ happens. We can bound the regret contribution

of AxisParallelLeftist using Lemma 22 which implies that the regret across all d dimensions is at most of order

$$\min\left\{\frac{d\log T}{\Delta}, \sqrt{dT}\right\}.$$

Next, using Lemma 27, Lemma 29, and Lemma 30 for the bounds for DimensionReduction, ParallelNoisyBinarySearch, and UCB, we get that the regret is

$$\begin{aligned}
R_T &= [\text{AxisParallelLeftist}] + [\text{DimensionReduction}] + [\text{ParallelNoisyBinarySearch}] + [\text{UCB}] \\
&= O\left(\min\left\{\frac{d\log T}{\Delta}, \sqrt{dT}\right\}\right) + O\left(\frac{d\log T}{\Delta}\right) + O\left(\frac{d\log^2 T}{\Delta}\right) \\
&+ O\left(\min\left\{\frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*}\rangle|}, \sqrt{T\log T}\right\}\right) \\
&= O\left(\frac{d\log^2 T}{\Delta} + \min\left\{\frac{\log T}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*}\rangle|}, \sqrt{T\log T}\right\}\right).
\end{aligned}$$

□

Proof of Theorem 20. If $\Delta = \Omega\left(\log(T)\cdot\sqrt{\frac{d}{T}}\right)$, a condition which implies that DimensionReduction will be called, the regret is bounded as

$$\begin{aligned}
R_T &= [\text{AxisParallelLeftist}] + [\text{DimensionReduction}] + [\text{ParallelNoisyBinarySearch}] + [\text{UCB}] \\
&= O\left(\sqrt{dT}\right) + O\left(\sqrt{dT}\right) + O\left(\log(T)\sqrt{dT}\right) + O\left(\min\left\{\frac{\log(T)}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*}\rangle|}, \sqrt{T\log(T)}\right\}\right) \\
&= O\left(\log(T)\sqrt{dT} + \min\left\{\frac{\log(T)}{|\Delta - \langle \boldsymbol{v}, \boldsymbol{\tau}^{i^*}\rangle|}, \sqrt{T\log(T)}\right\}\right).
\end{aligned}$$

If Δ is very small, that is $\Delta \le \frac{1}{2}\log(T)\cdot\sqrt{\frac{d}{T}}$, then it is implied that $r_\Delta + 1 > r_{\max}$. With probability $1 - 1/T$, DimensionReduction will not be called. We can apply Lemma 26 to bound the reget contribution from AxisParallelLeftist and include the regret incurred from committing to

arm $\mathbf{0}$. Let N be the number of pulls up to epoch $r_{\max}$. The regret is at most

$$\begin{aligned}
&[\text{APL}] + [\text{Commit}] \\
&= O\left(\sqrt{dT}\right) + \sum_{t=N+1}^{T} \Delta - \left\langle \boldsymbol{v}, \boldsymbol{\tau}^{\boldsymbol{i}^*} \right\rangle \\
&= O\left(\sqrt{dT}\right) + T\Delta \\
&= O\left(\log(T)\sqrt{dT}\right)
\end{aligned}$$

□

Finally, we consider the case where arm $\mathbf{0}$ is optimal and DimensionReduction is called and briefly show the cumulative regret of AxisParallelLeftist in this case.

Proof of Theorem 21. Assume that arm $\mathbf{0}$ is optimal and that DimensionReduction is called. We do not assume that $\rho \leq r_\Delta + 3$; in fact, ρ may be significantly larger than $r_\Delta + 3$. Therefore, we can simply bound the number of pulls N using the largest possible epoch, $r = r_{\max}$.

First, the regret from AxisParallelLeftist when arm $\mathbf{0}$ is optimal is given by Lemma 23. Next, the regret from DimensionReduction comes from the two elimination rounds, since pulling arm $\mathbf{0}$ incurs no regret. Summing over all d dimensions in the worst case, we can bound the regret of DimensionReduction as

$$\begin{aligned}
R_{\text{DR}} &= \sum_{i=1}^{d} N \left\langle \boldsymbol{v}, \boldsymbol{R_i} \right\rangle + \sum_{i=1}^{d} N \left\langle \boldsymbol{v}, \boldsymbol{R'_i} \right\rangle \\
&= O\left(\sum_{i=1}^{d} \frac{\log(\frac{1}{\delta})}{\varepsilon_r^2} \cdot v_i \cdot 2^{r_{\text{i},i}-r} + \sum_{i=1}^{d} \frac{\log(\frac{1}{\delta})}{\varepsilon_r^2} \cdot v_i \cdot 2^{r_{\text{i},i}-r} \right) \\
&= O\left(\sum_{i=1}^{d} \log\left(\frac{1}{\delta}\right) 2^r \right) \\
&= O\left(\sum_{i=1}^{d} \log\left(\frac{1}{\delta}\right) \frac{\sqrt{T/d}}{\log T} \right)
\end{aligned}$$

Taking $\delta = 1/T^2$, we get that

$$\begin{aligned} R_{\text{DR}} &= O\left(\sum_{i=1}^{d} \sqrt{T/d}\right) \\ &= O\left(\sqrt{dT}\right) \end{aligned}$$

In ParallelNoisyBinarySearch, the cost of any arm $\boldsymbol{M_i}$ is at most the cost of its corresponding right arm $\boldsymbol{R_i}$. Then, taking $|S_{\text{dim}}| = d$ in the worst case, we can bound the regret of ParallelNoisyBinarySearch as

$$\begin{aligned} R_{\text{PNBS}} &= \sum_{i \in S_{\text{dim}}} \sum_{q=1}^{\left\lfloor \log_2 \frac{2}{\varepsilon_{\text{pnbs}}} \right\rfloor} N \left\langle \boldsymbol{v}, \boldsymbol{R_i} \right\rangle \\ &= O\left(\sum_{i=1}^{d} \sum_{q=1}^{\left\lfloor \log_2 \frac{2}{\varepsilon_{\text{pnbs}}} \right\rfloor} \log\left(\frac{1}{\delta}\right) 2^r \right) \end{aligned}$$

For $\varepsilon_{\text{pnbs}} = 1/T$ and $\delta = 1/T^2$, we get that

$$\begin{aligned} R_{\text{PNBS}} &= d \cdot \log_2(T) \sqrt{T/d} \\ &= O\left(\log(T)\sqrt{dT}\right). \end{aligned}$$

Finally, combining the above bounds for AxisParallelLeftist, DimensionReduction, and ParallelNoisyBinarySearch with the problem independent bound of UCB from Lemma 30 yields a total regret of

$$\begin{aligned} R_T &= [\text{AxisParallelLeftist}] + [\text{DimensionReduction}] + [\text{ParallelNoisyBinarySearch}] + [\text{UCB}] \\ &= O\left(\sqrt{dT}\right) + O\left(\sqrt{dT}\right) + O\left(\log(T)\sqrt{dT}\right) + O\left(\sqrt{T\log(T)}\right) \\ &= O\left(\log(T)\sqrt{dT}\right). \end{aligned}$$

□

Chapter 6

Conclusions and Future Work

Thresholded linear bandits represent a novel class of bandit problems that highlight the complexities introduced by the presence of a discontinuous reward function. This thesis presents a method for identifying optimal arms without needing to estimate the problem parameter θ, while achieving sublinear regret. This is accomplished through a structured exploration strategy that leverages the properties of axis-parallel halfspaces, effectively reducing the complexity of the problem to d one-dimensional subproblems.

While the algorithm demonstrates sublinear regret, it is important to note that it does not scale well with increasing dimensions d. Future work should focus on enhancing the algorithm to better exploit the multi-dimensional structure of the problem. Another promising research direction involves extending the geometrical exploration strategy to other problem variants mentioned in Section 3.3. These extensions could further enrich the applicability and effectiveness of thresholded linear bandit algorithms in various domains.

Appendix A

Appendix

A.1 Proof for Hoeffding's inequality for the first elimination round of DimensionReduction (Lemma 12).

Assume arm $\boldsymbol{R_i}$ is positive. Let the empirical difference of sample means $\bar{\Delta}_i$ be defined as

$$\begin{aligned}\bar{\Delta}_i &= \hat{p}_{\text{right,i}} - \hat{p}_0 \\ &= \frac{x_{i,1} + ... + x_{i,N}}{N} - \frac{x_{0,1} + ... + x_{0,N}}{N} \\ &= \frac{x_{i,1} + ... + x_{i,N} + (-x_{0,1}) + ... + (-x_{0,N})}{N},\end{aligned}$$

where $x_{i,j}$ and $x_{0,j}$, for $j \in [N]$, are the revenue samples from arms $\boldsymbol{R_i}$ and $\mathbf{0}$, respectively. By Hoeffding's inequality, for $N = 16\log(2/\delta)/\varepsilon_r^2$ and $\varepsilon' = \varepsilon_r/4$, we have

$$\begin{aligned}\Pr\left(|\bar{\Delta}_i - \Delta| > \varepsilon'\right) &\leq 2\exp\left(-\frac{2(2N)(\varepsilon')^2}{2^2}\right) \\ &= 2\exp\left(-\left(16\frac{\log(2/\delta)}{\varepsilon_r^2}\right)\left(\frac{\varepsilon_r}{4}\right)^2\right) \\ &= 2\exp\left(-\log\frac{2}{\delta}\right) \\ &= \delta.\end{aligned}$$

Note that $\bar{\Delta}_i$ is the empirical average of $2N$ samples, and each sample term is bounded within $[-1, 1]$. Hence the factor of 4 in the denominator of the exponential term.

A.2 Cumulative regret of AxisParallelLeftist for dimension i when $v_i > 1$

Take $\delta = 1/T^2$ and $n_r = \log\left(\frac{2}{\delta}\right)/2\varepsilon_r^2$. Suppose $v_i > 1$. Then $a_{0,i} = 1/v_i$ and Phase 1 ends after the first epoch. The cost of pulling arm $\boldsymbol{A_{r,i}}$ in any epoch $r \leq \rho$ is

$$v_i\left(\frac{1}{v_i}\cdot 2^{-r}\right) = 2^{-r}.$$

Consider the following cases and the corresponding regret incurred by AxisParallelLeftist for dimension i.

Case 1: Suppose arm 0 is optimal. We get that the regret for dimension i is of order

$$\begin{aligned}\sum_{r=0}^{r_{\max}} n_r \cdot 2^{-r} &= \sum_{r=0}^{r_{\max}} \frac{\log\frac{2}{\delta}}{2\varepsilon_r^2}\cdot 2^{-r}\\ &= \sum_{r=0}^{r_{\max}} \log\left(\frac{2}{\delta}\right)\cdot 2^{r+5}\\ &= O\left(\log\left(\frac{1}{\delta}\right)\cdot\frac{\sqrt{T/d}}{\log T}\right)\\ &= O\left(\sqrt{T/d}\right),\end{aligned}$$

where we set $\rho = r_{\max}$ and potentially overcount.

Case 2: Suppose arm $\boldsymbol{\tau^{i^*}}$ is optimal. If arm $\boldsymbol{\tau^{i^*}}$ is optimal, and $\tau_{i^*} \leq \Delta/v_i$ and $\tau_{i^*} \leq 1$, then $\rho \leq r_\Delta + 3$ by Lemma 7. The regret for dimension i is of order

$$\begin{aligned}\sum_{r=0}^{r_\Delta+3}\left(n_r\left(\Delta + 2^{-r}\right)\right) &= \sum_{r=0}^{r_\Delta+3}\left(\log\left(\frac{2}{\delta}\right)\left(2^{2r+5}\cdot\Delta + 2^{r+5}\right)\right)\\ &= O\left(\log\left(\frac{1}{\delta}\right)\left(\frac{1}{\Delta^2}\cdot\Delta + \frac{1}{\Delta}\right)\right)\\ &= O\left(\frac{\log T}{\Delta}\right).\end{aligned}$$

Case 3: Suppose arm $\boldsymbol{\tau}^{i^*}$ is optimal and $\Delta < 16\log(T)\sqrt{d/T}$. If $\Delta < 16\log(T)\sqrt{d/T}$ and AxisParallelLeftist exhaustes all epochs before detecting a lower bound on Δ, then $\rho = r_{\max}$. The regret for dimension i is of order

$$\begin{aligned}\sum_{r=0}^{r_{\max}} \left(n_r \left(\Delta + 2^{-r}\right)\right) &= \sum_{r=0}^{r_{\max}} \left(\log\left(\frac{2}{\delta}\right)\left(2^{2r+5}\cdot\Delta + 2^{r+5}\right)\right) \\ &= O\left(\log\left(\frac{1}{\delta}\right)\left(\frac{T/d}{\log^2(T)}\cdot\Delta + \frac{\sqrt{T/d}}{\log(T)}\right)\right).\end{aligned}$$

For $\Delta = O(\log(T)\sqrt{d/T})$, we get a regret of order $\sqrt{T/d}$.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2312–2320, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

Nicolò Cesa-Bianchi, Tommaso Cesari, and Vianney Perchet. Dynamic pricing with finitely many unknown valuations. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 247–273. PMLR, 22–24 Mar 2019.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 67–82. PMLR, 17–18 Aug 2018.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, pages 586–594, 2010.

Shweta Jain and Sujit Gujar. A multiarmed bandit based incentive mechanism for a subset selection

of customers for demand response in smart grids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2046–2053, 2020. doi: 10.1609/aaai.v34i02.5577.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: online computation and hashing. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 98–108. Curran Associates Inc., 2017. ISBN 9781510860964.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces, 2008.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985. ISSN 0196-8858. doi: 10.1016/0196-8858(85)90002-8.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737. PMLR, 2017.

Jérémie Mary, Romaric Gaudel, and Philippe Preux. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data*, pages 325–336, Cham, 2015. Springer International Publishing.

Nishant A. Mehta, Junpei Komiyama, Vamsi K. Potluru, Andrea Nguyen, and Mica Grant-Hagen. Thresholded linear bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 6968–7020. PMLR, 2023.

Kanishka Misra, Eric M. Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019. doi: 10.1287/mksc.2018.1129.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, May 2010. ISSN 0364-765X. doi: 10.1287/moor.1100.0446.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444.

Jia Yuan Yu and Shie Mannor. Unimodal bandits. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 41–48, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.