

2024

Constrained learning in the bandit setting: doubly optimistic strategies and fast rates

<https://hdl.handle.net/2144/48856>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
COLLEGE OF ENGINEERING

Dissertation

**CONSTRAINED LEARNING IN THE BANDIT SETTING:
DOUBLY OPTIMISTIC STRATEGIES AND FAST RATES**

by

TIANRUI CHEN

B.S., Tsinghua University, 2015

M.S., Tsinghua University, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2024

© 2024 by
TIANRUI CHEN
All rights reserved

Approved by

First Reader

Venkatesh Saligrama, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Second Reader

Alex Olshevsky, PhD
Associate Professor of Electrical and Computer Engineering
Associate Professor of Systems Engineering
Associate Professor of Computer Science

Third Reader

Ashok Cutkosky, PhD
Assistant Professor of Electrical and Computer Engineering
Assistant Professor of Computer Science
Assistant Professor of Systems Engineering

Fourth Reader

Brian Kulis, PhD
Associate Professor of Electrical and Computer Engineering
Associate Professor of Systems Engineering

Acknowledgments

I would like to first express my gratitude to my advisor Prof. Venkatesh Saligrama. Working with him has always been a great pleasure to me. He taught me how to conduct research by encouraging me to ask questions from all aspects, simplify problems to a naive level, and dig up into the root of problems. I am also truly grateful for his understanding and tolerance of my personal situations.

I would also like to thank my committee members Prof. Alex Olshevsky, Prof. Brian Kulis, and Prof. Ashok Cutkosky, for taking their time to read and provide valuable feedbacks.

I would like to thank my collaborator Dr. Aditya Gangrade for his devoted and detailed discussion on technical problems, and for helping me flesh out concepts, and improve writing.

I thank my labmates for their company on my PhD journey. I would like to especially thank Anil and Ruizhao for the help with setting up experiments, and Alp for his experience on various milestones.

Finally, I would like to thank my relatives and friends for their emotional support and encouragement. I especially thank my parent for giving me the courage and strength to always continue exploring.

CONSTRAINED LEARNING IN THE BANDIT SETTING: DOUBLY OPTIMISTIC STRATEGIES AND FAST RATES

TIANRUI CHEN

Boston University, College of Engineering, 2024

Major Professor: Venkatesh Saligrama, PhD
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

ABSTRACT

The (stochastic) bandit problem is a classic example used to address the challenge of balancing exploration and exploitation when dealing with bandit feedback. This dissertation focuses on stochastic bandits with constraints, where each action (or arm) bears both a reward and a safety risk. The objective is to maximize rewards while avoiding unsafe options. We explore the application of the optimistic principle in this context. Specifically, we study two performance metrics, the reward regret and the constraint violation. Our formulation penalizes deficit reward and excess risk in a per-round sense, thus pertinent to the safety critical scenario. We propose a doubly optimistic approach that aggressively determines which arms are feasible to be considered, along with several realizations of this strategy to specific scenarios. We instantiate this scheme on the multi-armed bandit, linear bandit, and generalized linear bandit cases. Our findings demonstrate that these algorithms achieve a regret bound that is faster than existing results. For the constrained multi-armed bandit problem, we achieve logarithmic regret for both the reward and safety risk; for linear and generalized linear bandit settings, we carry out a dual analysis based on linear programming sensitivity, identify the discreteness from the continuous problem, and manage to show

a logarithmic regret on reward and a $\log(T)$ to \sqrt{T} safety violation, which is the best to expect without prior information. The upper bounds are accompanied with lower bounds that are matching in terms of time horizon and gaps. We complement our study with illustrative simulations, and conclude with several future directions.

Contents

1	Introduction	1
1.1	Constrained Learning	1
1.2	Bandit Problems	3
1.3	Major Scheme of the Dissertation	3
1.4	Motivating Example: Clinical Trials	4
1.5	Chapter Overview	5
2	Problem Statement and Literature Review	6
2.1	Problem Setup	6
2.2	Literature Review	9
3	Constrained Multi-Armed Bandits	19
3.1	Problem Setup and Definitions	19
3.2	Doubly Optimistic Strategies on a Frequentist Perspective	22
3.3	Bayesian Methods	26
3.3.1	Thompson Sampling with Optimistic Safety Indices	27
3.3.2	Thompson Sampling with BAYESUCB	29
3.4	Lower Bound	30
4	Constrained Linear Bandits	32
4.1	Problem Setup and Definitions	32
4.2	Doubly Optimistic Play for SLB	35
4.2.1	Noise Scales and Confidence Sets	35
4.2.2	The DOSLB Algorithm	36

4.3	Lower Bound and Hardness of the Problem	38
4.4	Identifying Discreteness through Basic Index Sets	40
4.4.1	Basic Index Sets	42
4.5	Gaps Associated with Suboptimal BISs	43
4.5.1	Formal Definitions of the Gaps	47
4.6	Regret Bounds	49
5	Constrained Generalized Linear Bandits	57
5.1	Problem Setup	57
5.2	The DOSGLB Algorithm	59
5.2.1	Canonical Exponential Family and Generalized Linear Model	60
5.2.2	Maximum Likelihood Estimation of Parameters	60
5.2.3	Validity of the MLE	61
5.2.4	The DOSGLB Algorithm	62
5.2.5	Computational Efficiency	64
5.2.6	Reduction to Constrained Multi-Armed Linear Contextual Bandit Algorithms	65
5.3	Gaps under Generalized Linear Structure	66
5.4	Regret Analysis	71
6	Simulations	74
6.1	Simulations on SLB	74
6.1.1	Computationally Feasible Relaxation	74
6.1.2	Results	75
7	Conclusion and Future Work	80
A	Supplement for § 3	83
A.1	Notation and General Proof Strategy	83

A.2	Proof for Doubly Optimistic Confidence Bounds	86
A.2.1	Proof of Theorem 3.2.1	87
A.2.2	Proof of Theorem 3.2.2	88
A.2.3	Proof of Lemma A.2.1	90
A.3	Proofs for Thompson Sampling with Optimistic Safety Indices	93
A.4	Proofs for Thompson Sampling with BAYESUCB	99
A.5	Lower Bound	103
B	Supplement for § 4	106
B.1	Quantitative Bounds from the Theory of Online Linear Regression	106
B.2	Appendix on the Structural Behavior of DOSLB	109
B.3	Controlling the Play of Suboptimal BISs	113
B.3.1	Localizing Actions when a BIS is Activated	113
B.3.2	Proof of Noise Scale Lower Bound and the Finiteness of Spread	114
B.3.3	Limiting the Occurrence of Suboptimal BISs	116
B.4	Proofs of Bounds on Efficiency Regret and Safety Violations	117
B.4.1	The Efficiency of the Actions of DOSLB when Activating Optimal BISs	118
B.4.2	Proof of the Main Theorem	123
B.4.3	Proofs of Subsidiary Claims from §4.6	124
B.5	Proofs of Lower Bounds	127
B.5.1	Proof of Polynomial Lower Bound	128
B.5.2	Necessity of Dependency on Gaps.	129
C	Supplement for § 5	137
C.1	Auxiliary Results for the Regret Bound	137
C.2	Proofs on the Gaps	140
C.3	Proofs for the Regret Analysis	144

References	147
Curriculum Vitae	153

List of Tables

B.1 Description of BISs 131

List of Figures

4.1	Illustration of Gaps	45
6.1	Efficacy Regret and Safety Violation of DOSLB	76
6.2	Number of Suboptimal BIS Activated	77
6.3	Comparison between DOSLB and Safe-LTS	78
6.4	Comparison between L_∞ and L_1 Relaxations	79

List of Abbreviations

GLB	Generalized Linear Bandit
LB	Linear Bandit
LP	Linear Programming
MAB	Multi-armed Bandit
TS	Thompson Sampling
UCB	Upper Confidence Bound

Chapter 1

Introduction

1.1 Constrained Learning

Machine learning methods have been extensively researched and utilized across diverse real-world domains, such as engineering systems and clinical trials. Typically, a conventional machine learning problem involves optimizing a selected objective within a predefined function class. However, this approach overlooks the reality that many real-world systems operate under certain constraints. In the contemporary landscape of machine learning and artificial intelligence, the shift from conventional learning paradigms to constrained learning represents a significant move towards developing solutions that better align with the challenges posed by real-world scenarios. Although traditional learning approaches have led to notable progress, they often prove inadequate when confronted with the intricacies and limitations inherent in practical applications. The motivation behind constrained learning is rooted in the recognition and effective addressing of these complexities encountered in real-world contexts.

Standard learning paradigms often assume, or is simplified so that there is only a single objective to be optimized. Constrained learning, on the other hand, embraces the reality of limited resources, budget constraints, and contextual limitations, which addresses the critical need for resource-efficient or safety-critical decision-making. In scenarios where resources are finite—be it time, budget, or computational power—learning approaches that ignore constraints may lead to suboptimal solutions. Constrained learning methodologies aim to optimize decision-making processes by

explicitly considering and adhering to real-world constraints, ensuring efficiency and practicality in application.

Learning models trained in constrained environments are better equipped to generalize and adapt to the complexities of real-world situations. By incorporating constraints during the learning process, models are forced to navigate the challenges that arise in practical deployment, resulting in solutions that are more robust and reliable across diverse and dynamic contexts.

The ethical dimensions of machine learning methods demand that models not only perform well in controlled settings but also adhere to societal norms, legal constraints, and ethical guidelines. Constrained learning provides a framework for developing machine learning systems that not only excel in performance metrics but also operate within the bounds of ethical considerations, ensuring responsible and accountable deployment in real-world scenarios. Constrained learning methodologies actively contribute to mitigating biases and promoting fairness in machine learning systems. By incorporating constraints that reflect fairness requirements and ethical standards, these approaches strive to develop models that are not only technically proficient but also sensitive to the societal impacts of their decisions.

The motivation for embracing constrained learning over standard learning without constraints lies in its ability to bridge the gap between theoretical advancements and practical applicability. As we proceed towards the integration of machine learning into everyday life, the need for models that understand and navigate real-world constraints becomes increasingly imperative. Constrained learning thus serves as a new direction that pushes machine learning techniques towards more responsible, ethical, and effective side, surpassing the limitations of traditional learning paradigms.

1.2 Bandit Problems

Bandit problems constitute a category of sequential decision-making challenges present in diverse fields, including machine learning, economics, and operations research. The term *bandit* draws from the analogy of a gambler facing a row of slot machines (one-armed bandits) and having to decide where to allocate limited resources, such as time or money, to maximize overall rewards.

A bandit problem involves an agent, or player making a series of decisions over time faced with uncertainty. The agent is presented with a set of actions, often referred to as arms, each associated with an unknown reward distribution. The agent's goal is to learn and exploit the best-performing arm while simultaneously exploring other arms to gather information about their potential rewards.

The major challenge in bandit problems revolves around the exploration-exploitation trade-off. Exploitation entails selecting the arm estimated to be the best based on historical information, while exploration involves trying other arms to refine understanding of their reward distributions. Keeping a balance between these two aspects is crucial for optimizing cumulative rewards over time.

Bandit problems have applications in real-world scenarios such as clinical trials, online advertising, recommendation systems, and autonomous systems. Theoretical advancements and practical algorithms like epsilon-greedy, UCB (Upper Confidence Bound), and TS(Thompson Sampling) have been developed to address different facets of bandit problems.

1.3 Major Scheme of the Dissertation

In this dissertation, we study the bandit problem with constraints, where an agent is trying to solve a sequential decision-making problem under bandit information, and under unknown constraints. Bandit with constraints models the complexity of

constraints on top of the standard bandits, capturing the reality that in addition to the exploration-exploitation trade-off, it is also important to deal carefully with the reward-constraint trade-off in applications. In an intuitive sense, an agent can play more aggressively (in an extreme sense, as aggressive as non-constrained case) to accumulate rewards, while it must also retain the cautious to obey the constraints in face of stochastic feedback. In the next section, we motivate the constrained bandit problem with real-world examples.

1.4 Motivating Example: Clinical Trials

In clinical trials, the primary goal is to evaluate the effectiveness and safety of trial drugs. These drugs, while potentially curative for certain diseases, often bring with them a range of side effects, such as headaches and nausea. The challenge lies not only in ensuring the drug's effectiveness in treating the condition but also in minimizing these side effects to acceptable levels. This dual focus forms the basis for a critical decision-making problem: selecting the most appropriate drug and dosage for each patient that is both curative and safe.

This decision-making process is complex, as it must balance the positive outcomes against the potential negative side effects, ensuring that the latter remain below a certain safety threshold. Each patient's reaction to a drug is unique, which adds another layer of complexity and uncertainty. The effects of a drug on an individual can be thought of as random variables, with their average effects across the population providing a baseline for comparison. In this context, it's not sufficient to alternate between ineffective treatments (like placebos) and effective treatments that might be harmful. Safety needs to be a continuous consideration, ensuring that each patient receives a treatment that maximizes benefits while minimizing risks.

On one hand, ineffective treatments can lead to a lack of progress in combating

the disease; on the other hand, overly aggressive treatments can lead to unacceptable side effects. Thus, in a clinical trial, the objective becomes to find an optimal *sweet spot* where the drug is sufficiently effective in treating the disease without crossing the threshold of tolerable side effects for each individual patient in an overwhelming sense. This necessitates a nuanced approach to treatment assignment, one that accounts for the effectiveness and safety at the same time.

1.5 Chapter Overview

The remainder of this dissertation is arranged as follows. Chapter 2 formally sets up the problem and reviews related literature. Chapter 3 presents results for constrained multi-armed bandits. Chapter 4 incorporates the infinite armed case by considering the constrained linear bandits setting. Chapter 5 further extends the results to generalized linear model. Chapter 6 exhibits simulation results. Chapter 7 concludes the dissertation with future directions.

Chapter 2

Problem Statement and Literature Review

2.1 Problem Setup

The general form of bandit problem is a fundamental framework to study the exploration-exploitation trade-off. The game unfolds over rounds, during each of which the agent (also referred to as player or learner) selects an action (or arm) A_t from a set of available actions \mathcal{A}_t . The environment then provides feedback specific to the chosen action $f_t(A_t)$, and so the agent gathers information and refine its strategy before proceeding to the next round. The objective is to maximize rewards or minimize losses over time, with rewards being determined by the feedback in either a deterministic or stochastic manner.

To introduce the constrained (stochastic) bandit problem, we start with the most basic setting of standard bandits, the stochastic multi-armed bandit (MAB), where the action set $\mathcal{A}_t = \mathcal{A} := [1 : K]$ is time-invariant and finite, each arm i is associated with a fixed but unknown reward parameter μ_i , and the feedback upon pulling arm i is a noisy version of the associated reward parameter $f_t(i) = \mu_i + \eta_t$. There exist algorithms dealing with the stochastic MAB problem with both theoretical guarantee and practical implementation, including both frequentist (e.g. Upper Confidence Bound, see (Auer et al., 2002)) and Bayesian (e.g. Thompson Sampling, see (Russo and Van Roy, 2014)) approaches. Such problem is exceedingly well studied, and a plethora of methods with subtle differences have been established, as is well summarised in the recent textbook (Lattimore and Szepesvári, 2020).

The bandit problem, especially the MAB problem, has applications to diverse settings, including recommendation systems, clinical treatments, communication networks, and engineering design (Li et al., 2010; Villar et al., 2015; Hu et al., 2021). Although such setting has been thoroughly studied, many application domains such as the above have constraints accompanying the reward maximisation objectives. For instance, trial drugs have both positive (eg. curing a disease) and negative side-effects (headaches, nausea, etc) on a patient in a clinical trial, and it is as much in the interest of a patient to ensure that negative side effects are limited as it is to ensure that the drug is effective (Genovese et al., 2013). This scenario motivates the problem of choosing drug and dosage (arms) that have the maximum positive effect while ensuring that the side-effects remain below some safety threshold α . Since each patient responds differently, the observed response and the manifestation of side-effects for a specific patient can be modelled as random-variables, with the corresponding means representing population averages. Importantly, for such a scenario, safety must be accounted for in a per-round sense - it does no good to alternate between assigning ineffective placebos and effective but harmful doses; instead we need to ensure that individuals are not exposed to undue risk while accruing benefits.

We thus introduce the safety constrained multi-armed bandit problem, where each *arm*, $k \in [1 : K]$ is modelled by a tuple, consisting of a stochastic *reward*, of mean μ^k , and an associated stochastic *safety-risk*, of mean ν^k . Upon pulling an arm, the learner observes noisy instances of the reward and safety-risk. The learner is provided with a *tolerated risk level*, denoted α , and the goal of the *safe bandit problem* is to maximise the reward gained over the course of play, while ensuring that unsafe arms—those for which $\nu^k > \alpha$ —are not played too often. At each time step t , the play picks an action from the given action set $A_t \in [1 : K]$, and the environment generate stochastic feedbacks which are centered at the mean associated with the arm $r_t = \mu^{A_t} + \omega_t^0$,

$s_t = \nu^{A_t} + \omega_t^1$. The player then updates its strategy based on the observations, and the game proceeds to the next round.

We measure the performance of the bandit algorithm via two metrics, the cumulative regret on rewards \mathcal{E}_T , and the net constraint violation \mathcal{S}_T , defined as follows.

$$\mathcal{E}_T = \sum_{t=1}^T (\mu^* - \mu^t)_+ \quad (2.1)$$

$$\mathcal{S}_T = \sum_{t=1}^T (\nu^t - \alpha)_+ \quad (2.2)$$

where $(\cdot)_+ = \max\{\cdot, 0\}$, $\mu^* = \max_{i \in [1:K]} \mu^i$ s.t. $\nu^i \leq \alpha$, and $\mu^t = \mu^{A_t}$, $\nu^t = \nu^{A_t}$.

The key point of the regret formulation lies in the fact that it only count the positive part, thus encouraging algorithms that do not alternate between effective and safe actions.

The constrained MAB setting, although intuitive and easy to understand, bears the major limitation of the finiteness (discreteness) of the action set. In bandit literature, it is well studied that infinite (continuous) action set could be considered, when certain structure is imposed in the action space. (Stochastic) linear bandit is a representative setting, where the observation is linear in the action, and all of the actions share a fixed and hidden parameter. We thus generalize the safety constrained multi-armed bandits to the constrained linear bandit setting, specified as follows.

At each time step t , the player picks action x_t (here we change the notation to obey the convention of linear bandit literature) from a known polytope $\mathcal{X} = \{x \in \mathbb{R}^d : Bx \leq \beta\}$. The environment then generate stochastic feedback on the unknown reward $r_t = \langle \theta, x_t \rangle + \omega_t^0$, and on several unknown safety constraints, $s_t^i = \langle a^i, x_t \rangle + \omega_t^i$, where θ and $a^i, i \in [1 : U]$ are the unknown reward and risk parameters, $\omega_t^i, i \in [0 : U]$ are observation noises. An arm $x \in \mathcal{X}$ is deemed safe if $\langle a^i, x \rangle \leq \alpha^i$ for all $i \in [1 : U]$. To keep the notations short, we stack all all the unknown parameters and corresponding

thresholds into a matrix A and vector α , so that the unknown constraints can be expressed as $Ax \leq \alpha$, and the safety polytope is naturally introduced as $\mathcal{S} = \{x \in \mathcal{X} : Ax \leq \alpha\}$. As the case in safe MAB, we measure the performance of an algorithm through the following metrics

$$\mathcal{E}_T := \sum_{t=1}^T (\langle \theta, x^* - x_t \rangle)_+,$$

$$\mathcal{S}_T := \sum_{t=1}^T \max_i (\langle a^i, x_t \rangle - \alpha_i)_+,$$

where $x^* = \arg \max_{x \in \mathcal{X}} \langle \theta, x \rangle$ s.t. $Ax \leq \alpha$.

The goal in both safe MAB and safe LB is to design an algorithm that achieves sublinear rates in both \mathcal{E}_T and \mathcal{S}_T .

2.2 Literature Review

The bandit problem, originated from the seminal work ([Thompson, 1933](#)), has been extensively studied and widely applied. We briefly summarize some important and most pertinent work, while the curious readers are referred to a series of representative surveys and textbooks, including ([Cesa-Bianchi and Lugosi, 2006](#); [Bubeck et al., 2012](#); [Lattimore and Szepesvári, 2020](#)).

There are mainly two types of bandit problems considered in the literature, the stochastic bandit and the adversarial bandit, categorised by the nature of the environment. In stochastic bandit, the distribution of the feedback associated with an arm is fixed but unknown, and the agent needs to estimate and exploit these distributions; while in adversarial bandit, the environment provide *any* feedback, oftentimes adversarial, and hence the agent needs to deal with worst case feedback structure. In this dissertation, we mainly consider the stochastic bandit setting. For the sake of simplicity, we omit the terminology 'stochastic' hereinafter.

The simplest setting of bandit problems is the multi-armed bandit (MAB) problem, where the candidate actions form a finite and discrete set. The MAB problem is properly introduced in (Robbins, 1952), and thoroughly developed by a series of follow-up work. Classic approaches of dealing with MAB problems include frequentist methods like explore-then-commit (ETC) and upper confidence bound (UCB), and Bayesian methods like Thompson Sampling (Agrawal and Goyal, 2012). Refined analysis are further studied, including (Kaufmann et al., 2012a; Kaufmann et al., 2012b), providing tighter regret bounds on the regret.

Beyond multi-armed setting, linear bandit (LB) (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011) is also well-established to study the bandit problem with infinite arms under linear structure. The confidence set introduced in (Abbasi-Yadkori et al., 2011) is widely used as a building block for many linear bandit algorithms.

In this dissertation, we study a family of bandit problems called (safety) constrained bandits. A key point of our setting is that the constraints need to be considered in a per-round sense, such that any constraint violation in a specific round should be noted and counted towards a cumulative objective. We briefly summarize the literature on constrained bandits, with comparison to our setup. To begin with, let's discuss previous methods for dealing with constrained bandit problems in terms of their formulation. One significant observation is that earlier approaches often impose constraints on overall play in a cumulative manner. This becomes problematic when our objective is to guarantee safety on a per-round basis. Next, we'll provide context for our proposed methodologies in relation to the existing body of work. Finally, we'll discuss the topic of pure exploration within the context of safe bandit scenarios, an area that has garnered recent attention.

In the literature on constrained bandit problems, there are two main categories

based on how constraints are handled: aggregate constraints and round-wise constraints.

A significant portion of constrained bandits has focused on aggregate constraint enforcement. This involves controlling the overall regret, denoted as \mathcal{R}_T , and aggregate violation, \mathcal{A}_T . The regret \mathcal{R}_T is defined as the cumulative sum of the differences in expected rewards between the optimal action (x^*) and the action actually taken (x_t) across all rounds up to T . Mathematically, it is represented as $\mathcal{R}_T = \sum_{t \leq T} \mu^* - \mu^t$ in the MAB case, and $\mathcal{R}_T := \sum_{t \leq T} \langle \theta, x^* - x_t \rangle$ in the LB case. While \mathcal{A}_T measures the maximum total violation of constraints over time and is expressed as $\mathcal{A}_T := \sum_{t \leq T} \nu^t - \mu^i$ $\mathcal{A}_T := \max_i (\sum_{t \leq T} \langle a^i, x_t \rangle - \alpha^i)$. The concept of bandits with aggregate constraints was first introduced in the work (Badanidiyuru et al., 2013) and later expanded upon by (Badanidiyuru et al., 2014; Agrawal and Devanur, 2014; Agrawal and Devanur, 2016; Agrawal et al., 2016; Sankararaman and Slivkins, 2021). These theories apply to the contexts where the total number of adverse effects needs to be constrained, while still attempting to match the performance of an optimal dynamic policy that has complete knowledge of all mean values. In more specific terms, (Badanidiyuru et al., 2013) considers a budget constraint, which can be mathematically represented as $\sum s_t - \alpha T \leq 0$, given that the total budget is $B = \alpha T$. On the other hand, (Agrawal and Devanur, 2014) offers a more relaxed approach. Instead of a hard constraint, they define a form of regret, which represents the maximum of zero and the excess of the total safety risk over αT . This is formulated as $\max(0, \sum s_t - \alpha T)$. The goal in this relaxed scenario is to ensure that this regret, which quantifies the extent to which the safety threshold is exceeded, remains minimal. These approaches represent different strategies in handling aggregate constraints in bandit problems, with the former emphasizing strict adherence to limits, and the latter allowing for some flexibility while still aiming to minimize risk overruns. These approaches ensure

that on average, over time, actions are both safe and effective. However, they do not guarantee safety in each individual round. The methods may alternate between very safe but ineffective actions and very unsafe but effective ones. Such a methodology might be suitable for scenarios where the average outcome over time is the primary concern, such as in resource management or budget allocation problems. However, in settings where safety is critical in each decision, like in healthcare or autonomous vehicle navigation, this approach is not appropriate.

The issue with these aggregate safety formulation in constrained bandit problems could lead to a significant violation of the constraints, which could be illustrated through the following example of 2-armed safe bandit. Arm 1 has a mean reward μ^1 of $1/2$ and no safety risk ($\nu^1 = 0$), while arm 2 has a higher mean reward ($\mu^2 = 1$) but comes with a significant safety risk ($\nu^2 = 1$). In the flavor of aggregate constraints, the optimal strategy would involve pulling arm 2, which has a higher reward but also a higher risk, for αT rounds – where α represents the safety threshold and T the total number of rounds. After reaching this threshold, the strategy would shift to pulling Arm 1, which is safer but less rewarding. Consequently, a low regret algorithm would require pulling arm 2 a proportionate number of times to T ($\Omega(T)$). This strategy, while optimal in the context of a global constraint, has a significant drawback: it exposes the agent to a large number of rounds with the high-risk arm 2, which is not desirable from a safety standpoint. To address this, our formulation proposes penalizing each use of arm 2 with a cost of $1 - \alpha$, effectively reducing its frequency of selection. As a result, a more effective approach under our model would involve playing arm 2 significantly fewer times, specifically less than proportionally to T (sublinearly).

Aggregate constraint formulation, particularly in the context of bandit problems and Markov Decision Processes (MDPs), continue to be a vibrant area of research. These methods generally focus on balancing the goal of achieving high rewards while

controlling the overall cost or risk involved, assessed over the entire course of action or decision-making trajectory. A notable segment of this research, often referred to as 'conservative bandits,' emphasizes strategies that are cautious or 'conservative' in nature. This line of work, introduced in (Wu et al., 2016), and developed by (Kazerouni et al., 2017; Garcelon et al., 2020a; Garcelon et al., 2020b), focuses exclusively on rewards and implements a continuous aggregate constraint throughout the process. Specifically, at any given round t , the cumulative reward up to that point, denoted as $\sum_{s \leq t} \mu^{A_s}$, is required to be at least $(1 - \alpha)t\mu^{k_0}$. Here, μ^{A_s} represents the mean reward of the action taken in round s , α is the predefined threshold, and μ^{k_0} is the mean reward of a baseline action k_0 . This conservative approach ensures that the accumulated reward at any point in the sequence does not fall significantly below a certain fraction of what would have been obtained by consistently choosing a baseline action. The objective is to maintain a steady level of reward performance while exploring other options. However, applying a similar running aggregate constraint to safety risks, as in the conservative bandit problem, could lead to issues similar to those observed in globally constrained formulations. If such a constraint were applied to safety risks, it would allow for the possibility of accumulating a significant amount of budget in earlier rounds, so that larger risk in later rounds could be offset by these safer choices. This approach could result in rounds being exposed to high levels of risk, which is undesirable in scenarios where maintaining safety in each individual round is crucial.

In constrained MDPs, the objective is to learn policies that not only yield high rewards but also keep the aggregate cost or risk within acceptable limits throughout the decision-making trajectory. Recent works like (Vaswani et al., 2022) contribute to this area by developing algorithms and strategies that balance reward maximization with cost control over the entire policy execution path. Some other studies, like (Yu

et al., 2017), approach these problems from a more optimization-focused angle. These works often integrate techniques and insights from the field of optimization to enhance the efficiency and effectiveness of bandit algorithms under constraints. In summary, aggregate constraint methods are crucial in scenarios where the overall performance over time, rather than the outcome of individual decisions, is the key focus. They find applications in various fields, from resource allocation to automated decision systems, where a long-term perspective on costs and benefits is essential.

An important aspect of our approach is the per-round application of the constraint, as opposed to an aggregated approach over all rounds. This means that the optimal dynamic policy in our scenario is more focused on consistently selecting a single, safer arm, rather than alternating between a safe and risky one. This per-round constraint enforces a more rigorous adherence to safety in each individual decision, contrasting with the aggregate approach that could allow for significant exposure to risk in the pursuit of higher rewards. The per-round formulation can be further divided into hard and soft enforcement strategies. The hard round-wise constraints require the knowledge of a very safe action and a safety margin. These methods build a pessimistic set of actions, ensuring high probability safety in each round. They pick actions by maximizing an optimistic reward index within this safe set. This approach was first considered in linear bandits (Amani et al., 2019) and further refined by (Moradipari et al., 2021). The main challenge here is the requirement of knowing a safe starting point and the dependence of performance on the safety margin.

(Pacchiano et al., 2021) presents a novel approach to the safe bandit problem, which differs significantly from previous methods in the following two key aspects.

- **Action Space Lifted to Policy Space:** Instead of focusing on individual arms, they consider policies over arms. These policies are represented as distributions over arms, denoted by π_t . This shift to policy-based decision-making allows for

a more nuanced approach to selecting actions, where a policy at any given time t dictates the probability distribution from which an arm is chosen.

- **"Hard" Per-Round Constraint:** The model enforces a stringent constraint on each round, formulated as $\langle \pi_t, \nu \rangle \leq \alpha$. Here, $\langle \pi_t, \nu \rangle$ represents the expected safety risk associated with the policy π_t , and α is a pre-defined safety threshold. This constraint ensures that the expected safety risk in each round does not exceed the threshold, prioritizing safety on a per-round basis.

The concept of regret in this model is also distinct. It is defined as the sum of the differences in expected rewards between the optimal static safe policy, π^* , and the actual policy used in each round, π_t . The optimal policy π^* is identified as the one that maximizes the expected reward $\langle \pi, \mu \rangle$ while ensuring that the expected safety risk $\langle \pi, \nu \rangle$ remains within the acceptable limit α .

Exploration within this framework is facilitated by incorporating a known safe arm k_s and utilizing the flexibility provided by the difference between the actual safety risk of k_s and the threshold α as a margin for exploration in the policy π_t . This strategy allows for controlled exploration while maintaining adherence to the safety constraint.

Overall, the approach by (Pacchiano et al., 2021) offers a more dynamic and flexible method for addressing the safe bandit problem, with a strong emphasis on maintaining safety on a round-by-round basis and leveraging policy-based decisions for effective exploration within safety constraints. However, despite being constrained on a per-round basis, this formulation encounters challenges akin to those seen in globally constrained frameworks. Although it appears to ensure safety in each round, the optimal static policy (π^*) actually achieves safety only when viewed in the aggregate sense over time. This can be problematic, as illustrated by the previous example where the optimal policy π^* ends up being $(1 - \alpha, \alpha)$. In such a scenario, a low regret algorithm would be compelled to assign a significant probability to the high-risk arm

2 in most rounds. Consequently, this leads to a situation where approximately $\Omega(T)$ rounds (linear in time) are exposed to the high-risk arm, mirroring the issue found in globally constrained models.

In a related vein, the works ([Amani et al., 2019](#); [Moradipari et al., 2021](#)) in the linear bandit setting explore similar themes, but without adopting the policy action space. Their research also focuses on hard round-wise safety constraints. A key aspect of their approach is the utilization of a known safe action and the continuity of the action space to facilitate adequate exploration within the safety constraints.

The pure-exploration approach in safe bandit problems is also an active area of research, focusing on identifying arms that are not only nearly-optimal in terms of rewards but also nearly-safe. This approach is particularly relevant in contexts where safety is a critical factor, and the goal is to explore and identify the best possible options within safety constraints. ([Katz-Samuels and Scott, 2018](#)) explore this concept in the formulation of finite-armed bandits. Their work aim to identify the top arms that meet specific safety thresholds. This approach is particularly useful in scenarios where there are a limited number of options (arms), each with distinct characteristics and associated risks or rewards. ([Wang et al., 2021](#)) extend this exploration to a more complex setting where each arm comes with a continuous parameter that influences both the reward and safety. This addition of a continuous parameter adds a layer of complexity and realism to the model, as it mirrors real-world scenarios where choices exist on a spectrum. ([Camilleri et al., 2022](#)) further explore the concept in the linear bandit framework, which aligns closely with the structure discussed in our context. In their study, they focus on identifying the best feasible arm, assuming a finite and known set of possible actions. This research is significant in linear bandits, where actions and outcomes have a linear relationship, making the exploration for safe and effective arms more nuanced. A critical aspect to note in these studies, particularly

in (Camilleri et al., 2022), is that even though safety might not be strictly enforced during the learning phase, the methods developed are still geared towards identifying arms that balance safety and effectiveness. However, these methods typically provide safety assurances only to a certain level of precision, indicating a trade-off between exploration speed and the precision of safety guarantees. In summary, the pure-exploration approach in safe bandit problems is essential for scenarios where the primary objective is to explore and identify the best possible actions within given safety constraints.

Beyond the perspective of formulation, there is also the categorization of pessimistic-optimistic approach and doubly-optimistic approach. In the safety constrained bandit problem, the majority of existing solutions adopt a pessimistic-optimistic (PO) approach, focusing on hard constraint satisfaction. This means they aim to ensure with high probability that the total safety violation (\mathcal{E}_T) remains zero. Key contributions in this domain include (Amani et al., 2019; Moradipari et al., 2021; Pacchiano et al., 2021; Bernasconi et al., 2022). While this approach provides strong safety guarantees, there are notable limitations and trade-offs involved. One significant constraint of PO methods is the prerequisite of an explicitly defined, nontrivially sized safe region, known beforehand. This requirement can be a substantial assumption, especially in situations where such prior knowledge is not readily available or hard to ascertain. In addition, the efficacy guarantees provided by these methods are typically of the form $\mathcal{E}_T = O((M^s)^{-1}\sqrt{T})$, where M^s represents the measure of the size of the known safe region. This relation implies that the efficacy of the method is inversely proportional to the size of the safe region and directly proportional to the square root of the time horizon T . The focus of this dissertation, on the other hand, is to explore alternative methods that aim to improve upon these limitations. Specifically, we are interested in approaches that:

- **Reduce the dependency on having a predefined, nontrivially sized safe region** This could potentially broaden the applicability of the methods to a wider range of scenarios where such prior knowledge is limited or unavailable.
- **Improve upon the efficacy guarantees**, aiming for bounds that are more favorable and less dependent on the size of the safe region or the time horizon.

However, it's important to note that in pursuing these improvements, there is a trade-off in terms of safety guarantees. The methods we explore may offer 'soft' safety guarantees, where the total safety violation \mathcal{S}_T grows slower than the time horizon T , specifically $\mathcal{S}_T = o(T)$. This means that while the methods may allow for some safety violations, the frequency and severity of these violations diminish over time.

To conclude, the field of bandit problems with unknown constraints is diverse, with various approaches each suited to different aspects of the exploration-exploitation trade-off and the safety-reward balance. Our focus on round-wise constraint enforcement aims to address the limitations of aggregate constraint methods, particularly in scenarios where safety is a critical concern in every decision-making instance. Our methodology aims to strike a balance between the strength of safety guarantees and the practical limitations of requiring prior knowledge of safe regions, while also seeking to improve the efficacy of the solutions in the constrained bandit problem context.

Chapter 3

Constrained Multi-Armed Bandits

3.1 Problem Setup and Definitions

In the context of the (safety) constrained multi-armed bandit problem, we define an instance by specifying a risk level denoted as α within the range of $[0, 1]$. Additionally, we have a constant K , which represents the number of arms available, and a corresponding vector of probability distributions denoted as $(\mathbb{P}^k)_{k \in [1:K]}$. Each entry in this vector is supported within the range $[0, 1]^2$ and can be represented as a pair (R, S) , where R stands for reward, and S represents the safety risk. Furthermore, we associate two vectors, μ and ν , both within the range $[0, 1]^K$, which correspond to the mean reward and safety risk of each arm. In mathematical terms: for each arm k , we have $(\mu^k, \nu^k) := \mathbb{E}_{(R,S) \sim \mathbb{P}^k}[(R, S)]$. It's important to note that the reward R and safety risk S components of each arm do not necessarily need to be independent, and our results are designed to handle various dependence structures.

The scenario unfolds in rounds, denoted as $t \in \mathbb{N}$. At each time step t , the learner, represented by an agent implementing some algorithm for the bandit problem, makes a decision denoted as A_t , which corresponds to choosing an arm to "pull". Upon selecting an action, the learner receives samples $(R_t, S_t) \sim \mathbb{P}^{A_t}$ independently of the history. The learner's knowledge at time t is characterized by the information set $\mathcal{H}_{t-1} = \{(A_s, R_s, S_s) : s < t\}$, and the action A_t must be adapted to the filtration induced by this set. The learner has no knowledge of any specific properties of the probability distributions \mathbb{P}^k other than the fact that they are supported within the

range $[0, 1]^2$.

The competitor, representing the *best feasible arm* given the safety constraint and the mean vectors, is determined as:

$$k^* = \arg \max_{k \in [1:K]} \mu^k \text{ s.t. } \nu^k \leq \alpha,$$

whose corresponding mean reward and safety risk are denoted as μ^* and ν^* . We will use this notation throughout - for any symbol \mathfrak{h}^k , we set $\mathfrak{h}^* = \mathfrak{h}^{k^*}$. Without loss of generality, we assume that the optimization is feasible, and that k^* is unique. We define the efficiency gap Δ^k and the feasibility gap Γ^k for playing arm k as:

$$\Delta^k := 0 \vee (\mu^* - \mu^k), \quad \Gamma^k := 0 \vee (\nu^k - \alpha),$$

where $a \vee b := \max(a, b)$, and we also use $a \wedge b := \min(a, b)$ later on. It's worth noting that $\Delta^k \vee \Gamma^k > 0$ for $k \neq k^*$, according to the uniqueness of the optima arm.

The performance of a learner is assessed using the concept of (pseudo-) regrets as defined in Equation 2.1 and Equation 2.2. For the sake of simplicity, we further introduce the following concept of a "composed" regret, defined in Equation 3.1:

$$\mathcal{R}_T := \sum_{1 \leq t \leq T} \Delta^{A_t} \vee \Gamma^{A_t} \tag{3.1}$$

where Δ^{A_t} and Γ^{A_t} are defined as in the previous explanation. It is readily seen that $\max\{\mathcal{E}_T, \mathcal{S}_T\} \leq \mathcal{R}_T$. Hence upper bounding the composed regret will immediately give us an upper bound on both the efficiency and safety regret. We focus on bounding \mathcal{R}_T henceforth.

It's worth noting that the per-round regret, namely $\max(0, \mu^* - \mu^{A_t}, \nu^{A_t} - \alpha)$, has some appealing properties from a safety perspective. Specifically, it penalizes constraint violations on a per-round basis rather than globally. This means that violating the safety constraint at one time cannot be compensated for by being overly

cautious at another time, as would be the case if we considered the sum of excess safety risk $\sum(\nu^{A_t} - \alpha)$. Similarly, playing a suboptimal arm cannot be compensated for by playing an overly aggressive arm later, as would be the case if we studied the sum of suboptimality $\sum(\mu^* - \mu^{A_t})$. Controlling the regret \mathcal{R}_T implies that suboptimality and excess safety risk are small for most rounds individually, not just in aggregate.

Additionally, for each arm k , we introduce state variables N_t^k representing the number of times it has been played up to time t , and R_t^k and S_t^k representing the total rewards and safety risk incurred on those rounds. Specifically:

$$\begin{aligned} N_t^k &:= \sum_{s < t} \mathbb{I}\{A_s = k\}, \\ R_t^k &:= \sum_{s < t} \mathbb{I}\{A_s = k\} R_s, \\ S_t^k &:= \sum_{s < t} \mathbb{I}\{A_s = k\} S_s. \end{aligned}$$

Similarly, N_t^* , R_t^* , and S_t^* denote the corresponding variables for k^* . Note that $\mathcal{R}_t = \sum_{k \neq k^*} (\Delta^k \vee \Gamma^k) N_{t+1}^k$. We also use the notation $\hat{\mu}_t^k := R_t^k / N_t^k$ and $\hat{\nu}_t^k := S_t^k / N_t^k$ as the sample means.

To track the number of times an unsafe arm is played, we define \mathcal{U}_T as the sum of indicators that denote when $\nu^{A_t} > \alpha$:

$$\mathcal{U}_T := \sum_t \mathbb{I}\{\nu^{A_t} > \alpha\}.$$

Finally, we introduce the notation $d(a||b) := a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$ to represent the Kullback-Leibler (KL) divergence between Bernoulli distributions with means a and b . We also define the following one-sided values:

$$d_{<}(a||b) := d(a||b) \mathbb{I}\{a < b\},$$

$$d_{>}(a||b) := d(a||b) \mathbb{I}\{a > b\}.$$

Remark 3.1.1. While the formulation primarily focuses on a single safety constraint, it can be extended to handle multiple constraints. For example, one could introduce a safety-risk vector $S \in [0, 1]^d$ and require that the corresponding mean vectors ν^k should lie within a known safe set \mathcal{S} . Extensions of the methods described below would then control, for instance, $\sum \max(\mu^* - \mu^{A_t}, \text{dist}(\nu^{A_t}, \mathcal{S}))$. However, for clarity and ease of explanation, the focus here is on a single constraint.

3.2 Doubly Optimistic Strategies on a Frequentist Perspective

The concept of optimistic confidence bounds, as discussed in chapters 7-10 of ([Lattimore and Szepesvári, 2020](#)), is a foundational strategy in bandit algorithms. This approach is based on two key principles:

1. **Encouraging Exploration:** The method involves calculating an optimistic estimate of the potential reward for each arm, typically based on the observed data plus a confidence margin. By always selecting the arm with the highest optimistic estimate, the algorithm inherently encourages exploration. This is because arms with less certainty (i.e., fewer pulls) tend to have wider confidence intervals, resulting in higher optimistic estimates. Thus, arms that have not been explored much will naturally get selected, ensuring that the algorithm does not prematurely converge to a suboptimal arm due to lack of exploration.

2. **Efficiency Through Information Utilization:** As more data is collected for each arm, the confidence bounds become tighter, converging closer to the true mean rewards of the arms. This process allows the algorithm to gradually distinguish between optimal and suboptimal arms. Arms that are less rewarding will eventually show lower optimistic bounds as their estimates become more accurate, leading to less frequent selection. Consequently, over time, the algorithm becomes more efficient, focusing more on the arms that are more likely to yield higher rewards.

This principle is fundamental to many bandit algorithms, including the Upper Confidence Bound (UCB) algorithm, and has been proven to be effective in ensuring

Algorithm 1 Doubly Optimistic Confidence Bounds

```

1: Input:  $K$ , functions  $U, L$ .
2: Initialise:  $\mathcal{H}_0 \leftarrow \emptyset$ 
3: for  $t = 1, 2, \dots$  do
4:   if  $t \leq K$  then
5:      $A_t \leftarrow t$ 
6:   else
7:      $\forall k, L_t^k \leftarrow L(t, \mathcal{H}_{t-1}, k)$ .
8:      $\Pi_t \leftarrow \{k : L_t^k \leq \alpha\}$ .
9:      $\forall k \in \Pi_t, U_t^k \leftarrow U(t, \mathcal{H}_{t-1}, k)$ .
10:     $A_t \leftarrow \arg \max_{k \in \Pi_t} U_t^k$ .
11:   end if
12:   Pull  $A_t$ , receive  $(R_t, S_t) \sim \mathbb{P}^{A_t}$ .
13:   Update  $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$ .
14: end for

```

both robust exploration and efficient exploitation in various bandit problem settings.

The concept of doubly optimistic bounds follows a similar principle. In this approach, lower bounds on safety-risk denoted as L_t^k and upper bounds on rewards denoted as U_t^k are maintained such that $L_t^k \leq \nu^k$ and $U_t^k \geq \mu^k$ hold with high probability. A set of "permissible arms" Π_t is constructed, consisting of arms that are considered feasible based on the available information up to time t . The action A_t is chosen to maximize U_t^k among arms in Π_t . This optimism in Π_t encourages exploration for high rewards, but as L_t^k becomes more concentrated with increasing observations (N_t^k), it helps identify unsafe arms, which are then avoided.

The algorithm for doubly optimistic confidence bounds is summarized in Algorithm 1. It operates as follows: Initialize with an empty history \mathcal{H}_0 . For each time step t : If $t \leq K$, select action $A_t = t$ (initial exploration); If $t > K$, update L_t^k for all arms k based on the available history; Define Π_t as the set of arms that satisfy the safety constraint; Compute U_t^k for all arms k in Π_t based on the available history; Select action A_t as the arm with the highest U_t^k among those in Π_t ; Pull arm A_t and observe rewards (R_t, S_t) from \mathbb{P}^{A_t} ; Update the history \mathcal{H}_t with the new observations.

The analysis of this scheme is based on a similar idea of the UCB type of analysis. To control the play of unsafe arms, it can be argued that $\nu^k - L_t^k$ is bounded as $\sqrt{\log(T)/N_t^k}$, which implies that arms violating the safety constraint will fall out of Π_t after being played a sufficient number of times. Additionally, the bounds are shown to be consistent (namely, $L_t^* \leq \nu^*$ and $U_t^* \geq \mu^*$) and optimistic with high probability. As a result, arms are played only if their upper bounds exceed the mean reward of the best arm, and the number of plays is limited by the gap between the upper bound and the mean.

The scheme presented in the previous discussion is analyzed using confidence bounds based on KL-UCB (Kullback-Leibler Upper Confidence Bounds, see ([Garivier and Cappé, 2011](#))), which are known to provide optimal mean-dependent regret control for standard bandit problems. These KL-UCB bounds are a natural choice for random variables supported on $[0, 1]$ and offer a strong foundation for analyzing the performance of the algorithm in the context of safe bandit problems.

It is important to note that the field of bandit algorithms has seen extensive research on various types of confidence bounds, and different choices of bounds can be made depending on the problem at hand. While the analysis in this case uses KL-UCB bounds, it is acknowledged that there are other types of confidence bounds that can be utilized to achieve similar goals, such as Empirical-KL-UCB ([Cappé et al., 2013](#)) and UCBV ([Audibert et al., 2009](#)).

The choice of which type of confidence bounds to use depends on the specific characteristics of the problem and the underlying distributions of the rewards. Different bounds may offer better performance in different scenarios, and researchers often select the most suitable bounds for their particular application.

The KL-UCB type bounds is summarized as follows:

$$\begin{aligned}\gamma_t &:= \log t + 3 \log \log t, \\ U(t, \mathcal{H}_{t-1}, k) &:= \max\{q > \hat{\mu}_t^k : d(\hat{\mu}_t^k \| q) \leq \gamma_t / N_t^k\}, \\ L(t, \mathcal{H}_{t-1}, k) &:= \min\{q < \hat{\nu}_t^k : d(\hat{\nu}_t^k \| q) \leq \gamma_t / N_t^k\},\end{aligned}$$

where γ_t serves as a trade-off parameter that balances the width and consistency of the upper and lower bounds. These bounds are natural for Bernoulli random variables, and since these are the ‘least-concentrated’ law on $[0, 1]$, the fluctuation bounds extend to general random variables. We show the following result based on the KL bounds.

Theorem 3.2.1. *Algorithm 1 instantiated with KL-UCB type bounds attains the following for any T and any $\varepsilon > 0$.*

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k \neq k^*} \frac{(1 + \varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_{<}(\mu^k \| \mu^*) \vee d_{>}(\nu^k \| \alpha)} + \xi_k,$$

where $\xi_k = O(\log \log T + \varepsilon^{-2})$. Further, the number of times an unsafe arm is played is bounded as

$$\mathbb{E}[\mathcal{U}_T] \leq \sum_{k: \Gamma^k > 0} \left(\frac{(1 + \varepsilon) \log T}{d_{<}(\mu^k \| \mu^*) \vee d_{>}(\nu^k \| \alpha)} \right) + \xi_k.$$

The O in the above hides instance-dependent constants, the most pertinent of which is a dependence on $(\Delta^k \vee \Gamma^k)^{-3}$ with the ε^{-2} term. To ameliorate this, we also give a gap-independent analysis of the scheme.

Theorem 3.2.2. *Algorithm 1 instantiated with KL-UCB attains*

$$\mathbb{E}[\mathcal{R}_T] \leq \sqrt{28KT \log T} + 6K \log \log T + 32.$$

The statement that the discussed analysis using KL-UCB type bounds extends to standard bandits when sending $\alpha \rightarrow 1$ is indeed an interesting observation. This means that the same analysis and algorithm can be applied to standard bandit problems by setting the safety constraint α to be very close to 1. In standard bandit problems,

where there is no explicit safety constraint, the focus is typically on maximizing the cumulative reward without considering safety concerns. However, the mentioned analysis, originally designed for safe bandit problems, can still be used in the standard bandit setting as a generalization. This observation highlights the flexibility of the analysis and the algorithm based on KL-UCB type bounds. It implies that the same methodology can be applied to both standard and safe bandit problems, making it a versatile approach for various types of sequential decision-making tasks.

It’s worth noting that while the algorithm can be used in standard bandit problems, its performance and efficiency may depend on the specific problem instance and the degree of safety constraint enforced by α . Nevertheless, this insight provides a valuable connection between safe and standard bandit problems and underscores the adaptability of the analysis.

3.3 Bayesian Methods

Thompson Sampling (TS) is a pioneering method for addressing bandit problems, introduced by (Thompson, 1933). It promotes exploration through randomization. The core concept involves selecting a benign prior and playing arms based on their posterior probability of being optimal. Insufficiently explored arms maintain flat posteriors, ensuring a non-negligible chance of being chosen. A key advantage of TS is its exploitation of a posterior that may be more closely aligned with the underlying law \mathbb{P}^k compared to confidence bounds that are based on statistics. Empirical studies have shown that TS often outperforms comparable Upper Confidence Bound (UCB) methods in multi-armed bandit scenarios (Chapelle and Li, 2011).

In this section we investigate the application of Bayesian methods to safe bandits. We begin by substituting the KL-UCB -based arm selection in Algorithm 1 while keeping the construction of Π_t unchanged. We then explore a Bayesian approach for

selecting Π_t , drawing inspiration from the work of (Kaufmann et al., 2012a).

Our analysis focuses on Bernoulli bandits, where the laws \mathbb{P}^k dictate that rewards R and safety-risks S are distributed according to Bernoulli distributions with parameters μ^{A_t} and ν^{A_t} , respectively. It is important to note that the derived bounds, being dependent only on the means of rewards and safety-risks, are applicable to general laws supported on $[0, 1]^2$. As observed by (Agrawal and Goyal, 2012), an algorithm designed for Bernoulli bandits can be adapted for general laws by feeding it samples $\tilde{R}_t \sim \text{Bern}(R_t)$ and $\tilde{S}_t \sim \text{Bern}(S_t)$. These transformed variables \tilde{R}, \tilde{S} remain Bernoulli with identical means, thus extending any mean-dependent guarantees to the broader bandit problem. However, this approach may increase variances, potentially leading to inefficiency in cases with highly concentrated distributions.

For brevity, we omit explicit bounds on $\mathbb{E}[\mathcal{U}_T]$ and the gap-independent bounds in the subsequent discussion, as they are essentially equivalent to those in Theorems 3.2.1 and 3.2.2.

3.3.1 Thompson Sampling with Optimistic Safety Indices

For Bernoulli bandits, using the Beta distribution for priors is a natural choice due to its conjugacy properties. The standard TS approach initializes each arm with a non-informative prior, $\text{Beta}(1, 1) = \text{Unif}[0, 1]$. The corresponding posterior at time t can be easily derived as $\text{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$.

Algorithm 2 outlines the proposed strategy. We retain the optimistic lower bound from Algorithm 1 but modify the arm selection process within Π_t to follow a TS approach: for each arm in Π_t , we draw random scores ρ_t^k from their respective posteriors, and the arm with the highest score ρ_t^k is selected.

Analyzing this method is straightforward, provided an existing analysis of TS for standard bandits. As a matter of fact, we can control the selection of infeasible arms as in Algorithm 1. Moreover, as long as $k^* \in \Pi_t$ holds true with high probability, we

Algorithm 2 Thompson Sampling With Optimistic Safety Indices (TOPSI) for Bernoulli Bandits

- 1: **Input:** K , function L .
 - 2: **Initialise:** $\mathcal{H}_0 \leftarrow \emptyset$.
 - 3: **for** $t = 1, 2, \dots$ **do**
 - 4: **if** $t \leq N$ **then**
 - 5: $A_t \leftarrow t$
 - 6: **else**
 - 7: $\forall k, L_t^k \leftarrow L(t, \mathcal{H}_{t-1}, k)$.
 - 8: $\Pi_t \leftarrow \{k : L_t^k \leq \alpha\}$.
 - 9: $\forall k \in \Pi_t$, sample $\rho_t^k \sim \text{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$
 - 10: $A_t \leftarrow \arg \max_{k \in \Pi_t} \rho_t^k$.
 - 11: **end if**
 - 12: Pull A_t , receive $(R_t, S_t) \sim \mathbb{P}^{A_t}$.
 - 13: Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$.
 - 14: **end for**
-

can apply the following decomposition:

$$\mathbb{E}[N_{T+1}^k] \leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \mathbb{P}(A_t = k | k^* \in \Pi_t).$$

The first term in the above is controlled by the consistency of the lower confidence bound L_t^* . The second term, which is typical in standard bandit analyses, can be managed using any established TS analysis. We specifically employ the approach of (Agrawal and Goyal, 2013) to demonstrate the following result.

Theorem 3.3.1. *For Bernoulli Bandits, Algorithm 2 with a KL-UCB -type confidence bound achieves the following regret bound for any T and $\varepsilon > 0$:*

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k \neq k^*} \frac{(1 + \varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_{<}(\mu^k \| \mu^*) \vee d_{>}(\nu^k \| \alpha)} + \xi_k,$$

where $\xi_k = O(\log \log T + \varepsilon^{-2} \log(1/\varepsilon))$.

3.3.2 Thompson Sampling with BAYESUCB

While Algorithm 2 provides a rigorous analysis, it still relies on a potentially less precise frequentist bound for determining Π_t . Utilizing the posteriors on safety-risks might enhance performance.

One might consider applying the basic principles of Thompson Sampling, associating a posterior $P_{t,\nu}^k = \text{Beta}(S_t^k + 1, N_t^k - S_t^k + 1)$ with the safety risk, sampling safety scores $\sigma_t^k \sim P_{t,\nu}^k$, and setting $\Pi_t = \{k : \sigma_t^k \leq \alpha\}$. However, this approach is fundamentally flawed, primarily because it involves comparing scores to a fixed level α rather than among themselves. If $\nu^* = \alpha$, there is a constant probability that $\sigma_t^* > \alpha$, even when the empirical mean $\widehat{\nu}_t^*$ is accurate. This could result in a consistent selection of suboptimal arms, leading to linear regret. A similar problem was noted in the analysis of TS using a UCB-type framework (Kaufmann et al., 2012b), but here the issue lies in the scheme itself. Simulations confirm that when $\nu^* = \alpha$, this approach leads to linear expected regret.

To address this, one might introduce a slack variable, β_t^k , so that $\Pi_t = \{\sigma_t^k \leq \alpha + \beta_t^k\}$. This β_t^k should decrease as N_t^k increases but remain large enough to ensure $k^* \in \Pi_t$. This approach, akin to (Kaufmann et al., 2012b)'s analytical method, effectively designs a confidence bound, somewhat negating the original intent.

Our approach circumvents this issue by employing a *Bayesian confidence bound*, leveraging the BAYESUCB method of (Kaufmann et al., 2012a). We choose the δ_t^k th quantile of the posterior $P_{t,\nu}^k$ as the safety score, where δ_t^k decays with t . This method harnesses the adaptability of the posterior, while ensuring optimistic scores due to the small δ_t^k , thus maintaining a high likelihood of including $k^* \in \Pi_t$. As N_t^k increases, the scores for unsafe arms converge to ν^k , eventually excluding them. This method appears particularly suited for our context of arm filtering at a specified level. Algorithm 3 details the scheme, where $Q(P, \delta)$ denotes the δ th quantile of law P . We

Algorithm 3 Thompson Sampling with BAYESUCB (TSBU) for Bernoulli Bandits

- 1: **Input:** K , schedule δ_t^k .
 - 2: **Initialise:** $\mathcal{H}_0 \leftarrow \emptyset$.
 - 3: **for** $t = 1, 2, \dots$ **do**
 - 4: $\forall k$
 - 5: **if** $S_t^k = 0$ **then**
 - 6: $L_t^k \leftarrow 0$
 - 7: **else**
 - 8: $L_t^k \leftarrow Q(\text{Beta}(S_t^k, N_t^k - S_t^k + 1), \delta_t^k)$.
 - 9: **end if**
 - 10: $\Pi_t \leftarrow \{k : L_t^k \leq \alpha\}$.
 - 11: $\forall k \in \Pi_t$, sample $\rho_t^k \sim \text{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$
 - 12: $A_t \leftarrow \arg \max_{k \in \Pi_t} \rho_t^k$.
 - 13: Pull A_t , receive $(R_t, S_t) \sim \mathbb{P}^{A_t}$.
 - 14: Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$.
 - 15: **end for**
-

slightly bias the method for technical convenience.

The following summarizes our analysis of Algorithm 3.

Theorem 3.3.2. *For Bernoulli bandits, Algorithm 3, instantiated with $\delta_t^k = (\sqrt{8N_t^k t} \log^3 t)^{-1}$ attains the following regret bound for any $\varepsilon > 0$ and any T :*

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k \neq k^*} \frac{(1 + \varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_{<}(\mu^k \|\mu^*) \vee^{2/3} \cdot d_{>}(\nu^k \|\alpha)} + \xi_k,$$

where $\xi_k = O(\log \log T + \varepsilon^{-2} \log(1/\varepsilon))$

3.4 Lower Bound

We wrap up our theoretical investigation with a lower bound for algorithms that achieve sub-polynomial regret across all bounded distributions. This foundation is built upon the technique of (Garivier et al., 2019), which leverages the chain rule of KL divergence and the data processing inequality. The following relation, directly applicable to our scenario, is established:

Lemma 3.4.1. *For any safe bandit algorithm, and any two safe bandit instances $\{\mathbb{P}^k\}, \{\tilde{\mathbb{P}}^k\}$, and any $T, k_0 \in [1 : K]$,*

$$\sum_k \mathbb{E}[N_{T+1}^k] D(\mathbb{P}^k \|\tilde{\mathbb{P}}^k) \geq d(\mathbb{E}[N_{T+1}^{k_0}/T] \|\tilde{\mathbb{E}}[N_{T+1}^{k_0}/T]).$$

This lemma facilitates a conventional strategy - selecting $\tilde{\mathbb{P}}$ such that $\tilde{\mathbb{E}}^k[(R, S)] = (\mu^k \vee \mu^* + \varepsilon, \nu^k \wedge \alpha)$, while maintaining the other \mathbb{P}^k s unaltered. For bandit algorithms with sub-polynomial regret, the right side of the inequality scales as $\log(T)$, and the left simplifies to $\mathbb{E}[N_{T+1}^k] D(\mathbb{P}^k \|\tilde{\mathbb{P}}^k)$. Although the optimal selection of $\tilde{\mathbb{P}}$ is nuanced and depends on the specifics of \mathbb{P} , we explore a straightforward case to demonstrate the robustness of our previous analyses.

Proposition 3.4.2. *Any algorithm that ensures that, uniformly over all instances of safe Bernoulli bandit problems with independent rewards and safety-risks, the mean number of plays of any suboptimal arm is bounded as $O(T^x)$ for every $x \in (0, 1)$ must satisfy*

$$\liminf_{T \nearrow \infty} \frac{\mathbb{E}[N_{T+1}^k]}{\log T} \geq \frac{1}{d_{<}(\mu^k \|\mu^*) + d_{>}(\nu^k \|\alpha)}$$

Given that mean regret is expressible in terms of $\mathbb{E}[N_{T+1}^k]$, this also establishes a lower bound for regret. It is noteworthy that the denominator uses a sum rather than a maximum, as seen in our upper bounds. This implies that for strictly dominated arms (i.e., arms k where $\Delta^k \Gamma^k > 0$), our bounds could be off by a factor of up to two. This discrepancy arises because our approach does not capitalize on potential dependencies between the reward R and safety-risk S , presenting a pathway for future research endeavors.

Chapter 4

Constrained Linear Bandits

4.1 Problem Setup and Definitions

Notations: For natural numbers $a \leq b$, let $[a : b] := \{a, \dots, b\}$. In \mathbb{R}^d , $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and ℓ_2 -norm, respectively. For a matrix $V \succ 0$, the norm $\|z\|_V$ is defined as $\sqrt{\langle z, Vz \rangle}$. For a $p \times q$ matrix M , and a set $\mathfrak{S} \subset [1 : p]$, $M(\mathfrak{S})$ denotes the $|\mathfrak{S}| \times q$ submatrix of M , preserving rows indexed in \mathfrak{S} only and discarding the rest.

Setting: The safe linear bandit (SLB) problem is parameterized by an objective vector θ , U unknown constraint vectors $\{a^i\}_{i \in [1:U]}$, and K known constraint vectors $\{b^j\}_{j \in [1:K]}$, all in \mathbb{R}^d . It also involves constraint levels $\{\alpha^i\}_{i \in [1:U]}$, $\{\beta^j\}_{j \in [1:K]}$. These elements define the principal linear program as follows:

$$\max \langle \theta, x \rangle \text{ s.t. } Ax \leq \alpha, Bx \leq \beta$$

The matrices A, B and vectors α, β stack the linear constraints. The known constraints $Bx \leq \beta$ arise from predetermined limits on x (e.g., dosage levels must not exceed toxicity thresholds determined in model organisms). We assume a bounded polytope and feasibility of the program, together with a unique optimum x^* . The levels α, β , and the matrix B are known to the learner, whereas the objective θ and the unknown constraints A are hidden. The *feasible set* is defined as $\mathcal{S} := \{x : Ax \leq \alpha, Bx \leq \beta\}$. Unlike prior works, we do not assume a known feasible solution (for example, a placebo

$x = 0$) since α may be negative.

Mechanism: The problem proceeds in rounds, indexed by time step t . In each round t , the learner chooses an action from the domain $x_t \in \mathcal{X} := \{x \in \mathbb{R}^d : Bx \leq \beta\}$, receiving reward feedback r_t and safety feedback $\{s_t^i\}_{i \in [1:U]}$:

$$r_t = \langle \theta, x_t \rangle + w_t^0, \quad \text{and} \quad s_t^i = \langle a^i, x_t \rangle + w_t^i,$$

where w_t^i are conditionally centered sub-Gaussian noises. The learner's information set at time t is \mathcal{H}_{t-1} , comprising previous actions and feedback.

Performance Metrics: We aim to control the cumulative *Efficiency Regret* and *Safety Violations*, defined earlier in § 2.1 (and repeated below for convenience). Penalizing only the positive part of round-wise inefficiency or safety violations is crucial in the SLB context to prevent inter-round tradeoffs of safety violations and efficiency gain.

$$\mathcal{E}_T := \sum_{t=1}^T (\langle \theta, x^* - x_t \rangle)_+,$$

$$\mathcal{S}_T := \sum_{t=1}^T \max_i (\langle a^i, x_t \rangle - \alpha_i)_+,$$

Assumptions. As per (Abbasi-Yadkori et al., 2011), we make the following assumptions, which are standard in linear bandit literature:

1. Boundedness: $\|\theta\| \leq 1, \|a^i\| \leq 1$ for all i , and $\{Bx \leq \beta\} \subset \{\|x\| \leq 1\}$.
2. Sub-Gaussianity: For all t, i , the w_t^i are conditionally centered and 1-sub-Gaussian given $\sigma(\mathcal{H}_{t-1}, x_t)$, that is that is $\forall i \in [0 : U]$, and $\mathcal{F}_t := \sigma(\mathcal{H}_{t-1}, x_t)$

$$\mathbb{E}[w_t^i | \mathcal{F}_t] = 0, \quad \mathbb{E}[\exp(\xi \eta_t^i) | \mathcal{F}_t] \leq \exp(\xi^2/2), \quad \forall \xi \in \mathbb{R}.$$

Subsequent results are valid under these assumptions. These assumptions are in fact streamlined versions of those typically found in linear bandit literature. For

the completeness, we take a deeper look into these assumptions and explore their implications.

Boundedness. This assumption encompasses two facets: the boundedness of underlying parameters (i.e., $\|\theta\|, \|a^i\| \leq 1$) and the bounded domain ($\|x\| \leq 1$ for all $x \in \mathcal{X} = \{Bx \leq \beta\}$). The bounded domain is crucial for ensuring that the value of the optimization problem is finite. While we use $\|x\| \leq 1$, this could be generalized to $\|x\| \leq L$ without significantly altering the conclusions. This modification primarily influences the choice of regularizer λ in our algorithm DOSLB, where $\lambda \geq L^2$ is required for the validity of Lemma 4.2.1. The bounds on parameters, while standard, can be adjusted to $\|\theta\|, \max_i \|a^i\| \leq S$, affecting only the additive term in the confidence radius ω_t . Adapting to varying norms of a^i and θ is possible, as shown in recent studies like (Gales et al., 2022).

SubGaussianity. The assumption of 1-subGaussian noise, while offering significant technical convenience, can be expanded to R -subGaussian conditions. Altering this assumption impacts ω_t , scaling the first term which grows with t . This scaling is more substantial than the changes induced by modifying parameter bounds, as it directly affects the dynamic component of ω_t .

Overall Confidence Radius with General Parameters. Under more generic conditions ($\|x\| \leq L, \|\theta\| \leq S, \|a^i\| \leq S$, and R -subGaussian noise), our analysis remains applicable but with expanded confidence radii:

$$\sqrt{\omega_t(\delta; L, S, R)} = R \sqrt{\frac{1}{2} \log \left(\frac{(U+1) \det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + S \lambda^{1/2},$$

under the condition $\lambda \geq L^2$. This adjustment would typically increase the regret bounds by a factor of $\max(R, S)$ and alter the logarithmic terms to $\log(1 + TL^2/\delta)$ instead of $\log(1 + T/\delta)$. For simplicity, our subsequent analysis will maintain the default parameters $R = S = L = 1$.

4.2 Doubly Optimistic Play for SLB

The optimism principle proves to be highly effective in bandit problems, as demonstrated by (Lattimore and Szepesvári, 2020). Specifically, it shines in the domain of stochastic linear bandits, as shown in the works (Dani et al., 2008) and (Abbasi-Yadkori et al., 2011). The quintessential approach for addressing such problems is the 'OFUL' algorithm. This algorithm constructs confidence sets denoted as \mathcal{C}_t^0 for the objective parameter vector θ . It then selects the action x_t by maximizing the optimistic objective $\max_{\theta \in \mathcal{C}_t^0} \langle \theta, x \rangle$. The success of these methods can be attributed to their ability to adaptively explore the reward: when an inefficient choice of x_t is made, it indicates that the corresponding direction in the action space has not been explored adequately, leading to significant improvements in parameter estimates.

Our method for stochastic linear bandits (SLBs) similarly relies on constructing confidence sets, but for both the reward and constraint vectors. We will begin by describing the confidence sets we utilize and then discuss the core algorithm.

4.2.1 Noise Scales and Confidence Sets

Our confidence sets draw inspiration from (Abbasi-Yadkori et al., 2011). They are constructed based on an analysis of the noise scales associated with regularized least squares (RLS) estimators. For a chosen value of $\lambda \geq 1$, we define $V_t = \lambda I + \sum_{\tau=1}^t x_\tau x_\tau^\top$. The RLS estimates for the parameters are given by:

$$\hat{\theta}_t = (X_{1:t}^\top X_{1:t} + \lambda I)^{-1} X_{1:t}^\top R_{1:t},$$

$$\hat{a}_t^i = (X_{1:t}^\top X_{1:t} + \lambda I)^{-1} X_{1:t}^\top S_{1:t}^i.$$

For convenience, we define $\sqrt{\omega_t(\delta)} := \sqrt{\frac{1}{2} \log \left(\frac{(U+1) \det(V_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2}$. The

confidence ellipsoids are then defined as:

$$\mathcal{C}_t^0(\delta) := \{\tilde{\theta} : \|\tilde{\theta} - \hat{\theta}_{t-1}\|_{V_{t-1}} \leq \sqrt{\omega_{t-1}(\delta)}\},$$

$$\mathcal{C}_t^i(\delta) := \{\tilde{a}^i : \|\tilde{a}^i - \hat{a}_{t-1}^i\|_{V_{t-1}} \leq \sqrt{\omega_{t-1}(\delta)}\},$$

for each $i \in [1 : U]$. The term $\|z\|_{V_t}$ is small for z aligned poorly with the past actions, which results in wider confidence sets \mathcal{C}_t for under-explored directions. Additionally, we define the 'matrix confidence set' as $\mathbf{C}_t(\delta) := \{\tilde{A} \in \mathbb{R}^{U \times d} : \forall i, \tilde{A}_{i,\cdot} \in \mathcal{C}_t^i(\delta)\}$ for notational convenience. These sets are primarily known for their consistency (Theorem 2 of (Abbasi-Yadkori et al., 2011)), implying that with high probability, θ is contained in \mathcal{C}_t^0 and A is within \mathbf{C}_t . We formally present this result as the following lemma.

Lemma 4.2.1. *The confidence sets are consistent, i.e.,*

$$\forall \lambda \geq 1, \delta \in (0, 1), \quad \mathbb{P}(\forall t, \theta \in \mathcal{C}_t^0(\delta), A \in \mathbf{C}_t(\delta)) \geq 1 - \delta.$$

In the following, we will often omit the dependence of $\mathcal{C}_t^i(\delta)$, $\mathbf{C}_t(\delta)$, $\rho_t(x; \delta)$ on δ , where

$$\rho_t(x; \delta) := 2\sqrt{\omega_{t-1}(\delta)}\|x\|_{V_{t-1}^{-1}}.$$

4.2.2 The DOSLB Algorithm

Algorithm 4 Doubly-Optimistic Safe Linear Bandit (DOSLB) (λ, δ)

Input: $\lambda > 0, \delta \in (0, 1)$
for $t = 1, 2, \dots$ **do**
 Construct $\tilde{\mathcal{S}}_t(\delta)$ as in (4.1).
 Optimize (4.2) and play x_t .
 Observe $r_{t,x_t}, \{s_{t,x_t}^i\}$
 Update $X, R, \{S^i\}, V, C$
end for

Now, let's discuss our algorithm, Doubly-Optimistic Safe Linear Bandit (DOSLB; Algorithm 4). This scheme keeps track of confidence sets $\mathcal{C}_t^0(\delta)$ for θ and $\mathbf{C}_t(\delta)$ for A

(as discussed in §4.2.1). Using these sets, it constructs an optimistic ‘permissible set’ of actions x that are safe with respect to at least one set of constraints in \mathbf{C}_t . This permissible set is defined as:

$$\tilde{\mathcal{S}}_t(\delta) := \{x : \exists \tilde{A} \in \mathbf{C}_t(\delta) \text{ s.t. } \tilde{A}x \leq \alpha, Bx \leq \beta\}. \quad (4.1)$$

In other words, $\tilde{\mathcal{S}}_t$ encompasses all potential actions that could be deemed safe based on the accumulated knowledge. The algorithm then selects the action x_t optimistically from this permissible set as follows:

$$(\tilde{\theta}_t, x_t) \in \arg \max_{\tilde{\theta} \in \mathcal{C}_t^{\theta}(\delta), x \in \tilde{\mathcal{S}}_t(\delta)} \langle \tilde{\theta}, x \rangle. \quad (4.2)$$

The primary differentiation between the Doubly-Optimistic (DO) and Pessimistic-Optimistic (PO) approaches lies in the construction of the permissible set, as mentioned in §2.2. The DO approach, as we’ve discussed, adopts an optimistic stance when forming the permissible set. This optimistic $\tilde{\mathcal{S}}$ promotes a much more vigorous exploration strategy, ultimately resulting in enhanced efficiency performance, as we will delve into in the subsequent analysis. Naturally, this increased aggressiveness comes at the expense of safety, and there is an inherent trade-off between the two. However, we will demonstrate through subsequent regret analysis that the DOSLB algorithm effectively manages and controls this trade-off, ensuring that the cost in terms of safety is well-contained.

It’s essential to note that this optimistic construction of the permissible set sets DOSLB apart from the Prior Optimism approach, which creates a more conservative permissible set (as discussed in §2.2). This optimistic $\tilde{\mathcal{S}}$ approach encourages more aggressive exploration, resulting in improved efficiency, as we will analyze further below. Naturally, this approach comes at the cost of safety performance, but it can be shown through subsequent regret analysis that DOSLB effectively manages this

trade-off.

4.3 Lower Bound and Hardness of the Problem

In standard linear bandits, where the safety set \mathcal{S} is known, adopting an optimistic approach leads to instance-dependent logarithmic regret bounds for a large T , as shown (Abbasi-Yadkori et al., 2011). These results hinge on the fact that when \mathcal{S} is known, any action chosen by an optimistic method lies within the *finite* set of extreme points of \mathcal{S} . Consequently, there exists a positive value Δ such that for any suboptimal action x chosen by the method, $\langle \theta, x^* - x \rangle \geq \Delta$. This nontrivial separation directly leads to regret bounds of $O(\log^2(T)/\Delta)$. However, in our case, the scenario becomes more complex when some constraints are unknown. It is a natural question to ask: can we still achieve logarithmic bounds in such cases?

Addressing this question requires a redefining of our expectations. Since \mathcal{S} is not fully known, we must consider not just the efficiency gap but also eliminate potentially unsafe points outside \mathcal{S} . We define the set of extreme points \mathcal{E} , which includes points meeting both known and unknown constraints as follows

$$\mathcal{E} := \{x : Bx \leq \beta, \exists \mathfrak{U} \subset [1 : U], \mathfrak{K} \subset [1 : K] : \text{rank} \begin{pmatrix} A(\mathfrak{U})^\top & B(\mathfrak{K})^\top \end{pmatrix} = d, A(\mathfrak{U})x = \alpha(\mathfrak{U}), B(\mathfrak{K})x = \beta(\mathfrak{K})\}.$$

In simpler terms, the finite set \mathcal{E} consists of points in the bounding polytope that activate d linearly independent known and unknown constraints. Notably, $x^* \in \mathcal{E}$. (Since there are only finite number of known and unknown constraints, the set \mathcal{E} is a finite set.) We will show, in this and the following section, that this set is of vital importance in identifying the discreteness in this continuously problem. We present a negative result (lower bound in this section) and a positive result (upper bound in the next section) to show that a quantity closely related to this set controls the rate of

the regret in this case.

An SLB instance is termed Δ -well-separated if every suboptimal point in \mathcal{E} is either inefficient or unsafe by a margin of at least Δ , which is to say, $\forall x \in \mathcal{E} \setminus \{x^*\}, \max\{\langle \theta, x^* - x \rangle, \max_i(\langle a^i, x \rangle - \alpha^i)\} \geq \Delta$. This means that every suboptimal point in \mathcal{E} is either nontrivially inefficient or unsafe. In this case, if all the potential extreme points are well separated, and the set \mathcal{E} is provided to the learner, it is immediate that we can derive $O(\log^2(T)/\Delta)$ bounds on both the efficiency regret and the safety violation. However, an SLB is in the face of unknown \mathcal{E} due to the unknown constraints, thus the question of interest is refined as: can we achieve logarithmic rate for both the efficiency regret and the safety violation? It turns out that the answer is counter-intuitively negative. We demonstrate this through the following theorem and example.

Theorem 4.3.1. *For any SLB algorithm, there exists a 1/8-well-separated instance where the algorithm incurs $\max(\mathbb{E}[\mathcal{E}_T], \mathbb{E}[\mathcal{S}_T]) = \Omega(\sqrt{T})$.*

This result asserts that for any SLB algorithm, there exists a 1/8-well-separated instance for which the algorithm must incur at least $\Omega(\sqrt{T})$ in either efficiency or safety. This means that logarithmic bounds cannot be universally achieved for all well-separated instances. Theorem 4.3.1 highlights a fundamental challenge in SLBs: the inability to precisely refine the location of the optimal point within \mathcal{S} . This difficulty is distinct from the challenges in standard bandits, where the noiseless problem often involves suboptimal points with similar performance.

The underlying obstacle is illustrated in the following example, which represents a simplified one-dimensional problem. In this problem, we maximize x subject to known constraints $0 \leq x \leq 1$ and an unknown constraint $ax \leq 1/4$, with $\theta = 1$, and $a \in \{(1 + \kappa)/2, (1 - \kappa)/2\}$, where $\kappa \in (0, 1/4)$ is a parameter to be determined later. It is straightforward to verify that this instance is indeed well-separated so long as $\kappa < 1/4$. Depending on the value of a , the extreme points of the sets differ. While

the estimation of a could be at best spread a segment of $1/\sqrt{t}$, it is unreasonable to eliminate any of the two options of a when $t < \kappa^{-2}$. In such scenario, if we eliminate $(1 - \kappa)/2$ when it's actually the truth, we suffer at least 2κ inefficiency; and reversely, if we wrongly eliminate $(1 + \kappa)/2$, we suffer at least 2κ safety violation. Thus $\max\{\mathcal{E}_T, \mathcal{S}_T\} = O(\kappa \cdot \min\{T, \kappa^{-2}\})$, which lead to a \sqrt{T} rate by picking $\kappa = 1/\sqrt{T}$. This inability to refine the precise location of the optimal point, rather than the presence of suboptimal points with similar performance, is the fundamental challenge preventing logarithmic control in SLBs compared to the standard linear bandits.

This result contrasts sharply with existing minimax lower bounds in standard bandits, which typically require setting $\Delta \sim T^{-1/2}$ to achieve $\Omega(\sqrt{T})$ bounds. In SLBs, the challenge is more about accurately pinpointing the optimal action within the feasible set, a problem exacerbated by the unknown constraints.

The subsequent sections will explore how the polytopal structure of SLBs, despite this fundamental challenge, induces a form of discreteness in the action space through index sets of constraints, which we will leverage in our analysis.

4.4 Identifying Discreteness through Basic Index Sets

Once we are at Theorem 4.3.1, an immediate question to ask is what rate is achievable. We shall now proceed to establish precise formal definitions for both the basic index sets (BIS) that enable us to encapsulate the inherent discreteness of DOSLB's actions, along with the gaps inherent in these index sets. These gaps ultimately result in a crucial implication: that 'suboptimal' index sets cannot be played frequently by DOSLB.

The construction and reasoning based on basic index sets (BIS) is inspired by the concept of basic feasible solution in linear programming (LP) theory. For a detailed treatment of LP theory, we refer the readers to the classic textbook ([Bertsimas and](#)

[Tsitsiklis, 1997](#)). In terms of LP theory, a basic feasible solution (BFS) is, in the non-degenerate case, a point where d linearly independent constraints meet. Each BFS corresponds to a vertex or an extreme point, from the geometric point of view. LP theory tells us that when we are considering a feasible, bounded, non-degenerate linear program, there must exist a BFS that attains the optimal value. Hence it suffices to examine the BFSs since there is only a finite amount. And indeed there exists even more efficient approach in this vein, called the simplex method.

Our study is essentially a linear program, with the difference that the feedbacks are noisy. The noisy observation structure of the unknown constraints adds to the complexity of the problem: if all the constraints are known, one could enumerate all the possible combinations of d constraints, figure out their intersection point, and verify one-by-one whether these points are BFS. While in our case, these points are not all known in advance, and under noisy feedback, the following could happen: 1) our estimated constraints (let's call them noisy constraints from now on) could intersect outside of the true domain \mathcal{S} ; 2) noisy constraints that intersect may not intersect at all were they noiseless; 3) noisy constraints that do not intersect may actually correspond to BFS in the noiseless program, etc.

To capture these subtlety, we need to introduce a series of concepts that deal with noise scales introduced by the stochastic bandit problem. The general idea is that, we measure the optimality of a set of constraints according to its original (noiseless) intersection point(s). If it's inefficient, the intersection should have an efficiency gap; if it's infeasible, it should have a feasibility gap. In order for a suboptimal BIS to be selected, it must be the case that the estimation noise already exceeds these gaps, so that it is mistaken as optimal. The subsequent developments are hinged on the core idea of controlling the noise scale.

4.4.1 Basic Index Sets

We define \mathcal{E} previously as a gentle start. To depict the behavior of these points in a quantitative way, we associate them with indices of constraints. Thus comes the definition of index set and basic index set.

Definition 4.4.1. *An index set is an ordered pair of sets $(\mathfrak{U}, \mathfrak{K})$ such that $\mathfrak{U} \subseteq [1 : U], \mathfrak{K} \subseteq [1 : K]$. An index set $I = (\mathfrak{U}, \mathfrak{K})$ is called a basic index set (BIS) if $|\mathfrak{U}| + |\mathfrak{K}| = d$.*

As a comparison to LP terminology, an index set is similar to the candidates for BFS, where we pick the solutions to all possible combinations of d constraints. While the BIS definition resembles that of BFS. (Note that to include all possibilities under noise, we drop the linear independence in the definition of BIS, but save it for later.) To capture the benignity of the BISs, we examine the performance of their associated points, defined as follows. It is worth noting that according to the definition of BIS, there is a possibility that each BIS corresponds to multiple points. Later on, based on this concept, we identify the points that are noisily associated with a BIS, taking the observation noise into consideration, and naturally derive a control over the noise scale.

Definition 4.4.2. *The set of points that activate an index set $I = (\mathfrak{U}, \mathfrak{K})$ is defined as*

$$\mathcal{X}^I := \{x \in \mathcal{S} : A(\mathfrak{U})x = \alpha(\mathfrak{U}), B(\mathfrak{K})x = \beta(\mathfrak{K})\}.$$

Note that we require the activation points to be feasible, meaning they must reside within the \mathcal{S} . With this requirement, the set \mathcal{X}^I could potentially be empty (no intersection within \mathcal{S}), consist of a single point (a unique intersection, and within \mathcal{S}), or even form an affine segment (infinite intersection within \mathcal{S}). The following terminology will prove to be beneficial in our discussion.

Definition 4.4.3. *A BIS I is called*

1. feasible if $\mathcal{X}^I \neq \emptyset$ and infeasible otherwise;
2. suboptimal if $x^* \notin \mathcal{X}^I$ and optimal otherwise;
3. full rank if the vectors $\{a^i\}_{i \in \mathcal{U}} \cup \{b^j\}_{j \in \mathcal{R}}$ span \mathbb{R}^d .

Throughout the bandit game, we encounter a challenge where we are compelled to operate not with the actual constraint matrix A , but rather with its imprecise (estimated-under-noise) counterparts, denoted as the \tilde{A} s. To address this uncertainty in the constraints, we broaden the concept of BIS (strict) activation to noisy activation. Note that the set $\tilde{\mathcal{X}}_t^I \subset \tilde{\mathcal{S}}_t$.

Definition 4.4.4. *The set of points that noisily activates an index set $I = (\mathcal{U}, \mathcal{R})$ at time t is*

$$\tilde{\mathcal{X}}_t^I := \{x \in \tilde{\mathcal{S}}_t : \exists \tilde{A} \in \mathcal{C}_t, \tilde{A}(\mathcal{U})x = \alpha(\mathcal{U}), B(\mathcal{R})x = \beta(\mathcal{R}), \tilde{A}x \leq \alpha, Bx \leq \beta\}.$$

The set $\tilde{\mathcal{X}}_t^I$, which we call the candidate set, is of vital importance, since our development are pinned upon this concept. To be more specific, the following claim asserts that our algorithm DOSLB must always pick elements in the $\tilde{\mathcal{X}}_t^I$ candidate sets. With Proposition 4.4.5, we say that I is *played* at time t if x_t noisily activates the BIS I at time t . Hence we draw the connection between an BIS (which is a concept for the noiseless program) and an arm that is actually played by the algorithm (under noisy observation).

Proposition 4.4.5. *The actions of DOSLB must noisily activate at least one BIS, i.e. $\forall t, \exists I_t : x_t \in \tilde{\mathcal{X}}_t^{I_t}$.*

4.5 Gaps Associated with Suboptimal BISs

Based on the definition of BIS, we claim that *if DOSLB noisily activates a suboptimal BIS at time t , then the noise scale $\rho_t(x_t; \delta)$ must be substantial.* To establish this, we introduce two key concepts for suboptimal BISs: the *feasibility gap* and the *efficiency*

gap, which respectively exploit the admissibility and optimism of x_t . Recall that in the safe multi-armed bandit setting, a suboptimal arm should suffer either a positive efficiency gap or a positive feasibility gap (or both). In this section we aim to derive a similar result. Our findings will provide a lower bound for ρ_t , determined by the *greater* of these gaps when suboptimal BISs are played.

The overall methodology primarily relies on reducing the problem to linear programming sensitivity analysis. However, we face an additional complication due to the necessity of handling general perturbations in constraint values, rather than exclusively examining the local behavior of the optimal value under minor perturbations. This complexity arises from our lack of knowledge regarding the constraints in matrices A or θ ; perturbations in this matrix, as indicated by \tilde{A} , can indeed cause the optimal solution x^* to appear suboptimal.

We first introduce some basic properties of points noisily associated with a BIS. Lemma 4.5.1 essentially describes the noise scale of points played by DOSLB must be large enough, so that suboptimal arms could be picked.

Lemma 4.5.1. *Suppose the confidence sets are consistent, and that the action of DOSLB at time t , x_t , noisily activates the BIS $I = (\mathfrak{U}, \mathfrak{R})$. Then the following relations hold true.*

$$Ax_t \leq \alpha + \rho_t \mathbf{1}, \quad Bx \leq \beta \quad (4.3)$$

$$A(\mathfrak{U})x_t \geq \alpha(\mathfrak{U}) - \rho_t \mathbf{1}, \quad B(\mathfrak{R})x_t = \beta(\mathfrak{R}), \quad (4.4)$$

$$\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \rho_t. \quad (4.5)$$

This is a straightforward inference from the noise scale lemma.

Lemma 4.5.2. *If the confidence sets are consistent, i.e., if $A \in \mathcal{C}_t(\delta)$ and $\theta \in \mathcal{C}_t^0(\delta)$, then $\forall x \in \mathcal{X}$,*

$$\forall i \in [1 : U], \max_{\tilde{a}^i \in \mathcal{C}_t^i(\delta)} |\langle \tilde{a}^i - a^i, x \rangle| \leq \rho_t(x; \delta), \quad \text{and} \quad \max_{\tilde{\theta} \in \mathcal{C}_t^0(\delta)} |\langle \tilde{\theta} - \theta, x \rangle| \leq \rho_t(x; \delta).$$

Next we proceed to the definitions of the gaps, and accompany the definitions with a quantitative example.

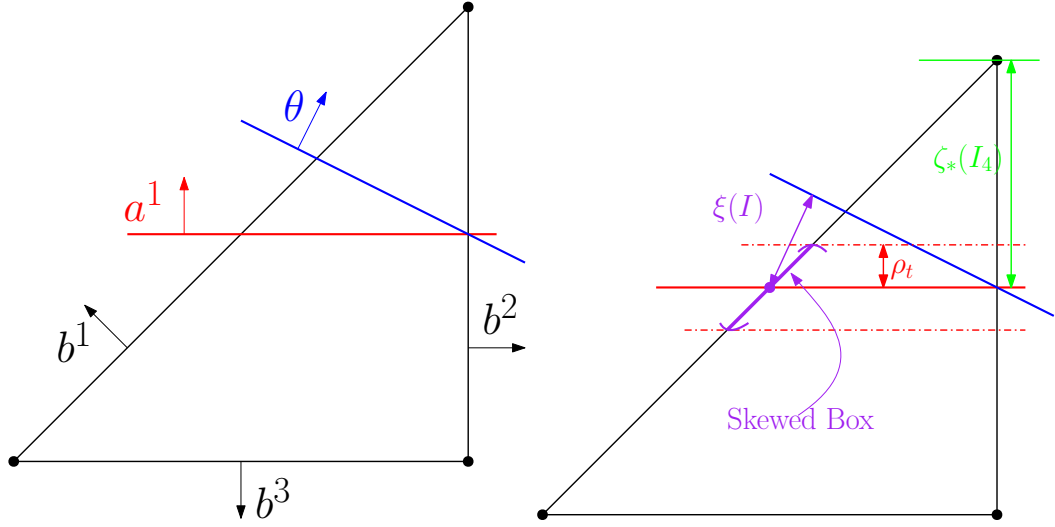


Figure 4.1: Left: Example Setup; Right: Illustration of Gaps

Example 4.5.3. To illustrate these definitions, consider the LP

$$\max x_1 + 2x_2 : \underbrace{x_2 \leq 1/2}_{\text{unknown}}, \underbrace{x_1 \geq x_2, x_1 \leq 1, x_2 \geq 0}_{\text{known}}.$$

Foregoing normalisation for clarity, we have $U = 1, K = 3$ and the parameters $\theta = (1, 2), a^1 = (0, 1), b^1 = (-1, 1), b^2 = (1, 0), b^3 = (0, -1), \alpha = (0.5), \beta = (0, 1, 0)$. There are $\binom{1+3}{2} = 6$ index sets,

$$\begin{aligned} I_1 &= (\{1\}, \{1\}), I_2 = (\{1\}, \{2\}), I_3 = (\{1\}, \{3\}), \\ I_4 &= (\emptyset, \{1, 2\}), I_5 = (\emptyset, \{1, 3\}), I_6 = (\emptyset, \{2, 3\}). \end{aligned}$$

Of these, I_2 is optimal, and the rest suboptimal, with $x^* = (1, 1/2)$. Further, I_3 is rank-deficient while the rest are full-rank. Finally, I_3 and I_4 are infeasible, while the rest are feasible.

In Example 4.5.3, we focus on BIS $I_1 = (\{1\}, \{1\})$. It is immediately observed that I_1 is feasible, full-rank and suboptimal, while the activating point is unique: $x^{I_1} = (1/2, 1/2)$. To calculate the gaps, we need to introduce several quantities. To begin with, the efficiency separation is defined as $\xi(I) := \langle \theta, x^* - x^I \rangle$. In our case, $\xi(I_1) = 1/2$.

Efficiency Gap When noisy activation of I occurs, point x_t may deviate from x^I , but not to a significant extent. As per Lemma 4.5.1, x_t must reside within a (skewed) ℓ_∞ box in close proximity to x^I , with a 'radius' of ρ_t . Specifically, for I_1 , this box can be described as $\{x : x_1 = x_2, x_2 \in 1/2 \pm \rho_t\}$. However, this imposes a constraint on how large $\langle \theta, x_t \rangle$ can become. In fact, there exists a constant $\mathfrak{s}(I)$, such that moving within a unit box of this type can only increase $\langle \theta, x \rangle$ by at most $\mathfrak{s}(I)$ —effectively, $\mathfrak{s}(I)$ measures how well the geometry of this box aligns with θ . Consequently, we can conclude that $\langle \theta, x_t \rangle \leq \langle \theta, x^I \rangle + \rho_t \mathfrak{s}(I)$. For I_1 , $\mathfrak{s}(I)$ corresponds to the inner product between $(1, 1)$ and θ , resulting in $\mathfrak{s}(I_1) = 3$. Given that $\langle \theta, x^I - x^* \rangle = -\xi(I)$, the previous discussion reveals a discrepancy with (4.5). Combining these observations, we arrive at a significant lower bound: $\rho_t \geq \eta_*(I) := \xi(I)/(1 + \mathfrak{s}(I))$, which is referred to as the *efficiency gap* of I . In the case of Example 4.5.3, $\eta_*(I_1) = 1/8$.

Now let's take a look at another BIS $I_4 = (\emptyset, \{1, 2\})$. This resembles the case in safe multi-armed bandit where an arm is unsafe.

Feasibility Gap Certainly, it is also possible for x_t to noisily activate an infeasible BIS, as demonstrated in Example 4.5.3 with $I_4 = (\emptyset, \{1, 2\})$. In such cases, a conflict arises between (4.3) and (4.4): when ρ_t is small, any point satisfying (4.3) is in close proximity to the safe set, but points that meet (4.4) are distant from safety when I is infeasible. In the context of Example 4.5.3, only $(1, 1)$ satisfies (4.4) for I_4 , whereas points that satisfy (4.3) must lie below $x_2 = 1/2 + \rho_t$. This implies the existence of a minimal perturbation $\zeta_*(I)$, known as the *feasibility gap* of I . Therefore, if an infeasible BIS is noisily activated, it must hold that $\rho_t \geq \zeta_*(I)$. In the case of Example 4.5.3, $\zeta_*(I_4) = 1/2$.

These two examples illustrate the two fundamental gaps in selecting suboptimal BISs. If the BIS is infeasible, then its activation necessitates ρ_t to be greater than or equal to its feasibility gap. Conversely, if it is feasible but suboptimal, activating the

BIS requires ρ_t to exceed its efficiency gap.

Below, we provide a unified treatment of these aspects by examining the behavior of a parameterized linear program (LP) with a feasible set determined by (4.3) and (4.4), but with ρ_t replaced by a parameter denoted as ζ . The $\zeta_*(I)$ then corresponds to a feasibility condition for this parameterized LP, $\xi(I)$ represents the minimum finite value achievable by such programs, and $\mathfrak{s}(I)$ quantifies the sensitivity of the program concerning ζ .

4.5.1 Formal Definitions of the Gaps

In this section, the text formally defines various concepts related to basic index sets (BISs) and the associated gaps. These definitions are crucial for the analysis in the next section.

Definition 4.5.4. *For a BIS I and $\zeta \geq 0$, the activation polytope of scale ζ induced by I is defined as*

$$\mathcal{T}(\zeta; I) := \{x : Ax \leq \alpha + \zeta \mathbf{1}, A(\mathfrak{U})x \geq \alpha(\mathfrak{U}) - \zeta \mathbf{1}(\mathfrak{U}), Bx \leq \beta, B(\mathfrak{K})x \geq \beta(\mathfrak{K})\}.$$

Further the optimistic LP at scale ζ induced by I is defined as $P(\zeta; I) := \sup\{\langle \theta, x \rangle : x \in \mathcal{T}(\zeta; I)\}$.

As shown in the example, the activation polytope guarantees that if x_t noisily activates I , then the action is within the activation polytope of scale ρ_t , to be more specific, $x_t \in \mathcal{T}(\rho_t; I)$, and $\langle \theta, x_t \rangle \leq P(\rho_t; I)$. Note that the optimistic LP's value $P(\zeta; I)$ is right-continuous in its scale ζ , since its feasible set is a closed polytope growing with ζ . Together with the fact that $\zeta = 0$ corresponds to the noiseless program, we conclude that $P(0; I) = -\infty$ for infeasible BIS. This naturally leads to Definition 4.5.5 that the feasibility gap represents the minimum value of the parameter ζ such that the optimistic LP is not equal to negative infinity. In other words, it captures the minimum perturbation required to make the associated constraints

feasible.

Definition 4.5.5. We define the feasibility gap of a BIS I as $\zeta_*(I) := \inf\{\zeta \geq 0 : P(\zeta; I) > -\infty\}$.

Now let's turn to the reasoning around efficiency gap. In accordance with the consistency of confidence sets, we have $x^* \in \tilde{\mathcal{S}}_t$ and $\theta \in \mathcal{C}_t^0$. Given this, and due to the optimistically-selected x_t , there exists some $\tilde{\theta} \in \mathcal{C}_t^0$ for which $\langle \tilde{\theta}, x_t \rangle \geq \langle \theta, x^* \rangle$. Additionally, Lemma 4.5.2 implies that $\langle \tilde{\theta}, x_t \rangle - \rho_t \leq \langle \theta, x_t \rangle$. However, considering that x_t belongs to $\mathcal{T}(\rho_t; I)$, if ρ_t is sufficiently small, then $\langle \theta, x_t \rangle$ cannot exceed $\langle \theta, x^* \rangle - \xi(I)$ by a significant margin. Therefore, the noise scale must be sufficiently large to render x_t seemingly optimal, yet small enough to suggest the activation of I . This delicate balance is captured in the optimistic LP

$$P(\zeta; I) := \max\{\langle \theta, x \rangle : x \in \mathcal{T}(\zeta; I)\}.$$

By definition and consistency of the confidence set, $\langle \theta, x_t \rangle \leq P(\rho_t; I)$. A critical observation here is that the growth rate of $P(\zeta; I)$ in relation to ζ is bounded. Specifically, there exists a constant, referred to as $\mathfrak{s}(I)$ and termed the *spread of I* , such that

$$P(\zeta; I) \leq P(0; I) + \zeta \mathfrak{s}(I).$$

This constraint intuitively arises because an increase in ζ leads only to a linear expansion in the activation polytope \mathcal{T} , which, in turn, can only linearly increase the value. This limitation on the growth of P provides an additional lower bound for ρ_t :

$$\langle \theta, x^* \rangle - \rho_t \leq \langle \theta, x_t \rangle \leq P(\rho_t; I) \leq \langle \theta, x^* \rangle - \xi(I) + \mathfrak{s}(I)\rho_t \implies \rho_t \geq \frac{\xi(I)}{1 + \mathfrak{s}(I)},$$

a value we refer to as the *efficiency gap* of I . This concept can be extended to infeasible BISs by controlling the growth of P for $\zeta \geq \zeta_*(I)$. These ideas are further delineated in the definitions that follow. Note that the spread term seems artificial, but it is

rooted in the sensitivity analysis.

Definition 4.5.6. *The efficiency separation of I is defined as $\xi(I) := \langle \theta, x^* \rangle - P(\zeta_*(I); I)$. The spread of I is defined as $\mathfrak{s}(I) := \inf\{C : \forall \zeta \geq \zeta_*(I), P(\zeta; I) \leq P(\zeta_*(I); I) + C(\zeta - \zeta_*(I))\}$. Accordingly, the efficiency gap of I is defined as $\eta_*(I) := \frac{\xi(I)}{1+\mathfrak{s}(I)} + \frac{\zeta_*(I)\mathfrak{s}(I)}{1+\mathfrak{s}(I)}$.*

For infeasible BISs, it's notable that $\eta_*(I) > \zeta_*(I) > 0$ occurs when $\xi(I) - \zeta_*(I) > 0$. This implies that if a minor alteration in the constraints results in $P(\zeta; I)$ becoming feasible, yet the derived solutions are significantly less effective, then the efficiency gap of I becomes more critical than its feasibility gap. For feasible BIS, i.e. $\zeta_*(I) = 0$, a similar reasoning holds, leading to a general insight: if x_t noisily activates a suboptimal BIS I , then it follows that $\rho_t \geq \eta_*(I)$. This relationship underscores the impact of the efficiency and feasibility gaps in determining the lower bounds of the noise scale ρ_t . We summarize this reasoning into Proposition 4.5.7.

Proposition 4.5.7. *For any suboptimal BIS I , the spread is finite, i.e., $\mathfrak{s}(I) < \infty$. Therefore, $\max(\zeta_*(I), \eta_*(I)) > 0$.*

Proposition 4.5.7 presents the important claim, that the above development leads to non-trivial gaps (any suboptimal BIS has a strictly positive gap). This resembles exactly the case of safe multi-armed bandit problem. Therefore, it is legitimate to define the gap of an SLB instance. This gap is strictly positive, since it is a minimum over finite number of strictly positive numbers.

Definition 4.5.8. *The gap of an SLB instance is defined as $\Xi := \min\{\max(\zeta_*(I), \eta_*(I)) : I \text{ is suboptimal}\}$.*

4.6 Regret Bounds

Before stating the main theorems, it is important to point out that the analysis hinges on the following lemma: a lower bound of the noise scale. Under Lemma 4.6.1, a suboptimal BIS can only be picked when its noise scale is above the larger one of the

corresponding efficiency and safety gaps. Theorem 4.6.2 further controls the total amount of suboptimal BIS quantitatively.

Lemma 4.6.1. *If at time t , the confidence sets are consistent and the action of DOSLB noisily activates the suboptimal BIS I , then $\rho_t \geq \max(\zeta_*(I), \eta_*(I))$.*

Theorem 4.6.2. *Let $\{x_t\}$ denote the actions of DOSLB(δ) on a safe linear bandit problem. Then, with probability at least $1 - \delta$, if at any time t , x_t noisily activates a suboptimal BIS, then $\rho_t > \Xi$. Further, the total number of times suboptimal BISs are played is bounded as*

$$\sum_t \mathbb{1}\{\exists \text{suboptimal BIS } I : x_t \in \tilde{\mathcal{X}}_t^I\} = O\left(\frac{d^2 \log^2 T + d \log(T) \log(U/\delta)}{\Xi^2}\right).$$

This outcome indicates that, for the majority of the time, DOSLB selects actions that activate the same noisy constraints as those saturated by x^* . Essentially, although the method might not pinpoint x^* with a precision finer than $O(1/\sqrt{t})$, it is adept at identifying the binding constraints. Consequently, the actions chosen by DOSLB predominantly aim at accurately activating these specific constraints. This ability to focus on relevant constraints is a significant aspect of the algorithm's performance, as it ensures that the chosen actions are aligned with the critical constraints that define the optimum solution, thereby enhancing the overall efficiency and safety scores of the algorithm.

Moving to the main result of this chapter, the preceding development established that suboptimal BISs are not frequently selected, which effectively manages a 'dual' type of regret. We now aim to extend these findings to establish bounds on the 'primal' quantities, \mathcal{E}_T (efficiency) and \mathcal{S}_T (safety). To achieve this, it is essential to consider time steps when only optimal BISs (i.e., BISs for which $x^* \in I$) are activated. The behavior during these intervals can be regulated under the following weak nondegeneracy condition at the optimum.

Assumption 4.6.3. *Every optimal BIS (i.e., $I : x^* \in \mathcal{X}^I$) is full-rank. Further, $\forall i$,*

the noise w_t^i is generic in the sense that the probability that w_t^i lies in any subspace of less than d dimensions is zero.

The role of the non-degeneracy condition as stated in Assumption 4.6.3 is relatively modest in our analysis. Essentially, it is sufficient if x_t noisily activates a set of indices such that the true parameter θ can be represented as a linear combination of the true constraint vectors corresponding to these indices. In scenarios where this condition does not hold, our proof encounters challenges. Specifically, there might be instances where certain constraints necessary to express θ are not noisily activated by x_t , even though they are activated by x^* . This discrepancy disrupts the equality of the various programs we constructed in our analysis, reducing our conclusions to mere lower bounds. These bounds would depend on the constraints active at x^* but not noisily active at x_t , as well as those that are noisily active. Moreover, it's not immediately apparent whether x_t should also optimize this modified lower bound. Nevertheless, we conjecture that this requirement is more a limitation of our proof technique than a fundamental aspect of the problem.

As stated in the previous reasoning, when suboptimal BISs are activated, it must be the case where a nontrivial gap is endured. Since this scenario is already taken care of, we focus on dealing with the case where an optimal BIS is activated. Lemma 4.6.4 guarantees that in such cases, the action x_t in our framework is designed to be effective and not excessively unsafe. An important aspect to highlight is that the parameter ε is not an input to the algorithm; rather, our results apply concurrently for any choice of $\varepsilon > 0$. This versatility in handling various values of ε equips us to examine both the efficiency and safety aspects of DOSLB 's decisions, even when a certain level of finite precision slack is incorporated in the constraint levels.

To quantify this, we define the ε -precision safety violation, denoted as $\mathcal{S}_T^\varepsilon$, which is the sum of the maximum overages beyond the safety threshold α^i and the tolerance

ε across all constraints, formally represented as:

$$\mathcal{S}_T^\varepsilon := \sum_{t \leq T} \max_i (\langle a^i, x_t \rangle - \alpha^i - \varepsilon)_+ .$$

Here, the notation $(\cdot)_+$ signifies taking the positive part of the expression, effectively measuring the extent of safety violations beyond the permissible limits.

Lemma 4.6.4. *Under assumption 4.6.3, if the confidence sets are consistent, $t \geq d+1$, and the action x_t of DOSLB(δ) is that x_t only noisily activates the optimal BIS, then $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$. Further, for any $\varepsilon > 0$, if $\rho_t(x_t) < \varepsilon$, then for every i , $\langle a^i, x_t \rangle \leq \alpha^i + \varepsilon$.*

Integrating Lemma 4.6.4 and Theorem 4.6.2, we obtain insights into how DOSLB manages these safety violations while maintaining efficiency. The lemma ensures that if x_t activates the optimal BIS, then the chosen action is at least as good as the optimum x^* in terms of efficiency. The theorem, on the other hand, restricts the frequency of suboptimal BIS activations, thereby controlling the 'dual' type of regret. This combination effectively ensures that the actions of DOSLB are both safe, within the defined ε -precision, and effective over time. This is summarized in Theorem 4.6.5 and the interpretation that follows.

Theorem 4.6.5. *Under assumption 4.6.3, $\forall \varepsilon, \delta > 0$, w.p. $\geq 1 - \delta$, the actions of DOSLB($1, \delta$) satisfy*

$$\mathcal{E}_T = O\left(\frac{d^2 \log^2(T) + d \log(T) \log(\frac{U}{\delta})}{\Xi}\right), \quad \mathcal{S}_T^\varepsilon = O\left(\frac{d^2 \log^2(T) + d \log(T) \log(\frac{U}{\delta})}{\min(\Xi, \varepsilon)}\right).$$

Safety Properties. The concept of ε -precision safety violation, denoted as $\mathcal{S}_T^\varepsilon$, is crucial for evaluating the safety performance of the algorithm over time. This measure is sublinear if, in most rounds, the actions chosen by the algorithm fall within an ε margin of the feasible set \mathcal{S} , assessed in a strict ℓ_∞ sense. Essentially, $\mathcal{S}_T^\varepsilon$ quantifies the extent of safety violations, considering an allowable precision tolerance of ε in the constraint levels α^i . This metric becomes particularly significant in practical scenarios

such as drug trials like those discussed in §1.4. Here, setting ε as a proportion of the constraint levels α^i provides a framework to ensure that any potential over-exposure to drugs is kept minimal. This is reflected in two key aspects: the overall excess violation and the count of instances where patients are exposed to doses exceeding the safety thresholds by more than the ε margin. A critical feature of this approach is its simultaneous applicability for every $\varepsilon > 0$. Notably, the algorithm does not require ε as a predetermined parameter. This flexibility allows the algorithm to adapt to the specific precision requirements of different domains or applications. For instance, in a domain where even slight deviations from the safety thresholds are critical, a smaller ε value would be implicitly considered, ensuring stringent safety compliance. Conversely, in scenarios where there is a bit more leeway in safety constraints, a larger ε could be accommodated, allowing the algorithm more freedom to optimize other performance metrics without compromising overall safety.

Finite Precision Constraint Parameters. In addition to managing precision in the constraint levels α^i , it's important to consider scenarios where the constraint parameters a^i themselves are confined to a finite precision grid. This situation is particularly relevant in drug discovery contexts, where constraints often pertain to whether a drug binds to a specific set of receptors. These so-called 'coverage' constraints are typically binary, representing the presence or absence of binding, as noted in sources like (Radhakrishnan and Tidor, 2008). In such cases, the finite precision of the constraint matrix A implies a certain granularity in the decision-making process. With DOSLB adapted to this context, it means there exists a function $\varepsilon(\pi)$, dependent on the precision π of the constraint matrix, which determines a lower bound on unsafe actions. Specifically, whenever the algorithm selects an action that violates the constraints, it does so by a margin that is at least $\varepsilon(\pi)$. This characteristic is crucial for ensuring safety. By controlling $\mathcal{J}_T^{\varepsilon(\pi)/2}$ —the cumulative measure of safety violations within a

tolerance of $\varepsilon(\pi)/2$ —we can exert direct influence over the total safety violations \mathcal{S}_T . This approach allows for a more nuanced and precise handling of safety considerations, especially in fields like drug discovery where the implications of constraint violations can be significant. It acknowledges the reality that in many practical scenarios, constraints and decision variables may not exhibit continuous variability but are instead subject to discrete levels or thresholds, necessitating algorithms that can effectively navigate within these structured environments.

Efficiency Properties. The doubly optimistic strategy, at a price of slack in the safety measure, as described by the ε -precision concept, leads to significant gains in efficiency. This is quantitatively reflected in the logarithmic control exerted on \mathcal{E}_T , the measure of efficiency. Notably, this logarithmic behavior signifies an exponential improvement compared to the \sqrt{T} efficiency regret typically associated with PO methods. This improvement has profound implications, especially in systems where a marginal degree of tolerance is acceptable. In such contexts, even though the performance is evaluated against the nominal design values of the constraints, the flexibility to slightly deviate from these strict bounds allows for more efficient operation. The algorithm’s ability to act optimistically in situations with uncertain safety parameters, therefore, yields considerable benefits in terms of operational efficiency. For practitioners, this characteristic underscores the value of adopting a slightly more flexible approach in safety-critical applications. Instead of adhering to an overly conservative strategy, which might limit the system’s efficiency, allowing for a controlled degree of risk can lead to significantly better performance outcomes. This balance between safety and efficiency is crucial, particularly in dynamic environments where decisions must be made under uncertainty and where the cost of over-conservatism can be high in terms of missed opportunities or reduced operational effectiveness.

Exact Violations. The simultaneity of Theorem 4.6.5 in ε also lets us derive control on \mathcal{S}_T . Indeed, note that $\mathcal{S}_T \leq \mathcal{S}_T^\varepsilon + \varepsilon T$. Optimising over ε in the bound of Theorem 4.6.5 yields

Corollary 4.6.6. *Assuming 4.6.3, with high probability, the actions of DOSLB(δ) satisfy*

$$\mathcal{E}_T = O(d^2 \log^2 T / \Xi), \quad \text{and} \quad \mathcal{S}_T = \tilde{O}(d\sqrt{T}).$$

Observe that this is optimal in light of Theorem 4.3.1. Thus, up to logarithmic terms, doubly-optimistic play in SLBs saturates the tradeoff inherent in this lower bound in favour of minimal efficiency regret.

Tightness of Dependence on Ξ . Exploiting a subtle reduction of safe Multi-Armed Bandits problems to SLB problems, we show that the inverse dependence on Ξ is necessary.

Theorem 4.6.7. *Fix a $c \in (0, 1)$. For any $\Xi \leq 1/16$, and any method that ensures that in every SLB instance, $\max(\mathcal{E}_T, \mathcal{S}_T) = O(T^{1-c})$, there exists an instance of the SLB problem with gap at least Ξ , such that $\liminf \frac{\max(\mathbb{E}[\mathcal{E}_T], \mathbb{E}[\mathcal{S}_T])}{\log T} \geq c/108 \cdot \Xi^{-1} \log T$.*

This result applies to DOSLB, since it achieves $\max(\mathcal{E}_T, \mathcal{S}_T) = \tilde{O}(\sqrt{T})$ in general.

Finite Action Spaces. In linear bandits, it is a common assumption that the learner is presented with a *finite set* of actions denoted as \mathcal{A} . The learner’s task is to ensure that at each time step, the chosen action x_t belongs to this set \mathcal{A} , as outlined in (Abbasi-Yadkori et al., 2011). This particular scenario greatly simplifies our analysis in various ways. One significant advantage is that there is no longer a need to resolve the optimal point, which streamlines the entire analysis process and allows us to conduct it in the primal space. In fact, we can tailor our approach by adjusting DOSLB. It should make its optimistic choice from the intersection of $\tilde{\mathcal{S}}_t$ and \mathcal{A} . To further characterize this situation, we introduce two essential concepts: the *finite-arm*

efficiency gap for any action $x \in \mathcal{A}$, denoted as Δ_x and defined as $\Delta_x := \langle \theta, x^* - x \rangle_+$, and the *finite-action safety gap* for any action $x \in \mathcal{A}$, represented as Σ_x and defined as $\Sigma_x := \max_i (\langle a^i, x \rangle - \alpha^i)_+$. Additionally, we define the overall gap of the problem as Γ , which is calculated as $\Gamma := \min_{x \in \mathcal{A}} \max(\Delta_x, \Sigma_x)$. This comprehensive description allows us to analyze and address the problem effectively.

Proposition 4.6.8. *With probability at least $1 - \delta$, the modified finite-action DOSLB achieves the following bounds in the finite-armed SLB setting: $\max(\mathcal{E}_T, \mathcal{S}_T) = O(d^2 \log^2 T / \Gamma)$.*

Chapter 5

Constrained Generalized Linear Bandits

Although constrained linear bandit problem indeed encompass the advantage of infinite action space, the linear model is somewhat restrictive. Thus we are intrigued to explore a generalized linear model for the constrained bandit problem. Constrained bandit problems within the framework of generalized linear models have been relatively understudied. Surprisingly, there has been minimal focus on this area. The existing body of research, exemplified by the work of (Amani et al., 2020), does have its shortcomings. This research introduces certain constraints by necessitating the knowledge of a matrix B that links the reward and constraint, which places restrictions on the problem's generality and increases its complexity. Furthermore, the analysis provided in this work is relatively preliminary and narrative in nature, yielding a regret rate on the reward that scales at approximately $\tilde{O}(T^{2/3})$. To address this gap in the literature, we study the Safety-constrained Generalized Linear Bandit (SGLB) problem.

5.1 Problem Setup

In this study, we focus on a constrained optimization problem defined as follows:

$$\max_x \mu(\langle \theta, x \rangle) \text{ s.t. } \nu(Ax) \leq \alpha, \nu(Bx) \leq \beta \quad (5.1)$$

This optimization problem is conducted under stochastic bandit feedback, where $\theta \in \mathbb{R}^d$ represents the unknown reward parameter, $A = [a^1, \dots, a^U]^\top$ compiles the unknown constraint parameters, $\alpha = (\alpha^1, \dots, \alpha^U)$ represents the corresponding risk

levels, μ and ν are known link functions, and B and β define the known constraints.

The game of SGLB unfolds in rounds. At each time step t , the player selects an action x_t from the bounding polytope $\mathcal{X} = \{x \in \mathbb{R}^d : \nu(Bx) \leq \beta\}$. The player then receives observations $r_t = \mu(\langle \theta, x_t \rangle) + \omega_t^0$ and $s_t^i = \nu(\langle a^i, x_t \rangle) + \omega_t^i$ for $i \in [1 : U]$, where $\omega_t^i, i \in [0 : U]$, represent observational noises.

The objective is to maximize cumulative reward while minimizing total constraint violation. In bandit literature, regret is a key performance metric. We consider two metrics: cumulative regret on reward and total constraint violation. These metrics are defined as follows:

$$\begin{aligned} \mathcal{E}_T &:= \sum_{t \leq T} (\mu(\langle \theta, x^* \rangle) - \mu(\langle \theta, x_t \rangle))_+ \\ \mathcal{S}_T &:= \sum_{t \leq T} \max_i (\nu(\langle a^i, x_t \rangle) - \alpha_i)_+ \end{aligned}$$

In the above formulas, x^* represents the solution to the noiseless optimization problem (5.1). \mathcal{E}_T aims to ensure that the player's actions are close to the best possible actions satisfying the constraints. \mathcal{S}_T prevents actions that violate constraints excessively.

It's noteworthy that in both performance metrics, only the positive parts are considered, excluding the offsetting effect of negative instantaneous regret. This makes the problem inherently more challenging and interesting, as discussed further in the later sections.

Assumptions We introduce the following assumptions, which are common and technically convenient in Generalized Linear Model (GLM) literature to ensure the existence of optimality, among other properties.

Assumption 5.1.1 (Nice link functions). *The link functions μ and ν are continuously differentiable ($\mu, \nu \in C^1$) and Lipschitz with constants k_μ and k_ν respectively, and satisfy $c_\mu = \inf \mu'(\langle \theta, x \rangle) > 0$, and $c_\nu = \min_i \inf \nu'(\langle a^i, x \rangle) > 0$.*

Assumption 5.1.2 (Boundedness). $\forall x \in \mathcal{X}$, $\|x\|_2 \leq 1$. $\|\theta\|_2 \leq 1$, $\|a^i\|_2 \leq 1$ for all $i \in [1 : U]$.

Assumption 5.1.3 (Sub-Gaussian noise). Let \mathcal{F}_t be the filtration induced by information up to time step t , i.e. $\{x_{1:t}, r_{1:t}, \{s_{1:t}^i\}_{i \in [1:U]}\}$. w_t^i are conditionally zero mean 1-sub-Gaussian noises, for all $i \in [0 : U]$ and all t , i.e.

$$\mathbb{E}[\exp(\eta w_t^i) | \mathcal{F}_{t-1}] \leq \exp(\eta^2/2)$$

for all $\eta > 0$.

Remark 5.1.4. These assumptions are standard in bandit and optimization literature. Assumption 5.1.1 guarantees the niceness and generality of the link functions, including commonly used functions like the logistic function. Assumption 5.1.2 ensures the existence of the optimization problem. Assumptions 5.1.3 restrict the noise and feedback within general yet manageable families, providing technical convenience without overly restricting the problem's scope.

5.2 The DOSGLB Algorithm

In this section, we present our algorithm for the safety-constrained generalized linear bandit problem, termed DOSGLB. This algorithm draws inspiration from the OFUL algorithm for linear bandits and adapts it to the constrained bandit problem while considering the non-linearity introduced by the link function. The core idea behind DOSGLB is to construct an estimation of the hidden parameters using historical data and then perform optimistic play based on these estimations. Before delving into the algorithm, we provide background information on the generalized linear model and the estimation of its parameters.

5.2.1 Canonical Exponential Family and Generalized Linear Model

A random variable r is said to belong to a canonical exponential family parameterized by $\theta \in \Theta$ if its density is given by:

$$f(r; \theta) = \exp(r\theta + H(r) + q(\theta))$$

Here, H and q satisfy certain regularity conditions. This canonical exponential family encompasses various commonly used distributions, including Gaussian, Gamma, and Poisson. For a random variable r from this family, its mean and variance can be expressed in terms of the function q : $\mathbb{E}[r] = -q'(\theta)$ and $\text{Var}(r) = -q''(\theta)$.

When we model the parameter as an inner product of an action x with an unknown parameter, the expectation becomes $\mathbb{E}[r|x] = -q'(\langle \theta, x \rangle)$. Denoting $\mu(\cdot) = -q'(\cdot)$ as the (inverse) link function, the relationship $r = \mu(\langle \theta, x \rangle)$ constitutes a Generalized Linear Model (GLM). Without μ , the inner product represents a standard linear model. By introducing the non-linear function μ , the model's expressive power is enhanced.

5.2.2 Maximum Likelihood Estimation of Parameters

With the preceding formulation, given the observations $(x_1, r_1), \dots, (x_{t-1}, r_{t-1})$, the log-likelihood function is expressed as:

$$\sum_{k=1}^{t-1} \log f(r_k; \theta | x_k) = \sum_{k=1}^{t-1} r_k \langle \theta, x_k \rangle + H(r_k) + q(\langle \theta, x_k \rangle)$$

Setting the derivative of this log-likelihood function to 0, we obtain the estimating equation (EE):

$$\sum_{k=1}^{t-1} (r_k - \mu(\langle \theta, x_k \rangle)) x_k = 0 \tag{5.2}$$

Here, we utilize the fact that $q'(\cdot) = -\mu(\cdot)$. Let $\hat{\theta}_t$ denote the solution of (5.2). $\hat{\theta}_t$ serves as the maximum likelihood estimator (MLE) of the parameter θ . It possesses favorable properties such as consistency and asymptotic normality under certain regularity conditions. Solving the EE can be made efficient using methods like Newton's method.

It's essential to note that the same development applies to the safety scores s_t^i . Specifically, \hat{a}_t^i represents the solution to the following EE for a^i :

$$\sum_{k=1}^{t-1} (s_k^i - \nu(\langle a^i, x_k \rangle)) x_k = 0 \quad (5.3)$$

5.2.3 Validity of the MLE

In bandit problems, a fundamental principle is optimism in the face of uncertainty. This principle has been substantiated through algorithms like the UCB algorithm for multi-armed bandits (Auer et al., 2002) and the OFUL algorithm for linear bandits (Abbasi-Yadkori et al., 2011). The essence of these algorithms lies in obtaining a reasonable estimate of the underlying parameter(s) and then adopting an optimistic strategy—overestimating the reward or underestimating the risk—to balance exploration and exploitation effectively.

In the context of SGLB problems, the initial step is to ensure the validity of the Maximum Likelihood Estimator (MLE) $\hat{\theta}_t$ concerning the true parameter. We establish the validity of this estimation through the following lemma, adapted from Lemma 3 in (Li et al., 2017).

Lemma 5.2.1. *If $\lambda_{\min}(V_\tau) \geq 1$, then for any $t \geq \tau$ and any $\delta \in (0, 1)$*

$$\mathbb{P} \left(\|\theta - \hat{\theta}_t\|_{V_{t-1}^{-1}} \leq \frac{1}{c_\mu} \sqrt{\frac{d}{2} \log(1 + \frac{2t}{d}) + \log(\frac{1}{\delta})} \right) \geq 1 - \delta$$

$$\mathbb{P} \left(\|a^i - \hat{a}_t^i\|_{V_{t-1}^{-1}} \leq \frac{1}{c_\nu} \sqrt{\frac{d}{2} \log(1 + \frac{2t}{d}) + \log(\frac{1}{\delta})} \right) \geq 1 - \delta$$

Here, $V_t = \sum_{k=1}^{t-1} x_k x_k^\top$, and $\lambda_{\min}(V)$ represents the minimum eigenvalue of a positive semi-definite matrix V . For convenience, we denote the coefficients by $\sqrt{\beta_{t-1}(\delta)} = \sqrt{\frac{d}{2} \log\left(1 + \frac{2t}{d}\right) + \log\left(\frac{U+1}{\delta}\right)}$. This lemma essentially implies that with high probability, the MLE $\hat{\theta}_t$ (or \hat{a}_t^i) is close to the true parameter θ (or a^i), within a bounded distance determined by a slowly varying term.

To interpret this lemma, we introduce confidence sets. Specifically, we define $\mathcal{C}_t^0(\delta) := \{\tilde{\theta} : \|\tilde{\theta} - \hat{\theta}_{t-1}\|_{V_{t-1}^{-1}} \leq \frac{1}{c_\mu} \sqrt{\beta_{t-1}(\delta)}\}$, $\mathcal{C}_t^i(\delta) := \{\tilde{a}^i : \|\tilde{a}^i - \hat{a}_{t-1}^i\| \leq \frac{1}{c_\nu} \sqrt{\beta_{t-1}(\delta)}\}$, and $\mathcal{C}_t(\delta) := \{\tilde{A} \in \mathbb{R}^{U \times d} : \forall i \in [1 : U], \tilde{A}_i \in \mathcal{C}_t^i(\delta)\}$. Lemma 5.2.1 guarantees that, with probability at least $1 - \delta$, the reward parameter θ belongs to $\mathcal{C}_t^0(\delta)$, and the constraint parameters A belong to $\mathcal{C}_t(\delta)$. Formally,

$$\theta \in \mathcal{C}_t^0(\delta), A \in \mathcal{C}_t(\delta) \forall i \in [1 : U] \text{ w.p. } \geq 1 - \delta \quad (5.4)$$

This assurance allows us to focus solely on cases where the confidence sets are valid, containing the true parameters, and exceptions occur with a probability no greater than δ . For analytical simplicity, we also define the noise scale as $\rho_t(x_t) = 2\sqrt{\beta_{t-1}(\delta)} \cdot \|x_t\|_{V_{t-1}^{-1}}$. In subsequent discussions, we may omit the explicit mention of δ from \mathcal{C} , \mathcal{C} , and ρ , and the variable x from ρ when context permits.

5.2.4 The DOSGLB Algorithm

The DOSGLB algorithm draws inspiration from the DOSLB algorithm presented in (Chen et al., 2023). While sharing the same fundamental approach, DOSGLB is tailored to fit the GLM framework. This method upholds optimistic confidence sets for both reward and safety parameters and jointly optimizes these parameters along with the action selection process.

Algorithm 5 Doubly-Optimistic Safe Generalized Linear Bandit (DOSLB) (δ)

Input: $\lambda > 0, \delta \in (0, 1)$
 Play actions e_1, \dots, e_d , receive $r_1, \dots, r_d, s_1^i, \dots, s_d^i, i \in [1 : U]$
for $t > d$ **do**
 Calculate $\hat{\theta}_t, \hat{a}_t^i$ by solving (5.2) and (5.3)
 Construct $\tilde{\mathcal{S}}_t(\delta)$ as in (5.5)
 Optimize (5.6) and play x_t
 Receive $r_t, s_t^i, i \in [1 : U]$
end for

Permissible Set Construction: A vital aspect of DOSGLB is the construction of the permissible set, denoted as $\tilde{\mathcal{S}}_t(\delta)$. This set encompasses all actions that could potentially adhere to the given constraints, as defined in Equation 5.5.

$$\tilde{\mathcal{S}}_t(\delta) := \{x \in \mathcal{X} : \exists \tilde{A} \in \mathcal{C}_t(\delta) \text{ s.t. } \nu(\tilde{A}x) \leq \alpha\}. \quad (5.5)$$

Optimistic Action Selection: Within this permissible set, DOSGLB adopts an optimistic approach. It selects actions by optimizing the reward function with respect to both the reward parameter and the action simultaneously. This selection process is formalized in Equation 5.6, where $(\tilde{\theta}_t, x_t)$ represents the chosen pair of parameters and action at time step t .

$$(\tilde{\theta}_t, x_t) \in \arg \max_{\tilde{\theta} \in \mathcal{C}_t^0(\delta), x \in \tilde{\mathcal{S}}_t(\delta)} \mu(\langle \tilde{\theta}, x \rangle). \quad (5.6)$$

The detailed algorithmic steps are provided in Algorithm 5, encapsulating the essence of DOSGLB's approach.

In its structure, DOSLB preserves the advantages of optimistic exploration. If the selected action x_t proves to be highly inefficient or unsafe, it indicates that the exploration along this direction has been insufficient. In other words, the chosen action hasn't been thoroughly explored yet, as evidenced by its large V_{t-1}^{-1} -norm. Consequently, the feedback obtained from this action offers information enabling

refinement of the parameter estimates.

5.2.5 Computational Efficiency

Algorithm 6 Doubly-Optimistic Safe Generalized LinUCB (DOSGLinUCB) (δ)

Input: $\delta \in (0, 1)$

Play actions e_1, \dots, e_d , receive $r_1, \dots, r_d, s_1^i, \dots, s_d^i, i \in [1 : U]$

for $t > d$ **do**

 Calculate $\hat{\theta}_t, \hat{a}_t^i$ by solving (5.2) and (5.3)

 Play the action x_t as in (5.7)

 Receive $r_t, s_t^i, i \in [1 : U]$

end for

It has long been noted that confidence region-based algorithms like OFUL (Abbasi-Yadkori et al., 2011) lack computational efficiency in (generalized) linear bandit problems, primarily due to the challenging double optimization step in (5.6). This computational burden intensifies in constrained scenarios, where the permissible set 5.5 complicates matters further. To tackle this issue, we adopt the perspective introduced by (Filippi et al., 2010; Li et al., 2017) and implement a score-based (compared to parameter-based) algorithm DOSGLinUCB. The key modification lies in the following expression:

$$\begin{aligned} x_t &= \arg \max_{x \in \mathcal{X}} \mu(\langle \hat{\theta}_t, x \rangle) + \frac{1}{c_\mu} \sqrt{\beta_{t-1}(\delta)} \|x\|_{V_{t-1}^{-1}}, \\ &\text{s.t. } \nu(\hat{A}_t x) - \frac{1}{c_\nu} \sqrt{\beta_{t-1}(\delta)} \|x\|_{V_{t-1}^{-1}} \leq \alpha \end{aligned} \tag{5.7}$$

This algorithm, described in Algorithm 6, introduces a level of relaxation compared to DOSLB by allowing a broader range of feasible actions. It employs a reward upper bound to select among potentially safe actions, making it more optimistic. Notably, the objective in (5.7) is convex, enabling straightforward optimization using techniques like interior point methods. However, it's important to note that the constraints are non-convex, as the case in (5.5). To address this, a further relaxation is applied (which

is also applied to DOSLB), replacing the V_{t-1} -norm confidence sets with $L - \infty$ or $L - 1$ norm confidence sets, formulated as follows:

$$\mathcal{C}_{t,\infty}^i := \left\{ \tilde{a} : \|(\tilde{a} - \hat{a}^i)V_t^{1/2}\|_\infty \leq \frac{1}{c_\nu} \sqrt{\beta_{t-1}(\delta)} \right\},$$

$$\mathcal{C}_{t,1}^i := \left\{ \tilde{a} : \|(\tilde{a} - \hat{a}^i)V_t^{1/2}\|_1 \leq \frac{1}{c_\nu} \sqrt{d\beta_{t-1}(\delta)} \right\},$$

By incorporating these relaxations, the algorithm achieves enhanced computational efficiency, albeit at the expense of a reduced performance guarantee.

5.2.6 Reduction to Constrained Multi-Armed Linear Contextual Bandit Algorithms

In this subsection, we elucidate the connection between our approach and that of (multi-armed) linear contextual bandit, which is also known as contextual bandit with linear payoff (Chu et al., 2011). While our main focus has been on the generalized linear bandit with a continuous action space and shared parameters, it is worth noticing the shared design philosophy across different variants of the contextual bandit problem.

In the K -armed bandit with linear payoff setting, the agent observes contexts x_i for each arm $i \in [1 : K]$ and selects the arm that maximizes the reward while adhering to the constraints. To apply our algorithm in this context, one can replace the continuous action x_t with the context corresponding to the chosen arm x_{a_t} , where a_t selects from the discrete arms $[1 : K]$. It's crucial to recognize that due to the finite number of arms, a natural gap in the problem is established, obviating the need for the gap analysis detailed in §5.3. Unlike the continuous action problem discussed here, the finite-armed problem is computationally more tractable. The optimization here is always solvable, a marked departure from the complexities involved in solving problems with continuous action spaces.

This reduction underscores the adaptability and versatility of our approach, demonstrating its seamless integration into various contextual bandit settings, whether continuous or discrete. The underlying design principles remain consistent, showcasing the applicability of the optimistic methodology across different problem formulations.

5.3 Gaps under Generalized Linear Structure

Under Assumption 5.1.1, the link functions must be strictly increasing, ensuring the existence of an inverse function. Remarkably, the noiseless (non-linear) optimization problem (5.1) can be equivalently represented as a linear program:

$$\max_x \langle \theta, x \rangle \text{ s.t. } Ax \leq \nu^{-1}(\alpha), Bx \leq \nu^{-1}(\beta) \quad (5.8)$$

When all constraints are known, under noisy observations, (5.8) reduces to a standard linear bandit problem. Due to the polytopal structure of the domain, there exists a gap between different actions. Basic linear programming theory ensures that only the extreme points of the polytope could serve as candidates for the optimal solution. Therefore, calculating the gap of the problem involves comparing these extreme points, as demonstrated in prior works (Dani et al., 2008; Abbasi-Yadkori et al., 2011). Specifically, when the candidate actions are finite, it is well-established that logarithmic rates can be achieved, as evidenced in (Auer et al., 2002) for the unconstrained case and (Chen et al., 2022) for the constrained case.

However, in scenarios where the boundaries are blurred by noise, the finite-action property of the constrained linear bandit problem vanishes. The inherent challenge of this problem stems from the difficulty of estimation. Determining the exact location of the optimal solution becomes inherently difficult, constrained by the estimation complexity. This intricate relationship has been meticulously explored in (Chen et al., 2023), revealing the impossibility of achieving faster rates than \sqrt{T} on both the reward

and safety regret. Despite this, a form of discreteness can still be distilled from this continuous problem through the concept of basic index sets. Although a more rigorous definition can be found in (Chen et al., 2023), the basic premise is outlined here.

To grasp this concept, it's necessary to understand some fundamental principles in linear programming. A linear program (LP) can always be expressed in the form:

$$\max_{x \in \mathbb{R}^d} \langle \theta, x \rangle \text{ s.t. } Ax \leq \alpha \quad (5.9)$$

In this context, the i th constraint $\langle a^i, x \rangle \leq \alpha^i$ is deemed binding or active at x^0 (alternatively, that x^0 activates the i th constraint) if $\langle a^i, x^0 \rangle = \alpha^i$. A point x^0 in $\{x \in \mathbb{R}^d : Ax \leq \alpha\}$ is known as a Basic Feasible Solution (BFS) if it activates d linearly independent constraints. When the domain $\{x \in \mathbb{R}^d : Ax \leq \alpha\}$ forms a bounded polytope, such as a hypercube, all the basic feasible solutions are vertices/extreme points, and all the vertices/extreme points are basic feasible solutions. These terms are used interchangeably.

The optimal feasible solution of the LP (5.9), if it exists, must be attained at one of the basic feasible solutions (Bertsimas and Tsitsiklis, 1997). In cases where the domain is a bounded polytope, all basic feasible solutions are vertices, and so the linear bandit problem reduces to a multi-armed bandit problem (Auer et al., 2002). Consequently, logarithmic regret can be achieved due to the finite number of vertices, representing the discrete nature of the problem.

However, when unknown constraints are introduced, the situation becomes more intricate. The primary challenge arises from unknown constraints potentially intersecting at unknown extreme points, which cannot be determined solely by the known constraints. Consequently, the reduction to a multi-armed bandit no longer applies. Fortunately, this complexity does not signify a dead end. A critical observation from (Chen et al., 2023) illuminates that despite the constraints floating around the true

values, the algorithm's selected action point must noisily activate d constraints. Based on this insight, it is concluded that the room for the noise scale cannot be excessively large, as established through a dual sensitivity analysis (see §5.4). This room for the noise scale is encapsulated by the concept of various gaps, formalized as follows.

Similar to the role of BFS in the noiseless program, a concept representing the problem's discreteness is introduced: the basic index set (BIS).

Definition 5.3.1. *A basic index set (BIS) is an ordered pair of sets $(\mathfrak{U}, \mathfrak{K})$ where $\mathfrak{U} \subseteq [1 : U]$, $\mathfrak{K} \subseteq [1 : K]$, and $|\mathfrak{U}| + |\mathfrak{K}| = d$.*

In the context of the constrained generalized linear bandit problem, each BIS gives rise to a system of linear equations, the solutions of which activate the corresponding BIS. This concept is akin to the notion of BFS in certain scenarios.

Definition 5.3.2. *The set of points that activate an index set $I = (\mathfrak{U}, \mathfrak{K})$ is defined as*

$$\mathcal{X}^I := \{x \in \mathcal{X} : \nu(A(\mathfrak{U})x) = \alpha(\mathfrak{U}), \nu(B(\mathfrak{K})x) = \beta(\mathfrak{K})\}.$$

With these activating points identified, BIS can be categorized in the following ways, which are instrumental when discussing different gaps.

Definition 5.3.3. *A BIS I is called (i) feasible if $\mathcal{X}^I \neq \emptyset$ and infeasible otherwise; (ii) suboptimal if $x^* \notin \mathcal{X}^I$ and optimal otherwise; (iii) full rank if the vectors $\{a^i\}_{i \in \mathfrak{U}} \cup \{b^j\}_{j \in \mathfrak{K}}$ span \mathbb{R}^d .*

So far, we have been dealing with the noiseless program, where the development closely aligns with standard LP definitions. Next, we introduce a unique element in our gap analysis: the concept of noisily activating points, which arises solely in the stochastic bandit setting due to observation noises.

Recall from §5.2.3 that, with high probability, the MLEs are close to the true parameters, a fact summarized by the confidence sets introduced in the same section. We consider the scenario where all the confidence sets are valid from this point onwards.

In the presence of noisy observations, BISs could intersect at points different from the true ones. However, the concentration results mentioned above (confidence sets) constrain the variation room for these intersection points due to the noise. To capture this maximum variation room for the extreme points, we introduce the definition of noisy activating points.

Definition 5.3.4. *The set of points that noisily activates an index set $I = (\mathfrak{U}, \mathfrak{K})$ at time t is*

$$\begin{aligned} \tilde{\mathcal{X}}_t^I := \{x \in \mathcal{X} : \exists \tilde{A} \in \mathbf{C}_t, \nu(\tilde{A}(\mathfrak{U})x) = \alpha(\mathfrak{U}), \\ \nu(B(\mathfrak{K})x) = \beta(\mathfrak{K}), \nu(\tilde{A}x) \leq \alpha\} \end{aligned}$$

Here, $M(\mathcal{L})$ stacks the rows of matrix M indexed in \mathcal{L} to form a submatrix.

Thanks to the monotonicity and smoothness of the link function, we can adapt Proposition 8 in (Chen et al., 2023) to our generalized linear case. This structural result enables analysis centered on BISs. Additionally, we can derive the range of variation for the actions selected by our algorithm, as stated in Lemma 5.3.6.

Proposition 5.3.5. $\forall t, \exists$ BIS $I_t = (\mathfrak{U}_t, \mathfrak{K}_t) : x_t \in \tilde{\mathcal{X}}_t^{I_t}$.

Lemma 5.3.6. *Assume the event in (5.4) holds, and the action of DOSLB at time t , x_t , noisily activates the BIS $I = (\mathfrak{U}, \mathfrak{K})$. Then the following holds.*

$$\begin{aligned} Ax_t &\leq \nu^{-1}(\alpha) + \frac{\rho_t}{c_\nu} \mathbf{1}, & Bx &\leq \nu^{-1}(\beta) \\ A(\mathfrak{U})x_t &\geq \nu^{-1}(\alpha(\mathfrak{U})) - \frac{\rho_t}{c_\nu} \mathbf{1}, & B(\mathfrak{K})x_t &= \nu^{-1}(\beta(\mathfrak{K})), \\ \langle \theta, x_t \rangle &\geq \langle \theta, x^* \rangle - \frac{\rho_t}{c_\mu}. \end{aligned}$$

Proposition 5.3.5 ensures that our focus can be narrowed down to BISs. These BISs allow us to establish specific gaps for each of them. There are two types of gaps associated with a BIS: the efficacy gap and the feasibility gap. This observation aligns with findings in multi-armed and linear cases (Chen et al., 2022; Chen et al., 2023). The efficacy gap gauges the distance between a BIS and the optimal BIS. On the

other hand, the feasibility gap measures the margin a BIS needs to remain feasible. If the associated points of a BIS are outside the feasible set, its feasibility gap is positive. Conversely, if the associated points are within the feasible set, its efficacy gap is non-negative. In the stochastic bandit observation setting, it's necessary to consider the intersection points influenced by noise, as demonstrated in Lemma 5.3.6. This necessity leads to the introduction of the concept of an activation polytope.

Definition 5.3.7. *For a BIS I and $\zeta \geq 0$, the activation polytope at scale ζ induced by I is defined as*

$$\mathcal{T}(\zeta; I) := \{x \in \mathcal{X} : Ax \leq \nu^{-1}(\alpha) + \frac{\zeta}{c_\nu} \mathbf{1}, A(\mathfrak{L})x \geq \nu^{-1}(\alpha(\mathfrak{L})) - \frac{\zeta}{c_\nu} \mathbf{1}(\mathfrak{L}), B(\mathfrak{R})x \geq \nu^{-1}(\beta(\mathfrak{R}))\}$$

Further the optimistic LP at scale ζ induced by I is defined as $P(\zeta; I) := \sup\{c_\mu \langle \theta, x \rangle : x \in \mathcal{T}(\zeta; I)\}$.

The activation polytope and corresponding optimistic LP delineate the room for extreme points to vary due to noise. With these concepts in place, we can introduce the feasibility and efficacy gap. The feasibility gap captures the minimum scale for an activation polytope to be non-empty. Following convention, we assign the value of an infeasible program as $-\infty$.

Definition 5.3.8. *We define the feasibility gap of a BIS I as $\zeta_*(I) := \inf\{\zeta \geq 0 : P(\zeta; I) > -\infty\}$.*

Under the noisy activation of a BIS I , the actual point x_t could deviate from the associated point x^I , the extent of which is captured by the activation polytope. Within the activation polytope, moving the activation point around could lead to an increase in the objective value, characterized by the (efficacy) spread. We formally define the efficacy separation, spread, and efficacy gap as follows.

Definition 5.3.9. *The efficacy separation of I is defined as $\xi(I) := c_\mu \langle \theta, x^* \rangle -$*

$P(\zeta_*(I); I)$. We further define the spread of I as $\mathfrak{s}(I) := \inf\{C : \forall \zeta \geq \zeta_*(I), P(\zeta; I) \leq P(\zeta_*(I); I) + C(\zeta - \zeta_*(I))\}$.

Definition 5.3.10. *The efficacy gap of a BIS I is defined as $\eta_*(I) := \frac{\xi(I)}{1+\mathfrak{s}(I)} + \frac{\zeta_*(I)\mathfrak{s}(I)}{1+\mathfrak{s}(I)}$.*

As observed in the multi-armed version of the problem (Chen et al., 2022), each suboptimal BIS corresponds to both an efficacy gap and a feasibility gap, at least one of which must be strictly positive. This key observation is summarized in the following proposition.

Proposition 5.3.11. *For any suboptimal BIS I , $\max(\zeta_*(I), \eta_*(I)) > 0$.*

Therefore, we can define the gap of the problem, which represents the room needed to distinguish the closest suboptimal BIS from the optimal one.

Definition 5.3.12. *The gap of an SGLB instance is $\Xi := \min\{\max(\zeta_*(I), \eta_*(I)) : I \text{ is suboptimal}\}$.*

5.4 Regret Analysis

We now discuss main theoretical claims. Firstly, we conclude that suboptimal BISs cannot be selected too frequently, which we achieve through a dual sensitivity analysis. Subsequently, we convert these findings into bounds for the the regret quantities, specifically \mathcal{E}_T and \mathcal{S}_T . To do this, we must consider the time steps when only optimal BISs are chosen (i.e., I such that $x^* \in I$). We can manage the occurrences of such selections under a weak non-degeneracy condition at the optimum, which is commonly assumed in optimization literature.

Assumption 5.4.1. *Every optimal BIS is full-rank. Further, $\forall i$, the noise w_t^i is generic in the sense that the probability that w_t^i lies in any subspace of less than d dimensions is zero.*

Assumption 5.4.1 essentially demands that the underlying linear program and the noise adhere to normalcy. In practical terms, this means that an optimal BIS must

exclusively contain the optimal action. Translated into linear programming terminology, this implies that the optimal feasible solution is non-degenerate. Additionally, it necessitates that the noise be generic, indicating that it possesses some energy across all dimensions almost surely. Non-degeneracy serves to circumvent complex discussions involving corner cases caused by vacuous basic feasible solutions and is primarily for technical convenience, aligning with practices in the linear programming literature, such as (Bertsimas and Tsitsiklis, 1997). Generic noise ensures that when probing an action, the agent has observations from all dimensions almost surely, preventing situations where there is no signal for specific dimensions with a fixed probability. In such cases, exploring these dimensions with sublinear regret would be impossible. These assumptions, though weak, lay the foundation for our main technical result, controlling the noise level using the gaps defined in the previous section. Essentially, Lemma 5.4.2 establishes that whenever a suboptimal BIS is activated, the noise scale must be sufficiently large to escape its respective gaps. Consequently, due to concentration results, this scenario occurs rarely.

Lemma 5.4.2. *If at time t , the confidence sets are consistent and the action of DOSLB noisily activates the suboptimal BIS I , then $\rho_t \geq \max(\zeta_*(I), \eta_*(I))$.*

Having controlled the time steps when suboptimal BISs are (noisily) activated, we also provide a result regarding the behavior of optimal BISs. Lemma 5.4.3 states that whenever an optimal BIS is (noisily) activated, the corresponding action must be overly efficient, and safety violations are controllable within any small tolerance margin ε .

Lemma 5.4.3. *Under assumption 5.4.1, if the confidence sets are consistent, $t \geq d + 1$, and the action x_t of DOSLB(δ) is that x_t only noisily activates the optimal BIS, then $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$. Further, for any $\varepsilon > 0$, if $\rho_t(x_t) < \varepsilon$, then for every i , $\langle a^i, x_t \rangle \leq \alpha^i + \frac{k_\nu}{c_\nu} \varepsilon$.*

By combining Lemma 5.4.2 and Lemma 5.4.3, we obtain the primary regret bound,

as outlined in Theorem 5.4.4.

Theorem 5.4.4. *Under assumption 5.4.1, $\forall \delta > 0$, the actions of DOSLB(δ) satisfy with high probability*

$$\mathcal{E}_T = O(d^2 \log^2 T / \Xi), \quad \text{and} \quad \mathcal{S}_T = \tilde{O}(d\sqrt{T}).$$

To demonstrate the *tightness of the dependence on Ξ* , we refer to Theorem 23 from (Chen et al., 2023), which provides a lower bound in the linear case. Since the linear model is a special case of GLM, where in our notation $k_\mu = c_\mu = k_\nu = c_\nu = 1$, this theorem also serves as a lower bound for our setting.

Theorem 5.4.5. *For any $c \in (0, 1)$, for any Ξ small enough, and any method that ensures that in every SLB instance, $\max(\mathcal{E}_T, \mathcal{S}_T) = O(T^{1-c})$, there exists an instance of the SLB problem with gap $\geq \Xi$, such that*

$$\liminf \frac{\max(\mathbb{E}[\mathcal{E}_T], \mathbb{E}[\mathcal{S}_T])}{\log T} = \Omega(\Xi^{-1} \log T).$$

Indeed, Theorem 5.4.4 stands as optimal when viewed alongside Theorem 5.4.5. Consequently, accounting for logarithmic terms, the strategy of doubly-optimistic play in SGLB maximizes the tradeoff outlined in this lower bound, favoring minimal efficacy regret.

Chapter 6

Simulations

Although this dissertation focuses mainly on the theoretical aspects of constrained bandit problems, we briefly present several informative simulation results to complement the claims made in previous chapters. Since the performance of the doubly optimistic schemes among different settings are quite similar, we take the DOSLB in the SLB setting as a representative example to illustrate the empirical performance of our algorithms.

6.1 Simulations on SLB

We verify the theoretical study above with simulations over Example 4.5.3, and study the relative performance of DOSLB and the optimistic-pessimistic method Safe-LTS (Moradipari et al., 2021). These implementations are based on the following relaxation of Algorithm 4.

6.1.1 Computationally Feasible Relaxation

A well-known barrier to implementing Algorithm 4 is that even if all constraints were known, the program (4.2) is non-convex (Dani et al., 2008). In our case, this is further complicated by the fact that the set $\tilde{\mathcal{S}}_t$ needs to be determined, which too is computationally subtle.

We approach these issues by constructing *box confidence sets*. We consider two

relaxations to this, the L_∞ box and the L_1 box, as follows:

$$\mathcal{C}_{t,\infty}^i := \{\tilde{a} : \|(\tilde{a} - \hat{a}^i)V_t^{1/2}\|_\infty \leq \sqrt{\beta_t}\},$$

$$\mathcal{C}_{t,1}^i := \{\tilde{a} : \|(\tilde{a} - \hat{a}^i)V_t^{1/2}\|_1 \leq \sqrt{d\beta_t}\},$$

such that $\mathcal{C}_t^i \subset \mathcal{C}_{t,\infty}^i$ and $\mathcal{C}_t^i \subset \mathcal{C}_{t,1}^i$, since L_2 ball with radius r is contained in L_1 ball with radius r and L_∞ ball with radius \sqrt{dr} . Take any $\tilde{a} \in \mathcal{C}_{t,\infty}^i$, $\|(\tilde{a} - \hat{a}^i)V_{t-1}^{1/2}\|_\infty \leq \sqrt{\beta_{t-1}} \implies \|(\tilde{a} - \hat{a}^i)V_{t-1}^{1/2}\|_2 \leq \sqrt{d\beta_{t-1}}$, and the same holds for the L_1 relaxation as well. Thus replacing \mathcal{C}_t^i by $\mathcal{C}_{t,\infty}^i$ or $\mathcal{C}_{t,1}^i$, the only change in analysis will be from ρ_t to $\bar{\rho}_t = \sqrt{d}\rho_t$. Hence running DOSLB with these worsens regret bounds by at most $O(\sqrt{d})$ (and the relaxed regret bound by at most $O(d)$).

The principal advantage of using $\mathcal{C}_{t,\infty}$ lies in the fact that the box-confidence sets are polytopes. Due to this, the values of $\tilde{\theta} \in \mathcal{C}_{t-1,\infty}^0$ and $\tilde{a}^i \in \mathcal{C}_{t-1,\infty}^i$ that are active for the optimistic action x_t must lie at the extreme points of these sets. Since each set has only $2d$ extreme points, this allows us to determine x_t by solving $(2d)^{U+1}$ convex programs, which is computationally feasible so long as U is small.

To investigate the choice between these two relaxations, we simulated DOSLB with each of these on the instance of Example 4.5.3. We find that while both show very strong efficacy and acceptable safety violations, the L_1 relaxation appears to be more aggressive, and thus has weaker safety properties. See §6.1.2 for the observations.

6.1.2 Results

We implement DOSLB on with the L_∞ relaxation above on the instance of Example 4.5.3 over the horizon $T = 10^4$, and with the parameters $\lambda = 2, \delta = 1/(4T) = 2.5 \times 10^{-5}$. The noise in observations is independent and Gaussian, with variance 0.1. The key observations of our study are as follows. Notice that for this instance, $\Xi = 1/8$.

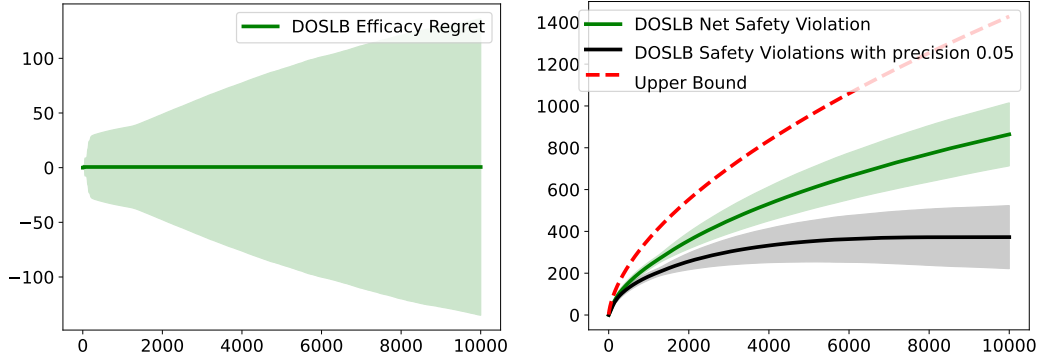


Figure 6.1: Efficacy Regret and Safety Violation of DOSLB . We plot averages and one standard deviation confidence regions over 30 runs for \mathcal{E}_T (left) and both \mathcal{S}_t and $\mathcal{S}_t^{0.05}$ (right). We also plot the upper bounds we show in the latter to contextualise the observations. Observe that the efficacy regret is marginal: the mean is essentially 0, and the variance limited. Further, observe that the growth of the net safety violation \mathcal{S}_t is well-controlled, and lies far below the bounds of §4.6. Further, the finite precision violations show a strong flattening, as is expected from Theorem 4.6.5.

DOSLB is very effective, and has well-controlled violations. Figure 6.1 shows the efficacy regret \mathcal{E}_t and both the arbitrary precision safety violations \mathcal{S}_t and the finite precision safety violations $\mathcal{S}_t^\varepsilon$ for the value $\varepsilon = 0.05 = 2\Xi/5$. The observations strongly validate our main claims of strong efficacy regret control, and well-behaved growth of safety violations. Indeed, observe that the efficacy regret is essentially zero over most of the runs (with rare runs rising to $\mathcal{E}_{10^4} \approx 100$). This property arises since DOSLB very rarely plays suboptimal BISs (see the following discussion and Figure 6.2), and when it plays the optimal BIS, it plays a ‘over-efficient’ but unsafe point. Further, the extent of the lack of safety of the actions chosen by DOSLB is well-controlled, as seen in the behaviour of \mathcal{S}_T . The finite precision regret shows even stronger control, with growth essentially halted at $t \approx 5000$, validating the analysis underlying Theorem 4.6.5.

DOSLB rarely activates suboptimal index sets. In Figure 6·2, we plot the number of times that DOSLB noisily activates a suboptimal BIS, i.e., any index set other than $I_2 = (\{1\}, \{2\})$. The main observation is that this occurs very rarely: indeed, over the horizon of 10^4 , most runs do not activate suboptimal BISs more than 100 times. This is far below the upper bound of Theorem 4.6.2. We also observe the curious fact that for $t < 1500$, such index sets are essentially never activated.

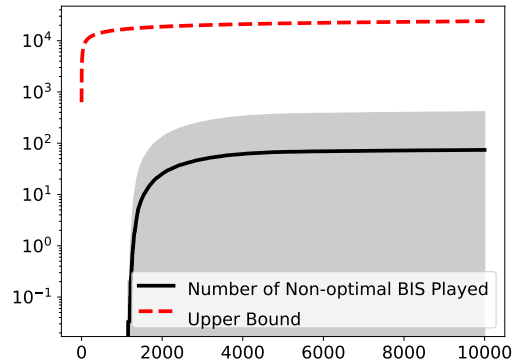


Figure 6·2: Number of times a suboptimal BIS is noisily activated by DOSLB in the instance of Example 4.5.3. Means over and one standard deviation over 30 runs are reported, and the vertical scale is logarithmic. Observe that the method activates such index sets very rarely, typically far less than 1% of the times. Also observe that the growth is essentially flat.

DOSLB compares favourably with pessimistic-optimistic methods. To contextualise our method, we also implement the PO-method *safe-LTS* due to (Moradipari et al., 2021) in the instance of Example 4.5.3. As previously discussed, *safe-LTS* constructs a pessimistic set of permissible points, Π_t , such that with high probability all points in Π_t must be safe. The method then selects actions optimistically, in this case by exploiting Thompson sampling. The safe point provided to *safe-LTS* is $x^s = (0, 0)$, which has the separation $M^s = 1/2$.

Figure 6·3 compares the behaviour of the *raw efficacy regret* $\sum \langle \theta, x^* - x_t \rangle$ (left) and the *raw safety violation* $\sum \max_i (\langle a^i, x_t \rangle - \alpha^i)$ (right) of *Safe-LTS* and *DOSLB*

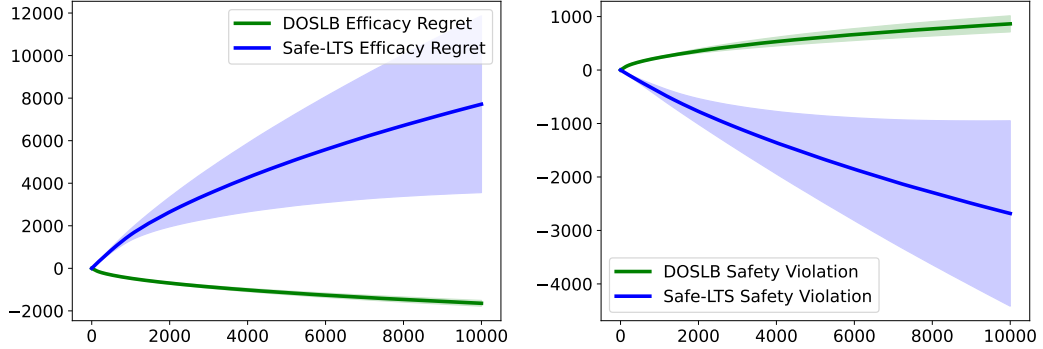


Figure 6.3: Comparing the behaviour of DOSLB and safe-LTS on the instance of Example 4.5.3. The left plot shows the *raw* efficacy regret, while the right plot is the *raw* safety violations of the two methods, and each reports means and one-standard deviation confidence regions over 30 runs.. Observe that the efficacy performance of safe-LTS is extremely poor, indicating that the algorithm is far from the boundary of the safe set \mathcal{S} for most of its runs. In contrast, the violation properties of DOSLB are well-controlled, and almost four times smaller than the efficacy regret of safe-LTS.

(since the efficacy regret of DOSLB , and the safety-violations of safe-LTS are both essentially 0, the raw behaviour elucidates more insight). As expected, safe-LTS suffers from 0 safety regret, since it plays in a pessimistic set. However, this is accompanied by a large efficacy regret, with the mean of over 7000 at the horizon $T = 10^4$. This arises due to the extreme conservatism of this method, which is evident from its safety violation property: the method has a strong negative (and decreasing still) violation, indicating that it continues to play deep in the interior of the domain for large T . Indeed, since over the domain, $\langle a, x \rangle - \alpha \in [-0.5, 0.5]$, and since the violation at $T = 10^4$ is roughly -3000 , this indicates that with a nontrivial probability, the method remains at least 0.25-separated from the boundary of the safe set.

In comparison, observe that the raw efficacy regret of DOSLB is negative, but not nearly as far as the violations of safe-LTS. This indicates that the method is shrinking towards the boundary of the safe set at a much better rate. Of course, this property is similarly illustrated by the violation behaviour: this nearly four times smaller than

the efficacy regret of safe-LTS, and concentrates strongly to $\approx 10^3$ at $T = 10^4$.

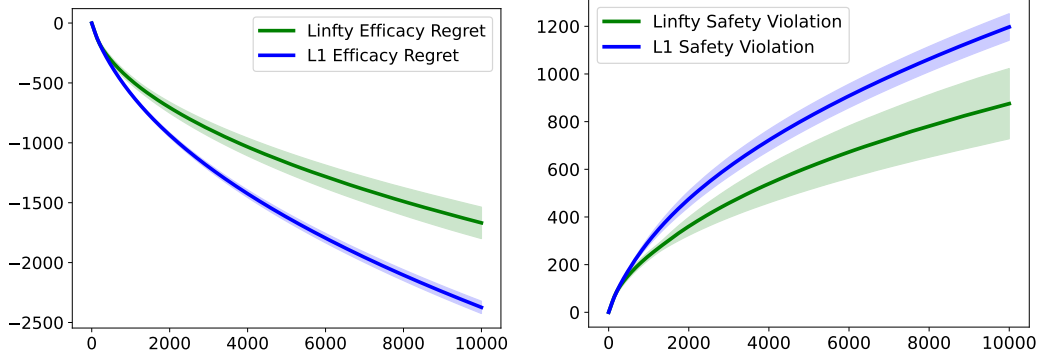


Figure 6.4: Comparison of L_∞ vs L_1 relaxations of DOSLB over 30 runs on the instance of Example 4.5.3. Observe that the L_1 relaxation yields more aggressive, and less safe play.

L_∞ versus L_1 relaxation. Finally, on a technical note, we simulate DOSLB with both the L_∞ and L_1 relaxations discussed in the previous section. Again, since each achieves 0 efficacy regret, we plot their raw efficacy and safety performances in Figure 6.4.

As we noted earlier, both relaxations are largely comparable, showing similar performance on both metrics. In particular, the qualitative observations made earlier remain the same. The difference between the methods is that the L_1 -relaxed DOSLB is somewhat more aggressive than the L_∞ -relaxed DOSLB, and thus suffers a lower raw efficacy regret and a higher safety violations. One reason for this may be that the L_1 -DOSLB has a larger magnitude in terms of the V_t -norm.

Chapter 7

Conclusion and Future Work

In this dissertation, we study the constrained bandit problems, in terms of multi-armed, linear, and generalized linear settings. We explore the usage of doubly optimistic approach in such scenario, and end up with meaningful findings. The doubly optimistic strategy encourages aggressive exploration, and thus enjoys fast rate in reward accumulation. For the safety risk, in terms of multi-armed setting, due to the discreteness of the action space, the doubly optimistic strategy is able to filter out the unsafe arms as fast as accumulating rewards, thus the safety violation is also at a logarithmic rate; for the continuous settings, we investigate the linear structure, and show that at least \sqrt{T} rate is achievable under no additional assumptions, which is already the best to expect in the provided setting, while in addition, we show that an arbitrary slackness in the safety measurement could lead to a logarithmic rate similar to the discrete case. These findings contextualize the potential usage of doubly optimistic strategies in applications.

Moving forward, the exploration of contextual bandit and reinforcement learning settings with the inclusion of states remains open for investigation. In these scenarios, the challenge lies in devising effective strategies under various constraints, while guaranteeing fast convergence rates. Doubly optimistic strategies have shown promise in addressing such challenges, yet there is still much ground to cover in fully understanding their potential.

One area of future research lies in the extension of doubly optimistic strategies

beyond the realm of bandit settings into broader machine learning contexts. For instance, there is a compelling avenue to explore how these constrained methods could enhance the efficiency of training neural networks. By integrating constrained optimization techniques, we may unlock novel approaches for accelerating neural network convergence and improving generalization performance. This could have implications across various domains, from image recognition to natural language processing.

Furthermore, the proposed doubly optimistic techniques could find its usage in real-world applications with complex feedback model. As an example, the dynamic and complex nature of wireless radio access networks(Olwal et al., 2016) presents various challenges, including resource allocation(Liang et al., 2019; Wei et al., 2020) and interference mitigation(Wei et al., 2021; Wei et al., 2019). By leveraging constrained optimization methods inspired by doubly optimistic strategies, we could potentially optimize those operations in real-time, leading to improved network performance and user experience. In traffic systems, the integration of constrained optimization techniques holds promise for alleviating congestion and enhancing traffic flow (Salkham et al., 2008; Arel et al., 2010). By applying doubly optimistic strategies to traffic management algorithms, we may develop adaptive systems capable of dynamically adjusting traffic signals and routing based on real-time constraints. This could result in reduced travel times, minimized environmental impact, and enhanced overall efficiency of transportation networks. Additionally, the adoption of constrained optimization methods inspired by doubly optimistic strategies could also be applicable to distributed machine learning systems. In scenarios where data is distributed across multiple nodes or devices, such as in edge computing environments, efficient coordination and optimization are of significant interests (Verbraeken et al., 2020; Wei et al., 2023a; Wei et al., 2023b). By applying constrained optimization techniques, we may develop

distributed learning algorithms that not only converge faster but also exhibit improved robustness and scalability ([Wei and Shen, 2022](#); [Boyd et al., 2011](#)).

In summary, the exploration of doubly optimistic (and in general, constrained learning) methods represents a rich area for future research spanning various domains. By pushing the boundaries of these methodologies, we can uncover new possibilities for optimization and learning techniques.

Appendix A

Supplement for § 3

A.1 Notation and General Proof Strategy

We commence by introducing some notations, followed by an overview of the general proof strategy.

We adopt the convention of using \mathcal{H}_{t-1} to denote both the history and the sigma algebra induced by it. Here, \mathcal{H}_0 corresponds to the trivial sigma algebra. Naturally, the sequence $\{\mathcal{H}_t\}$ forms a filtration. It is noteworthy that in the cases of Thompson sampling (TS), the laws of ρ_t^k are measurable with respect to \mathcal{H}_{t-1} . Bayesian methods also incorporate additional randomness, represented by the various ρ_t^k s.

A crucial observation for all our designed methods is that the permissible set Π_t is a *predictable* process, meaning it is determined based on \mathcal{H}_{t-1} . Each method utilizes an index derived from the history to determine Π_t , making it a deterministic function of the variables $\{(A_s, R_s, S_s) : s < t\}$. While not strictly necessary, this predictability serves as a convenient representation that we leverage in our proofs. For brevity, we denote the conditional laws $\mathbb{P}(\cdot|\mathcal{H}_{t-1})$ as \mathbb{P}_{t-1} .

Proof Strategy: The foundational breakdown of regret is expressed in terms of N_{T+1}^k . Due to the additive definition, we have

$$\mathbb{E}[\mathcal{R}_T] = \sum_{k \neq k^*} \mathbb{E}[N_{T+1}^k](\Delta^k \vee \Gamma^k).$$

Thus, the primary arguments focus on controlling $\mathbb{E}[N_{T+1}^k]$ for all suboptimal arms k .

Naturally, subsidiary claims regarding $\mathbb{E}[\mathcal{U}_T]$ also stem from these main arguments.

The arguments manage $\mathbb{E}[N_{T+1}^k]$ for both infeasible and inefficient arms separately, through a straightforward splitting of terms. In cases where arms exhibit both inefficiency and infeasibility, the more stringent control between these two arguments is applied. This approach defines the structure of the expressions in the main text.

Infeasible Arms: All our strategies incorporate a safety index L_t^k to populate the permissible set Π_t . We leverage the properties of this index to regulate the play of infeasible arms. This decomposition is expressed as

$$\mathbb{E}[N_{T+1}^k] = \sum_{t=1}^T \mathbb{P}(A_t = k) \leq \sum_{t=1}^T \mathbb{P}(L_t^k \leq \alpha).$$

The formulation of the two indices, whether through KL-UCB or BAYESUCB, guarantees that the likelihood of playing an infeasible arm more than $O(\log(T)/d(\nu^k \|\alpha))$ times is exceedingly small. For KL-UCB, this outcome directly follows from Chernoff's bound. In the case of BAYESUCB, the argument is derived from the KL-UCB approach, drawing a connection between the tails of Beta distributions and Binomials.

Inefficient Arms: Employing the conventional approach for confidence-bound-based index policies, the control of inefficient arms necessitates a reference index for comparison with the reward indices. Ideally, the index of k^* is desired. However, utilizing this index requires that k^* itself is permissible, as otherwise, the algorithm disregards its reward index when selecting an arm. This represents a key deviation from standard proofs.

Consider the case of KL-UCB. The strategy involves decomposing the expression as follows:

$$\begin{aligned}
\mathbb{E}[N_{T+1}^k] &= \sum_t \mathbb{P}(A_t = k) \\
&= \sum_t \mathbb{P}(A_t = k, k^* \notin \Pi_t) + \mathbb{P}(A_t = k, k^* \in \Pi_t) \\
&\leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \sum_t \mathbb{P}(A_t = k, k^* \in \Pi_t)
\end{aligned}$$

The initial step is to ensure that the first term is small, leveraging the consistency of L_t^* .

This allows us to proceed in a manner consistent with standard approaches. Specifically, for KL-UCB, we break down the second term as follows:

$$\begin{aligned}
\sum_t \mathbb{P}(A_t = k, k^* \in \Pi_t) &\leq \sum_t \mathbb{P}(U_t^* < \mu^*, k^* \in \Pi_t) \\
&\quad + \mathbb{P}(U_t^k \geq \mu^*, U_t^* \geq \mu^*, k^* \in \Pi_t, A_t = k) \\
&\leq \sum_t \mathbb{P}(U_t^* < \mu^*) + \mathbb{P}(U_t^k \geq \mu^*, A_t = k).
\end{aligned}$$

The final expression is the standard quantity controlled in regret proofs, and this argument can be reiterated without alteration. For completeness, we will provide a sketch of these proofs in the subsequent sections. In the case of KL-UCB, this essentially aligns with the argument by (Garivier and Cappé, 2011), while for the BAYESUCB bound, it follows the reasoning presented by (Kaufmann et al., 2012a), which is itself very similar to (Garivier and Cappé, 2011). To establish the efficiency of TS, we will utilize the argument outlined by (Agrawal and Goyal, 2013).

Remark on showing consistency of L_t^ :* It's worth noting that our choices of L_t^k are designed in a way that consistency proofs for U_t^* directly translate into those for L_t^* —this symmetry arises from the relevant functionals under the mappings $(S, \nu^k, \alpha) \mapsto (1 - S, 1 - \nu^k, 1 - \alpha)$. Following this transformation, $1 - L_t^k$ serves as an

upper bound of $1 - \nu^k$ in a U_t^k -type fashion. The argument for controlling $\sum \mathbb{P}(L_t^k \leq \alpha)$ for infeasible arms is essentially the same as that for controlling $\sum \mathbb{P}(U_t^k \geq \mu^*)$ for the standard bandit version of the respective method.

However, we observe a deviation in proving the consistency for BAYESUCB. Controlling standard regret in a Bayesian setting involves comparing two *random* indices, and (Kaufmann et al., 2012a) directly compare their index $U_{t,\text{BAYESUCB}}^*$ to μ^* to argue that N_t^* is at least logarithmically large. With this in hand, they assert that $U_{t,\text{BAYESUCB}}^*$ is at least $\mu^* - O(\sqrt{1/\log(T)})$ with high probability, suggesting that it is unlikely to be exceeded by suboptimal arms. However, to ensure the consistency of our (random) safety index L_t^* , we must compare it to a fixed value α . Consequently, the second argument utilizing a weakened consistency does not directly carry over. To address this, we adjust the quantiles δ_t^k sufficiently to ensure that the first argument alone is adequate to establish consistency. This introduces a gap, which might potentially be addressed with a more robust analysis.

Note on Dependency: The outlined overview does not leverage the potential dependence between the signals (R, S) . There is a possibility that such dependence could be exploited, and its significance might increase with an escalation in the number of safety constraints. We identify this as an area for potential exploration in future work.

A.2 Proof for Doubly Optimistic Confidence Bounds

The subsequent lemma essentially stems from the main result in the KL-UCB analysis by (Garivier and Cappé, 2011) and serves as a pivotal statement to establish our results. It is stated slightly more generically than in their paper, allowing us to employ the same result to demonstrate both gap-dependent and gap-independent bounds. We encountered this approach in the work of (Agrawal and Goyal, 2013).

Lemma A.2.1 (Adaptation of (Garivier and Cappé, 2011)). *For any suboptimal arm k , Algorithm 1, instantiated with the KL-UCB -type confidence bounds, achieves the*

following guarantees:

- If $\Delta^k > 0$, then for any $x \in (\mu^k, \mu^*)$,

$$\mathbb{E}[N_{T+1}^k] \leq \frac{\log T + 3 \log \log T}{d(x \|\mu^*)} + 6 \log \log T + \frac{2}{1 \wedge d(x \|\mu^k)} + 24. \quad (\text{A.1})$$

- If $\Gamma^k > 0$, then for any $y \in (\alpha, \nu^k)$,

$$\mathbb{E}[N_{T+1}^k] \leq \frac{\log T + 3 \log \log T}{d(y \|\alpha)} + \frac{2}{1 \wedge d(y \|\nu^k)}. \quad (\text{A.2})$$

We will first present the proofs of the two results using the above lemma and defer the proof of the lemma itself to the end.

A.2.1 Proof of Theorem 3.2.1

Proof. Consider an arbitrary arm k . If $\Delta^k > 0$, choose $x \in (\mu^k, \mu^*)$ such that $d(x \|\mu^*) = \frac{d(\mu^k \|\mu^*)}{1+\varepsilon}$. This choice exists as $d(x \|\mu^*)$ is a continuous function that monotonically decreases from $d(\mu^k \|\mu^*)$ to 0 as x varies in (μ^k, μ^*) . Our aim is to demonstrate that the third term in the bound of (A.1) is bounded by $O(\varepsilon^{-2})$. For small ε , we have $x = \mu^k + O(\varepsilon)$.

Let's abbreviate $d = d(\mu^k \|\mu^*)$, and observe that the derivative $d' := \partial_z d(z \|\mu^*)|_{z=\mu^k}$ is non-zero. Consequently, $x - \mu^k = \varepsilon \frac{d}{|d'|} + O(\varepsilon^2)$. Notably, since $d(z \|\mu^k)$ is minimized at $z = \mu^k$, we find that $d(x \|\mu^k) = \frac{1}{2} d'' \varepsilon^2 (d/d')^2 + O(\varepsilon^3)$, where $d'' = \partial_{zz}^2 d(z \|\mu^k)|_{z=\mu^k}$.

This leads to the conclusion that

$$\frac{2}{d(x \|\mu^k) \wedge 1} = O\left(\frac{d'^2}{d'' d^2 \varepsilon^2}\right),$$

which is a scaling of ε^{-2} by a problem-dependent constant.

Similarly, if $\Gamma^k > 0$, we proceed as above and choose $y \in (\alpha, \nu^k)$ such that $d(y \|\alpha) = \frac{d(\nu^k \|\alpha)}{1+\varepsilon}$. Using an entirely identical calculation, the final term of (A.2) is bounded as $O\left(\frac{f'^2}{f''} \frac{1}{d^2(\nu^k \|\alpha) \varepsilon^2}\right)$, where $f' = \partial_z d(z \|\alpha)|_{z=\nu^k}$, and $f'' = \partial_{zz}^2 d^2(z \|\nu^k)|_{z=\nu^k}$.

Utilizing both of these bounds, we conclude that

$$\begin{aligned}\mathbb{E}[N_{T+1}^k] &\leq \frac{1}{\mathbb{1}\{\mu^k < \mu^*\}} \left\{ \frac{(1+\varepsilon)\log T}{d(\mu^k\|\mu^*)} + \frac{(1+\varepsilon)3\log\log T}{d(\mu^k\|\mu^*)} \right. \\ &\quad \left. + 6\log\log T + 24 + O\left(\frac{(d'^2/\tilde{d}'')}{d^2(\mu^k\|\mu^*)\varepsilon^2}\right) \right\}, \\ \mathbb{E}[N_{T+1}^k] &\leq \frac{1}{\mathbb{1}\{\nu^k > \alpha\}} \left\{ \frac{(1+\varepsilon)\log T}{d(\nu^k\|\alpha)} + \frac{(1+\varepsilon)3\log\log T}{d(\nu^k\|\alpha)} + O\left(\frac{(f'^2/\tilde{f}'')}{d^2(\nu^k\|\alpha)\varepsilon^2}\right) \right\},\end{aligned}$$

where we set $1/\mathbb{1}\{p\} = \infty$ when the proposition p is untrue. Notably, recalling that $\mathbb{1}\{\mu^k < \mu^*\}d(\mu^k\|\mu^*) = d_{<}(\mu^k\|\mu^*)$ and similarly $d_{>}(\nu^k\|\nu^*)$, we can choose the tighter of the above bounds to obtain the result

$$\begin{aligned}\mathbb{E}[N_{T+1}^k] &\leq \frac{(1+\varepsilon)\log T}{d_{<}(\mu^k\|\mu^*) \vee d_{>}(\nu^k\|\nu^*)} \\ &\quad + O\left(\frac{\log\log T}{d_{<}(\mu^k\|\mu^*) \vee d_{>}(\nu^k\|\nu^*)} + \frac{1}{(d_{<}(\mu^k\|\mu^*) \vee d_{>}(\nu^k\|\nu^*))^2\varepsilon^2}\right)\end{aligned}$$

The claimed bounds now follow trivially. To control $\mathbb{E}[\mathcal{R}_T]$, simply multiply by the per-round regret of playing arm k , $\Delta^k \vee \Gamma^k$, and sum. To control $\mathbb{E}[\mathcal{U}_T]$, simply add up the above over the unsafe arms. □

Note that when the gaps Δ^k and Γ^k decrease, the last term in the expression scales as $1/(\Delta^k \wedge \Gamma^k)^4$, resulting in only a $T^{3/4}$ gap-independent bound.

A.2.2 Proof of Theorem 3.2.2

The gap-independent regret bounds follow from the observation that arms with very small gaps cannot incur substantial regret over T rounds. Let $\mathbf{M} > 0$ be a parameter to be determined later. Expressing regret as follows:

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k:\Delta^k > \Gamma^k \vee \mathbf{M}} \mathbb{E}[N_{T+1}^k]\Delta^k + \sum_{k:\Gamma^k > \Delta^k \vee \mathbf{M}} \mathbb{E}[N_{T+1}^k]\Gamma^k + \mathbf{M} \sum_{k:(\Delta^k \vee \Gamma^k) \leq \mathbf{M}} \mathbb{E}[N_{T+1}^k]. \quad (\text{A.3})$$

The last term is bounded by $\mathbf{M}T$, and thus, choosing \mathbf{M} of order $\sqrt{K \log T/T}$ controls regret. The remaining task is to show that $\mathbb{E}[N_{T+1}^k]$ is not too large for arms with large gaps. We develop bounds explicitly dependent on the gaps using (A.1) and (A.2).

Lemma A.2.2. *For any arm k with $\Delta^k > 0$,*

$$\mathbb{E}[N_{T+1}^k] \leq \frac{2 \log T + 6 \log \log T + 4}{(\Delta^k)^2} + 6 \log \log T + 24.$$

Similarly, for any arm k with $\Gamma^k > 0$,

$$\mathbb{E}[N_{T+1}^k] \leq \frac{2 \log T + 6 \log \log T + 4}{(\Gamma^k)^2}.$$

Proof. For an arm k with $\Delta^k > 0$, set $x = (\mu^k + \mu^*)/2 =: \bar{\mu}^k$. By Pinsker's inequality, $d(\bar{\mu}^k \parallel \mu^*) \geq 2(\mu^* - \bar{\mu}^k)^2 = (\Delta^k)^2/2$, and $d(\bar{\mu}^k \parallel \mu^k) \geq 2(\bar{\mu}^k - \mu^k)^2 = (\Delta^k)^2/2$. Plugging these into the bound yields the claim, noting that $(\Delta^k)^2/2 \leq 1$.

For arms with $\Gamma^k > 0$, a similar control results from (A.2) by setting $y = (\alpha + \nu^k)/2$. \square

With these bounds, we can now demonstrate the claim.

Proof of Theorem 3.2.2. The first term in (A.3) can be bounded as

$$\begin{aligned} & \sum_{\Delta^k > \Gamma^k \vee \mathbf{M}} \frac{2 \log T + 6 \log \log T + 2}{\Delta^k} + (6 \log \log T + 24) \Delta^k \\ & \leq K_\Delta \left(\frac{2 \log T + 6 \log \log T + 4}{\mathbf{M}} + 6 \log \log T + 24 \right), \end{aligned}$$

where $K_\Delta = |\{k : \Delta^k > \Gamma^k\}|$.

Similarly, the second term in (A.3) can be bounded as

$$\sum_{\Gamma^k > \Delta^k \vee \mathbf{M}} \frac{2 \log T + 6 \log \log T + 4}{\Gamma^k} \leq K_\Gamma \frac{2 \log T + 6 \log \log T + 4}{\mathbf{M}},$$

where $K_\Gamma = |\{k : \Gamma^k > \Delta^k\}|$.

Finally, observing that $K_\Gamma + K_\Delta \leq K$, we conclude that

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{K}{\mathbf{M}}(2 \log T + 6 \log \log T + 4) + (6 \log \log T + 24) \sum (\Delta^k \vee \Gamma^k) + T\mathbf{M}.$$

The claim follows by choosing $\mathbf{M} = \sqrt{K(2 \log T + 6 \log \log T + 4)/T}$, and noting that $2 \log T \geq 4$ for $T \geq 8$, and $2 \log \log T \leq \log T$ for all T . \square

A.2.3 Proof of Lemma A.2.1

Proof. We make the argument separately for infeasible and inefficient arms. The former is easier, so let us begin with it.

Infeasible arms

We follow the decomposition from §A.1. Recall that $L_t^k = \min\{q \leq \hat{\nu}_t^k : d(\hat{\nu}_t^k \| q) \leq \gamma_t/N_t^k\}$. Since $d(\hat{\nu}_t^k \| x)$ is a continuous decreasing function on $[0, \hat{\nu}_t^k]$, if $L_t^k \leq \alpha$ then it must either hold that $\hat{\nu}_t^k \leq \alpha$, or $d(\hat{\nu}_t^k \| \alpha) \leq d(\hat{\nu}_t^k \| L_t^k) = \gamma_t/N_t^k$. Either way, we have that $d_{>}(\hat{\nu}_t^k \| \alpha) \leq \gamma_t/N_t^k$.

Now, let $\hat{\nu}^k(s)$ denote the value of $\hat{\nu}_t^k$ after the s -th time we play the arm k . We observe that

$$\begin{aligned} \sum_t \mathbf{1}\{A_t = k\} &\leq \sum_{t=1}^T \mathbf{1}\{A_t = k, d_{>}(\hat{\nu}_t^k \| \alpha) \leq \gamma_t/N_t^k\} \\ &= \sum_{t=1}^T \sum_{s=1}^t \mathbf{1}\{A_t = k, sd_{>}(\hat{\nu}_t^k \| \alpha) \leq \gamma_t, N_t^k = s\} \\ &\leq \sum_{t=1}^T \sum_{s=1}^t \mathbf{1}\{A_t = k, N_t^k = s\} \cdot \mathbf{1}\{sd_{>}(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\} \\ &= \sum_{s=1}^T \mathbf{1}\{sd_{>}(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\} \cdot \sum_{t=s}^T \mathbf{1}\{A_t = k, N_t^k = s\} \\ &\leq \sum_{s=1}^T \mathbf{1}\{sd_{>}(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\}, \end{aligned}$$

where we have used that γ_t increases with T , and for any value s , there is at most one time step on which N_t^k is exactly s and we play the action k .

Now, we observe that for any $y \in (\alpha, \nu^k)$, the event $\{d_{>}(\hat{\nu}^k(s) \| \alpha) \leq d(y \| \alpha)\} = \{\hat{\nu}^k(s) \leq y\}$. Indeed, $d_{>}(u \| \alpha)$ is exactly equal to 0 for $u \leq \alpha$, and monotonically increasing for $u > \alpha$. Recalling Chernoff's bound (which applies since the random

variables are bounded in $[0, 1]$, $P(\hat{\nu}^k(s) \leq y) \leq \exp(-sd(y\|\nu^k))$. This sets up the following calculation.

Let $y \in (\alpha, \nu^k)$, and define $S(y) := \lfloor \gamma_T/d(y\|\alpha) \rfloor$, so that for all $s > S(y)$, $\gamma_T/s < d(y\|\alpha)$. Then

$$\begin{aligned}
\mathbb{E}[N_{T+1}^k] &= \sum_{t=1}^T \mathbb{P}(A_t = k) \\
&\leq \sum_{s=1}^T \mathbb{P}(sd_{>}(\hat{\nu}^k(s)\|\alpha) \leq \gamma_T) \\
&\leq S(y) + \sum_{s=S(y)+1}^T \mathbb{P}(d_{>}(\hat{\nu}^k(s)\|\alpha) \leq d(y\|\alpha)) \\
&\leq S(y) + \sum_{s=S(y)+1}^T e^{-sd(y\|\nu^k)} \\
&\leq S(y) + \frac{e^{-(S(y)+1)d(y\|\nu^k)}}{1 - e^{-d(y\|\nu^k)}} \\
&\leq S(y) + \frac{2}{1 \wedge d(y\|\nu^k)}, \tag{A.4}
\end{aligned}$$

where the last term uses that $(S(y) + 1)d(y\|\nu^k) \geq 0$, and $\frac{1}{1-e^{-u}} \leq \frac{2}{1 \wedge u}$. But $S(y) \leq \frac{\gamma_T}{d(y\|\alpha)} = \frac{\log T + 3 \log \log T}{d(y\|\alpha)}$.

Inefficient arms

Again, we follow the decomposition from §A.1, namely

$$\mathbb{E}[N_{T+1}^k] \leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \mathbb{P}(U_t^* < \mu^*) + \mathbb{P}(U_t^k \geq \mu^*).$$

Observe that $\mathbb{P}(k^* \notin \Pi_t) = \mathbb{P}(L_t^* > \alpha)$, which will be taken care of later.

As highlighted in Section A.1, the final term is regulated using the same approach as the inefficiency control. Specifically, considering $U_t^k = \max\{q \geq \hat{\mu}_t^k : d(\hat{\mu}_t^k\|q) \leq \gamma_t/N_t^k\}$, where $d(\hat{\mu}_t^k\|x)$ increases in the range $[\hat{\mu}_t^k, 1]$, if $U_t^k \geq \mu^*$, then either $\hat{\mu}_t^k \geq \mu^*$, or $d(\hat{\mu}_t^k\|\mu^*) \leq \gamma_t/N_t^k$. Employing a similar derivation, we find that

$$\sum_t \mathbb{1}\{A_t = k, U_t^k \geq \mu^*\} \leq \sum_{s=1}^T \mathbb{1}\{sd_{<}(\hat{\mu}^k(s)\|\mu^*) \leq \gamma_T\},$$

where $P(d_{<}(\hat{\mu}^k(s)\|\mu^*) \leq d(x\|\mu^*)) = P(\hat{\mu}^k(s) \leq x) \leq \exp(-sd(x\|\mu^k))$ for any $x \in$

(μ^k, μ^*) . The resulting sum yields the bound

$$\sum \mathbb{P}(U_t^k \geq \mu^*, A_t = k) \leq S(x) + \frac{2}{1 \wedge d(x \|\mu^k)},$$

where $S(x) \leq \frac{\gamma_T}{d(x \|\mu^*)}$.

The remaining task is to control $\sum \mathbb{P}(L_t^* > \alpha) + \mathbb{P}(U_t^* < \mu^*)$. To handle the second term, we first utilize the monotonicity of $d(\hat{\mu}_t^* \| q)$ on $[\hat{\mu}_t^*, 1]$. We note that

$$\{U_t^* < \mu^*\} = \{\max\{q > \hat{\mu}_t^* : d(\hat{\mu}_t^* \| q) \leq \gamma_t/N_t^*\} < \mu^*\} = \{\hat{\mu}_t^* < \mu^*, d(\hat{\mu}_t^* \|\mu^*) > \gamma_t/N_t^*\}.$$

The final event is addressed by (Garivier and Cappé, 2011, Theorem 10), which states that for any $z > 0$ and any k ,

$$\mathbb{P}(N_t^k d(\hat{\mu}_t^k \|\mu^k) > z) \leq e(z \log(t) + 1)e^{-z}. \quad (\text{A.5})$$

Applying (A.5) to $\hat{\mu}_t^*$ with $z = \gamma_t$, we find that

$$\mathbb{P}(U_t^* < \mu^*) \leq e(\gamma_t \log(t) + 1)e^{-\gamma_t},$$

which leads to

$$\begin{aligned} \sum_{t=3}^T \mathbb{P}(U_t^* < \mu^*) &\leq \sum_{t=3}^T \frac{e(\log^2 t + 3 \log t \cdot \log \log t + 1)}{t \log^3(t)} \\ &\leq e(\log \log T + 4). \end{aligned} \quad (\text{A.6})$$

Control on $\sum \mathbb{P}(L_t^* > \alpha)$ follows identically. By exploiting monotonicity twice,

$$\{L_t^* > \alpha\} = \{\hat{\nu}_t^* > \alpha, d(\hat{\nu}_t^* \|\alpha) > \gamma_t/N_t^*\} \subset \{\hat{\nu}_t^* > \nu^*, d(\hat{\nu}_t^* \|\nu^*) > \gamma_t/N_t^*\},$$

and thus, applying (A.5) to $\hat{\nu}_t^*$ with $z = \gamma_t$,

$$\sum_{t=3}^T \mathbb{P}(k^* \notin \Pi_t) = \sum_{t=3}^T \mathbb{P}(L_t^* > \alpha) \leq e \log \log T + 4e. \quad (\text{A.7})$$

Combining these results, we obtain

$$\mathbb{E}[N_{T+1}^k] \leq \frac{\log T + 3 \log \log T}{d(x \|\mu^k)} + 6 \log \log T + 24 + \frac{2}{1 \wedge d(x \|\mu^k)},$$

where $2e < 6$ and $8e + 2 < 24$. □

A.3 Proofs for Thompson Sampling with Optimistic Safety Indices

An initial observation is that, given the constancy of the safety index L_t^k , we can directly apply the proofs of Lemma A.2.1. This implies that the bounds (A.2) and (A.7) persist. In other words:

$$\mathbb{E}[N_{T+1}^k] \leq \inf_y \frac{1}{\mathbb{1}\{\alpha < y < \nu^k\}} \left(\frac{\log T + 3 \log \log T}{d(y|\alpha)} + \frac{2}{1 \wedge d(y|\nu^k)} \right),$$

$$\sum_{t=3}^T \mathbb{P}(k^* \notin \Pi_t) \leq e \log \log T + 4e.$$

The primary focus of the analysis is now to ensure the extension of the Thompson Sampling (TS) analysis to control the exploration of inefficient arms. This is achieved by leveraging the analysis of (Agrawal and Goyal, 2013) for TS, although alternative analyses such as that of (Kaufmann et al., 2012b) can be equivalently employed.

The central bound is summarized as follows:

Lemma A.3.1 (Adaptation of (Agrawal and Goyal, 2013)). *There exists a universal constant C such that if $\Delta^k > 0$, then for any u, v such that $\mu^k < u < v < \mu^*$:*

$$\sum_{t=1}^T \mathbb{P}(A_t = k, k^* \in \Pi_t) \leq \frac{\log T}{d(u|v)} + \frac{3}{1 \wedge d(u|\mu^k)} +$$

$$\frac{C}{(\mu^* - v)^2} \left(1 + \log \frac{1}{\mu^* - v} + \log \left(\frac{1}{1 - e^{-d(v|\mu^*)}} \wedge T(\mu^* - v) \right) \right) \quad (\text{A.8})$$

We will now demonstrate the result from the main text using the above Lemma.

Proof of Theorem 3.3.1. Let's establish the proof of the theorem.

For infeasible arms, instantiate (A.2) with a value y such that $d(y|\alpha) = d(\nu^k|\alpha)/(1 + \varepsilon)$. As mentioned earlier, the resulting $d(y|\nu^k)$ is on the order of $\Theta(\varepsilon^2)$.

Turning to inefficient arms, we decompose the expected number of pulls as follows:

$$\mathbb{E}[N_{T+1}^k] = \sum_{t=1}^T \mathbb{P}(A_t = k) \leq \sum_{t=1}^T \mathbb{P}(k^* \notin \Pi_t) + \sum_{t=1}^T \mathbb{P}(k^* \in \Pi_t, A_t = k).$$

The first term is bounded by $3 \log \log T$. For the second term, instantiate the bound (A.8) with u and v selected such that

1. $d(u \|\mu^*) = d(\mu^k \|\mu^*) / \sqrt{1 + \varepsilon}$
2. $d(u \|\nu) = d(u \|\mu^*) / \sqrt{1 + \varepsilon} = d(\mu^k \|\mu^*) / (1 + \varepsilon),$

both of which exist due to continuity.

Demonstrating the bound then necessitates controlling $u - \mu^k$ and $\mu^* - v$ (using the upper bound $d(a \| b) \geq 2(a - b)^2$). Accordingly, as in the proof of Theorem 3.2.1, notice that $u = \mu^k + \Theta(\sqrt{1 + \varepsilon} - 1) = \mu^k + \Theta(\varepsilon)$. Similarly, $v = \mu^* - \Theta(\varepsilon)$. Hence, $d(u \|\mu^k), d(v \|\mu^*) = \Theta(\varepsilon^{-2})$. Finally, since this ε^{-2} term does not grow with T , $(d(v \|\mu^*))^{-1} \wedge T = O(\varepsilon^{-2})$. We can now conclude the argument as in the proof of Theorem 3.2.1. □

Similarly to the situation with Algorithm 1, this approach also yields a gap-independent bound.

Proposition A.3.2. *Algorithm 2, instantiated with KL-UCB -type lower confidence bounds, achieves the gap-independent regret bound*

$$\mathbb{E}[\mathcal{R}_T] \leq O(\sqrt{KT \log T} + K \log \log T).$$

Proof. For infeasible arms, instantiate (A.2) with $y = (\alpha + \nu^k)/2$ to establish that

$$\mathbb{E}[N_{T+1}^k] \leq O\left(\frac{\log T}{(\Gamma^k)^2}\right)$$

For inefficient arms, instantiate (A.8) with $u = \mu^k + \Delta^k/3$, and $v = \mu^k + 2\Delta^k/3$. Then $\mu^* - v = v - u = u - \mu^k = \Delta^k/3$, and by observing that $d(v \|\mu^*)^{-1} \wedge T \Delta^k/3 \leq$

$T\Delta^k/3$, we have the upper bound

$$\begin{aligned}\mathbb{E}[N_{T+1}^k] &\leq O(\log \log T) + O\left(\frac{\log T}{(\Delta^k)^2} + \frac{1 + \log(1/\Delta^k) + \log(T\Delta^k)}{(\Delta^k)^2}\right) \\ &= O\left(\log \log T + \frac{\log T}{(\Delta^k)^2}\right).\end{aligned}$$

Taking the tighter of these bounds, and partitioning according to the size of $\Delta^k \vee \Gamma^k$, we obtain the bound

$$\mathbb{E}[\mathcal{R}_T] \leq \inf_{\mathbf{M} > 0} T\mathbf{M} + O\left(\frac{K \log T}{\mathbf{M}}\right) + O(K \log \log T),$$

confirming the claim upon optimization. \square

It remains to establish the key lemma. Once again, we emphasize that the fundamental ideas originate from (Agrawal and Goyal, 2013).

Proof of Lemma A.3.1. Let's fix a specific arm k . The values u and v essentially act as benchmarks against which we can compare the random scores ρ_t^* and ρ_t^k . To facilitate this, we introduce the 'good' events:

$$\begin{aligned}\mathcal{G}_t^{\mu,k} &:= \{\widehat{\mu}_t^k \leq u\}, \\ \mathcal{G}_t^{\rho,k} &:= \{\rho_t^k \leq v\}.\end{aligned}$$

Notably, $\mathcal{G}_t^{\mu,k}$ belongs to \mathcal{H}_{t-1} .

We initiate the argument with the decomposition:

$$\begin{aligned}\mathbb{P}(A_t = k, k^* \in \Pi_t) &= \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}) \\ &\quad + \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, (\mathcal{G}_t^{\rho,k})^c) \\ &\quad + \mathbb{P}(A_t = k, k^* \in \Pi_t, (\mathcal{G}_t^{\mu,k})^c) \\ &\leq \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}) \\ &\quad + \mathbb{P}(A_t = k, \mathcal{G}_t^{\mu,k}, (\mathcal{G}_t^{\rho,k})^c) \\ &\quad + \mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu,k})^c).\end{aligned}\tag{A.9}$$

Now, the last term in (A.9) is easily controlled - indeed, $\mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu,k})^c) =$

$\mathbb{P}(A_t = k, \hat{\mu}_t^k > u)$ is exponentially small if N_t^k is large. In fact, mirroring the approach of the proof of Lemma A.2.1, we find that:

$$\begin{aligned} \sum_{t \leq T} \mathbb{1}\{A_t = k, \hat{\mu}_t^k > u\} &= \sum_{t \leq T} \sum_{s \leq t} \mathbb{1}\{A_t = k, N_t^k = s, \hat{\mu}_t^k > u\} \\ &= \sum_s \mathbb{1}\{\hat{\mu}^k(s) > u\} \sum_{t \geq s} \mathbb{1}\{A_t = k, N_t^k = s\} \\ &\leq \sum_{s \leq T} \mathbb{1}\{\hat{\mu}^k(s) > u\}, \end{aligned}$$

where we set $\hat{\mu}^k(s)$ to be the value of $\hat{\mu}_t^k$ at the first t such that $N_t^k = s$. By Chernoff's bound, $P(\hat{\mu}^k(s) > u) \leq \exp(-sd(u\|\mu^k))$, yielding the bound:

$$\sum \mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu, k})^c) \leq \frac{2}{1 \wedge d(u\|\mu^k)}. \quad (\text{A.10})$$

The second term in (A.9) is also amenable to control, particularly when observing the highly concentrated nature of the posterior Beta distribution around $\hat{\mu}_t^k$ with a variance scale of $1/N_t^k$. To elucidate this point further, (Agrawal and Goyal, 2013) leverage the following insight: If $F(x; \text{Beta}(a, b))$ represents the cumulative distribution function (CDF) of a $\text{Beta}(a, b)$ random variable, and $G(k; \text{Bin}(n, p))$ denotes the CDF of a Binomial random variable, then for natural $n \geq k$, the relation holds:

$$1 - F(x; \text{Beta}(k + 1, n - k + 1)) = G(k; \text{Bin}(n + 1, x)).$$

This relation is most easily derived from the fact that $\text{Beta}(k + 1, n - k + 1)$ is the distribution of the $k + 1$ -th order statistic of $n + 1$ samples from the uniform distribution. The probability of this exceeding x is precisely the probability that the k smaller ones are at most x , and the rest are at least x , which is expressed by the Binomial distribution. Consequently, we infer that for any N_0 , the probability that ρ_t^k assumes a large value can be bounded akin to a Binomial tail. For any v , this leads us to the conclusion:

$$\mathbb{P}(\rho_t^k > v | N_t^k, \hat{\mu}_t^k) \leq \exp(-N_t^k d_{>}(v\|\hat{\mu}_t^k)),$$

which follows straightforwardly from Chernoff's bound. Using a similar reasoning, we

can deduce that for any N_0 and $u < v$:

$$\mathbb{P}(\rho_t^k > v | N_t^k > N_0, \hat{\mu}_t^k \leq u) \leq e^{-N_0 d(v||u)},$$

Opting for $N_0 = \log(T)/d(v||u)$, we can derive the bound as follows:

$$\begin{aligned} \sum_{t=1}^T \mathbb{P}(A_t = k, \rho_t^k > v, \hat{\mu}_t^k \leq u) &\leq \sum_{t=1}^T \mathbb{P}(A_t = k, N_{T+1}^k \leq N_0) \\ &\quad + \sum_{t=1}^T \mathbb{P}(N_{T+1}^k > N_0, \rho_t^k > v, \hat{\mu}_t^k \leq u) \\ &\leq N_0 + T e^{-N_0 d(v||u)} \\ &\leq \frac{\log T}{d(v||u)} + 1 \leq \frac{\log T}{d(v||u)} + \frac{1}{1 \wedge d(u||\mu^k)}. \end{aligned} \quad (\text{A.11})$$

This leaves us with the first term of (A.9), which is the most challenging to control and ultimately relies on a rigorous analysis of Binomial tails. The general strategy involves employing v as a lower index for ρ_t^* . Specifically, let $\mathbf{P}_t := \mathbb{P}(\rho_t^* > v | \mathcal{H}_{t-1}) = \mathbb{P}_{t-1}(\rho_t^* > v)$. Observe that

$$\begin{aligned} \mathbb{P}_{t-1}(A_t = k, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}, k^* \in \Pi_t) &= \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \mathbb{P}_{t-1}(A_t = k, \rho_t^k \leq v) \\ &\leq \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \mathbb{P}_{t-1}(\forall k \in \Pi_t, \rho_t^k \leq v) \\ &= \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} (1 - \mathbf{P}_t) \mathbb{P}_{t-1}(\forall k \neq k^* \in \Pi_t, \rho_t^k \leq v) \\ &= \frac{1 - \mathbf{P}_t}{\mathbf{P}_t} \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \\ &\quad \times \mathbb{P}_{t-1}(\forall k \neq k^* \in \Pi_t, \rho_t^k \leq v < \rho_t^*) \\ &\leq \frac{1 - \mathbf{P}_t}{\mathbf{P}_t} \mathbb{P}_{t-1}(A_t = k^*), \end{aligned}$$

where we have utilized the fact that $\mathcal{G}_t^{\mu,k} \in \mathcal{H}_{t-1}$ and Π_t is predictable. The key is to exploit the fact that \mathbf{P}_t exponentially approaches 1 as N_t^* increases. By expressing this probability in terms of the size of N_t^* and analyzing it, (Agrawal and Goyal, 2013) show in their Lemma 2 that

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}[(1 - \mathbf{P}_t) \mathbb{P}_{t-1}(A_t = k^*) / \mathbf{P}_t] \\ & \leq \frac{24}{\Delta_v^2} + C' \sum_{s \geq 8/\Delta_v}^{T-1} e^{-\Delta_v^2 s/2} + \frac{1}{e^{\Delta_v^2 s/4} - 1} + \frac{e^{-sd(v\|\mu^*)}}{(s+1)\Delta_v^2}, \end{aligned}$$

where $\Delta_v := (\mu^* - v)$, and C' is a constant. Notably, each term in the sum is monotonically decreasing. Therefore, upper bounds can be derived by comparison to an integral, yielding for the first and second terms that

$$\sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} e^{-\Delta_v^2 s/2} \leq \int_0^{\infty} e^{-\Delta_v^2 s/2} ds = \frac{2}{\Delta_v^2},$$

and

$$\begin{aligned} \sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} \frac{1}{e^{\Delta_v^2 s/4} - 1} & \leq \int_{7/\Delta_v}^T \frac{1}{e^{\Delta_v^2 s/4} - 1} ds \\ & = \frac{4}{\Delta_v^2} \int_{7/4\Delta_v}^{\Delta_v^2 T/4} \frac{1}{e^u - 1} du \\ & \leq \frac{4}{\Delta_v^2} \log \frac{1}{1 - e^{-7/4\Delta_v}} \\ & \leq \frac{4}{\Delta_v^2} \log \frac{2}{1 \wedge 7/4\Delta_v} \leq \frac{4}{\Delta_v^2} \left(\log \frac{1}{\Delta_v} + O(1) \right), \end{aligned}$$

where we have employed the previously established fact that $\frac{1}{1-e^{-x}} \leq \frac{2}{x \wedge 1}$.

For the final term, we can bound it in two ways - firstly, by observing that $e^{-sd} \leq 1$, we obtain the bound $\sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} \frac{1}{s+1} \leq \log(T\Delta/8)$. Additionally, we derive a T -independent bound as follows, wherein we abbreviate $d_v = d(v\|\mu^*)$.

$$\begin{aligned}
\sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} \frac{e^{-sd_v}}{(s+1)\Delta_v^2} &= e^{d_v} \sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} \frac{e^{-(s+1)d_v}}{s+1} \\
&= e^{d_v} \sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} \int_{u=d_v}^{\infty} e^{-(s+1)u} du \\
&\leq e^{d_v} \int_{u=d_v}^{\infty} \sum_{s=1}^{\infty} e^{-(s+1)u} du \\
&= e^{d_v} \int_{u=d_v}^{\infty} \frac{e^{-2u}}{1-e^{-u}} du \\
&\leq \log \frac{1}{1-e^{-d_v}} \leq \log \frac{2}{d_v} + O(1).
\end{aligned}$$

By taking the smaller of these two bounds, the final term is controlled by $4\Delta_v^{-2}[\log(T\Delta_v \wedge d_v^{-1}) + O(1)]$, and we obtain

$$\sum_{t=1}^T \mathbb{P}(A_t = k, \mathcal{G}_t^{\mu, k}, \mathcal{G}_t^{\rho, k}, k^* \in \Pi_t) \leq \frac{C}{\Delta_v^2} \left(1 + \log \frac{1}{\Delta_v} + \log (\Delta_v T \wedge -d(v\|\mu^*)^{-1}) \right). \tag{A.12}$$

The asserted bound is then realized by adding up (A.10, A.11, A.12). \square

A.4 Proofs for Thompson Sampling with BAYESUCB

As the procedure for selecting arms given Π_t remains unchanged from the previous case, our focus is on demonstrating that Π_t is effective, specifically, that the lower bound index L_t^k performs well. This essentially leverages the fact that the argument from the previous section only relies on Π_t being a predictable process, along with details of the Thompson scores ρ_t^k . Therefore, the second term of the decomposition

$$\sum_t \mathbb{P}(A_t = k) \leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \sum_t \mathbb{P}(k^* \in \Pi_t, A_t = k)$$

can be handled similarly to control the actions of inefficient arms on rounds where $k^* \in \Pi_t$, yielding (A.8).

We establish the following bound, utilizing the techniques of (Kaufmann et al., 2012a) as described in §A.1.

Lemma A.4.1. *In the setting of Theorem 3.3.2, the following hold:*

- If $\Gamma^k > 0$, then for any $x \in (\alpha, \nu^k)$,

$$\mathbb{E}[N_{T+1}^k] \leq \frac{3/2 \log T + 3 \log \log T + 3/2 \log 2}{d(x||\alpha)} + \frac{2}{1 \wedge d(x||\nu^k)} \quad (\text{A.13})$$

- The mean number of times the optimal arm is treated as impermissible is bounded as

$$\sum_{t=3}^T \mathbb{P}(k^* \notin \Pi_t) \leq e \log \log T + 4e.$$

The claimed bound is readily obtained by combining the relevant components of the proofs of Theorems 3.2.1 and 3.3.1.

Proof of Theorem 3.3.2. For inefficient arms, combining the second part of Lemma A.4.1 and (A.8), we conclude that if $\Delta^k > 0$, then

$$\mathbb{E}[N_{T+1}^k] \leq \frac{\log T}{d(u||v)} + \frac{3}{1 \wedge d(u||\mu^k)} + \frac{C}{(\mu^* - v)^2} (1 + (d(v||\mu^*)^{-1} \wedge \log T)) + e \log \log T + 4e.$$

Similarly, for infeasible arms, by using (A.13), we have the control

$$\mathbb{E}[N_{T+}^k] \leq \frac{\log T + 3 \log \log T + \log 2}{2/3 d(y||\alpha)} + \frac{2}{1 \wedge d(y||\nu^k)}.$$

Now choosing u, v, y as in the proof of Theorem 3.3.1 and proceeding along the same lines gives the claim. \square

The same strategy also yields the following gap-independent result. The proof is identical and therefore omitted.

Proposition A.4.2. *Algorithm 3 instantiated with BAYESUCB using $\delta_t^k = 1/\sqrt{8N_t^k t \log^3 t}$ also satisfies the bound*

$$\mathbb{E}[\mathcal{R}_T] = O(\sqrt{KT \log T} + K \log \log T).$$

We now proceed to demonstrate the main lemma.

Proof of Lemma A.4.1. The argument relies on the following estimate, essentially serving as a reduction to the analysis of KL-UCB. This result is a variation of Lemma 1 of (Kaufmann et al., 2012a).

Lemma A.4.3. *Define the quantities*

$$\begin{aligned} \underline{\varphi}_t^k &:= \mathbb{1}\{S_t^k > 0\} \min \left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d \left(\frac{S_t^k}{N_t^k} \parallel q \right) \leq \log((2t \log^2 t)^{3/2}) \right\} \\ \overline{\varphi}_t^k &:= \mathbb{1}\{S_t^k > 0\} \min \left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d \left(\frac{S_t^k}{N_t^k} \parallel q \right) \leq \log(t \log^3(t)) \right\}. \end{aligned}$$

Then for all t ,

$$\underline{\varphi}_t^k \leq L_t^k \leq \overline{\varphi}_t^k.$$

Proof. Firstly, since $L_t^k = 0$ whenever $S_t^k = 0$, this case is trivial. So assume $S_t^k \geq 1$.

The idea behind the bounds is to exploit the relationship between the CDFs of Beta and Binomial random variables to reduce the quantile estimation to that of a Binomial. Then, we use Chernoff's bound for the Binomial to control where the quantile can be. Let $Z \sim \text{Beta}(S_t^k, N_t^k - S_t^k + 1)$. Then, we know that

$$\mathbb{P}(Z \leq q) = \mathbb{P}(\text{Bin}(N_t^k, q) \geq S_t^k).$$

Further, by Chernoff's upper bound and by estimating the s th term in the Binomial series using Stirling's approximation, we may show the following result (where the lower bound holds generally, and the upper bound holds for any $s \geq nq$).

$$\frac{1}{\sqrt{8n}} \exp(-nd((s/n) \parallel q)) \leq \mathbb{P}(\text{Bin}(n, q) \geq s) \leq \exp(-nd((s/n) \parallel q)).$$

Now, recall that L_t^k is the δ_t^k th quantile of the law of Z , so that $\mathbb{P}(Z \leq L_t^k) = \delta_t^k$.

Lower bound Suppose $q \leq S_t^k/N_t^k$ is such that

$$\exp(-N_t^k d(S_t^k/N_t^k \| q)) \leq \delta_t^k.$$

Then it follows that $q \leq L_t^k$. Therefore,

$$\begin{aligned} L_t^k &\geq \max \left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d \left(\frac{S_t^k}{N_t^k} \| q \right) \geq \log(1/\delta_t^k) \right\} \\ &= \min \left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d \left(\frac{S_t^k}{N_t^k} \| q \right) \leq \log(1/\delta_t^k) \right\}, \end{aligned}$$

where the final equality is due to the continuity of $d(a \| \cdot)$.

Now observe that

$$\log(1/\delta_t^k) \leq (2(t+1))^{3/2} \log^3 t.$$

Therefore, replacing $\log(1/\delta_t^k)$ by the larger $\log(2(t+1))^{3/2} \log^3 t$ in the lower bound can only decrease it.

Upper bound Suppose that $q \leq S_t^k/N_t^k$ is such that the lower bound on the Binomial tail exceeds δ_t^k . Then L_t^k must be smaller than this q , and so

$$L_t^k \leq \min \left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d \left(\frac{S_t^k}{N_t^k} \| q \right) \leq \log \left(\frac{1}{\sqrt{8N_t^k \delta_t^k}} \right) \right\}.$$

But, by definition,

$$\frac{1}{\sqrt{8N_t^k \delta_t^k}} = t \log^3 t. \quad \square$$

Note that the bounds $\bar{\varphi}$ and $\underline{\varphi}$ exactly follow the KL-UCB bounds' structure, with a distinct value for γ_T . Thus, the same proofs can be replicated.

To establish (A.13), for an arm with a safety gap, $\{L_t^k \leq \alpha\} \subset \{\underline{\varphi}_t^k \leq \alpha\}$, and we can then employ the proof of Lemma A.2.1 to control this identically—the only change being the replacement of $\log(\gamma_T)$ in $S(y)$ with $\log((2t \log^2 t)^{3/2})$.

Moreover, the upper bound aligns precisely with the KL-UCB bound, and hence without modification, we can immediately conclude that

$$\sum_{t \geq 3} \mathbb{P}(L_t^* > \alpha) \leq \sum_{t \geq 3} \mathbb{P}(\bar{\varphi}_t^* > \alpha) \leq e \log \log t + 4e. \quad \square$$

It's worth noting that the final property in the proof of Lemma A.4.3 is the motivation for selecting δ_t as we did. Essentially, this choice corresponds to the $1/\gamma_t$

from KL-UCB, scaled down to ensure that the BAYESUCB bound is at least as optimistic as that of KL-UCB. In principle, this offers the potential for a more refined analysis by choosing a more nuanced δ_t through exploiting stronger bounds for the Binomial tails.

For instance, it is known (Jeřábek, 2004, Prop A.4, A.2) that there exists a constant C such that for $s \geq nq + \sqrt{nq(1-q)}$,

$$\frac{1}{C} \frac{qn - qs}{s - qn} \sqrt{\frac{n}{s(n-s)}} e^{-nd(s/n||q)} \leq \mathbb{P}(\text{Bin}(n, q) \geq s) \leq C \frac{qn - qs}{s - qn} \sqrt{\frac{n}{s(n-s)}} e^{-nd(s/n||q)},$$

while for $s \leq nq + \sqrt{nq(1-q)}$, it is bounded below by another constant C' . This suggests using $\delta_t \sim \min\left(C', \frac{1}{t \log^3 t} \cdot \sqrt{\frac{N_t^k}{S_t^k(N_t^k - S_t^k)}}\right)$, although it is unclear how to handle the $(qn - qs)/(s - qn)$ term properly. Assuming this is indeed handled, this should result in an improvement to $\bar{\varphi}$, replacing $t^{3/2}$ by something $O(t)$, while the lower bound should remain unchanged. Of course, this does not quite explain the success of $\delta_t = 1/t$ in the experiments, and it is possible that this approach simply serves to make BAYESUCB look more like KL-UCB, which defeats the purpose somewhat.

A.5 Lower Bound

We start by presenting the main Lemma.

Proof of Lemma 3.4.1. Consider a (potentially randomized) algorithm. Let $\{\mathbb{P}^k\}$ and $\{\tilde{\mathbb{P}}^k\}$ be two safe bandit instances, and let $\mathcal{H}_t := \{(A_s, R_s, S_s) : s \leq t\}$ denote the history of play. We use \mathbb{P} to represent laws in the first instance and $\tilde{\mathbb{P}}$ for laws in the second. Similarly, \mathbb{E} and $\tilde{\mathbb{E}}$ denote expectations under the two laws.

Take any function Z measurable with respect to $\sigma(\mathcal{H}_T)$, bounded in $[0, 1]$. From \mathcal{H}_T , a random bit can be generated by computing $Z(\mathcal{H}_T)$ and then sampling $B \sim \text{Bern}(Z)$. The mean of B equals that of Z . Using the data processing inequality, we have

$$D(\mathbb{P}_{\mathcal{H}_T} \| \tilde{\mathbb{P}}_{\mathcal{H}_T}) \geq D(\mathbb{P}_B \| \tilde{\mathbb{P}}_B) = d(\mathbb{E}[Z] \| \tilde{\mathbb{E}}[Z]).$$

Applying the chain rule of KL divergence for $t \geq 1$, we get

$$\begin{aligned} D(\mathbb{P}_{\mathcal{H}_t} \|\tilde{\mathbb{P}}_{\mathcal{H}_t}) &= D(\mathbb{P}_{\mathcal{H}_{t-1}} \|\tilde{\mathbb{P}}_{\mathcal{H}_{t-1}}) \\ &\quad + \mathbb{E}[D(\mathbb{P}_{A_t|\mathcal{H}_{t-1}} \|\tilde{\mathbb{P}}_{A_t|\mathcal{H}_{t-1}} | \mathcal{H}_{t-1})] \\ &\quad + \mathbb{E}[D(\mathbb{P}_{(R_t, S_t)|A_t, \mathcal{H}_{t-1}} \|\tilde{\mathbb{P}}_{(R_t, S_t)|A_t, \mathcal{H}_{t-1}} | A_t, \mathcal{H}_{t-1})]. \end{aligned}$$

The second term on the right is 0 due to causality, and the feedback (R_t, S_t) is independent of history given A_t , distributed according to \mathbb{P}^{A_t} and $\tilde{\mathbb{P}}^{A_t}$ in the two instances. This leads to the recurrence

$$D(\mathbb{P}_{\mathcal{H}_t} \|\tilde{\mathbb{P}}_{\mathcal{H}_t}) - D(\mathbb{P}_{\mathcal{H}_{t-1}} \|\tilde{\mathbb{P}}_{\mathcal{H}_{t-1}}) = \sum_k \mathbb{P}(A_t = k) D(\mathbb{P}^k \|\tilde{\mathbb{P}}^k).$$

Summing this up and noting that \mathcal{H}_0 is trivial, and recalling $\sum_t \mathbb{P}(A_t = k) = \mathbb{E}[N_{T+1}^k]$, it follows that

$$D(\mathbb{P}_{\mathcal{H}_T} \|\tilde{\mathbb{P}}_{\mathcal{H}_T}) = \sum_k \mathbb{E}[N_{T+1}^k] D(\mathbb{P}^k \|\tilde{\mathbb{P}}^k).$$

The conclusion follows by taking $Z = N_{T+1}^{k_0}/T$, which trivially lies in $[0, 1]$. \square

Next, we present the proof of Theorem 3.4.2.

Proof of Theorem 3.4.2. Select $\tilde{\mathbb{P}}^j = \mathbb{P}^j$ for $j \neq k$, and let $\tilde{\mathbb{P}}^k$ be any law on $\{0, 1\}^2$ with means $(\mu^k \vee \mu^* + \varepsilon, \nu^k \wedge \alpha)$. In the $\tilde{\mathbb{P}}$ -instance, arm k is optimal.

As the algorithm ensures that suboptimal arms are not played more than $C_x T^x$ times, we have $\mathbb{E}[N_{T+1}^k/T] \leq C_x T^{-(1-x)}$ and $\tilde{\mathbb{E}}[N_{T+1}^k/T] \geq 1 - C_x T^{-(1-x)}$ for any $x \in (0, 1)$. Therefore,

$$\begin{aligned} &d(\mathbb{E}[N_{T+1}^k/T] \|\tilde{\mathbb{E}}[N_{T+1}^k/T]) \\ &\geq \left(1 - \frac{\mathbb{E}[N_{T+1}^k]}{T}\right) \log \frac{1}{1 - \tilde{\mathbb{E}}[N_{T+1}^k/T]} - \log 2 \\ &\geq (1 - o(1))(1 - x) \log \frac{T}{C_x} - \log 2 = (1 - o(1))(1 - x) \log T. \end{aligned}$$

Since we are working with independent means and safety rewards, taking $\tilde{\mathbb{P}}^k$ to also have independent rewards, we get $D(\mathbb{P}^k \|\tilde{\mathbb{P}}^k) = d_{<}(\mu^k \|\mu^* + \varepsilon) + d_{>}(\nu^k \|\alpha)$.

Therefore, for any $x, \varepsilon \in (0, 1)$,

$$\frac{\mathbb{E}[N_{T+1}^k]}{\log T} \geq \frac{(1-x)(1-o(1))}{d_{<}(\mu^k \|\mu^* + \varepsilon) + d_{>}(\nu^k \|\alpha)},$$

and the claim follows by taking $\underline{\lim}_{T \nearrow \infty}$ and then taking limits as $x \rightarrow 0, \varepsilon \rightarrow 0$, exploiting the continuity of $d_{<}(a \| b)$. \square

Appendix B

Supplement for § 4

B.1 Quantitative Bounds from the Theory of Online Linear Regression

We conclude the preliminaries with the following generic statement, which holds due to a couple of applications of the matrix-determinant lemma. The result is standard - see the discussions of ([Abbasi-Yadkori et al., 2011](#), Lemma 11) for historical discussions.

Lemma B.1.1. *Let $\{x_t\}$ be the actions of DOSLB. Suppose that for all t , $\|x_t\| \leq 1$, and let $\lambda \geq 1$. Then for any T ,*

$$\sum_{t=1}^T \|x_t\|_{V_{t-1}}^2 \leq \frac{3}{2} \log \left(\frac{\det(V_T)}{\det(\lambda I)} \right) \leq \frac{3}{2} d \log \left(1 + \frac{T}{\lambda d} \right).$$

Proof of Lemma B.1.1. First notice that since $V_t = V_{t-1} + x_t x_t^\top$, by the matrix-determinant lemma,

$$\det(V_t) = \det(V_{t-1}) \det(I + V_{t-1}^{-1/2} x_t x_t^\top (V_{t-1}^{-1/2})^\top) = \det(V_{t-1}) (1 + \|x_t\|_{V_{t-1}}^2),$$

and induction yields

$$\det(V_T) = \det(\lambda I) \prod_{t=1}^T (1 + \|x_t\|_{V_{t-1}}^2).$$

where we have used that $V_0 = \lambda I$.

Now, notice that since $V_{t-1} \succ \lambda I$ for each t , it follows that $\|x_t\|_{V_{t-1}}^2 \leq \|x_t\|^2 / \lambda \leq 1$. But for $z \in [0, 1]$, $z \leq \frac{3}{2} \log(1 + z)$, which implies that

$$\sum \|x_t\|_{V_{t-1}}^2 \leq \frac{3}{2} \sum \log(1 + \|x_t\|_{V_{t-1}}^2) = \frac{3}{2} \log \frac{\det(V_T)}{\det(\lambda I)}.$$

Finally, note that since V_T is positive definite, by an application of the AM-GM inequality, $\det(V_T) \leq (\text{trace}(V_T)/d)^d$, and further, $\text{trace}(V_T) = d\lambda + \sum_t \|x_t\|_2^2 \leq d\lambda + T$. Further observing that $\det(\lambda I) = \lambda^d$, we conclude that

$$\log \frac{\det(V_T)}{\det(V)} \leq d \log \frac{(d\lambda + T)/d}{\lambda} = d \log \left(1 + \frac{T}{d\lambda} \right).$$

□

An immediate consequence of the above is the following pair of observations which we shall use frequently.

Lemma B.1.2. *Let $\{x_t\}$ be the actions of DOSLB run with the parameters λ, δ . For every $T > 0$,*

$$\sum_{t \leq T} \rho_t(x_t; \delta)^2 \leq 3d^2 \log^2 \left(1 + \frac{T}{\lambda d} \right) + 6d \log \left(1 + \frac{T}{\lambda d} \right) \left(\log \frac{U+1}{\delta} + 2\lambda \right), \quad (\text{B.1})$$

$$\sum_{t \leq T} \rho_t(x_t; \delta) \leq d\sqrt{3T} \log \left(1 + \frac{\log T}{d\lambda} \right) + \sqrt{3dT \log \left(1 + \frac{T}{\lambda d} \right)} \left(\sqrt{2\lambda} + \sqrt{\log \frac{U+1}{\delta}} \right). \quad (\text{B.2})$$

These bounds supply the core bounds needed to convert the control we develop on ρ_t in §4.5 and §4.6 into control on \mathcal{E}_T and \mathcal{S}_T . Observe that the main terms in the above results do not show dependence on the failure probability parameter δ .

Proof of Lemma B.1.2. Recall that $\rho_t(x_t; \delta) = 2\sqrt{\omega_t(\delta)} \cdot \|x_t\|_{V_{t-1}^{-1}}$. Further observe that ω_t is an increasing function of t . Immediately by Lemma B.1.1,

$$\sum \rho_t^2 \leq 4\omega_T(\delta) \sum \|x_t\|_{V_{t-1}^{-1}}^2 \leq 6d\omega_T(\delta) \log \left(1 + \frac{T}{d\lambda} \right).$$

Further, once again applying Lemma B.1.1, and noting that $(\sqrt{u} + \sqrt{v})^2 \leq 2u + 2v$,

$$\begin{aligned} \sqrt{\omega_T(\delta)} &= \sqrt{\lambda} + \sqrt{\frac{1}{2} \log \frac{U+1}{\delta} + \frac{1}{4} \log \frac{\det(V_T)}{\det(\lambda I)}} \\ \implies \omega_T(\delta) &\leq 2\lambda + \log \frac{U+1}{\delta} + \frac{d}{2} \log \left(1 + \frac{T}{\lambda d} \right). \end{aligned}$$

Multiplying these two bounds controls $\sum \rho_t^2$.
Further, by the Cauchy-Schwarz inequality,

$$\sum_{t=1}^T \rho_t \leq \sqrt{T} \cdot \sqrt{\sum_{t=1}^T \rho_t^2}.$$

The bound (B.2) follows upon applying the bound on $\sum \rho_t^2$ above, and then using the trivial relation $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$. \square

We state the consistency of confidence sets as follows, which is adopted from (Abbasi-Yadkori et al., 2011).

Lemma B.1.3. (Abbasi-Yadkori et al., 2011, Thm.2) *The confidence sets are consistent, i.e.,*

$$\forall \lambda \geq 1, \delta \in (0, 1), \quad \mathbb{P}(\forall t, \theta \in \mathcal{C}_t^0(\delta), A \in \mathcal{C}_t(\delta)) \geq 1 - \delta.$$

Lemma B.1.3 yields the following key bound in terms of the *noise scale at x at time t* , defined as

$$\rho_t(x; \delta) := 2\sqrt{\omega_{t-1}(\delta)} \|x\|_{V_{t-1}^{-1}}.$$

Finally, let us argue that the quantity $\rho_t(x; \delta)$ indeed controls the noise scale of the problem by showing Lemma B.1.4.

Lemma B.1.4. *If the confidence sets are consistent, i.e., if $A \in \mathcal{C}_t(\delta)$ and $\theta \in \mathcal{C}_t^0(\delta)$, then $\forall x \in \mathcal{X}$,*

$$\forall i \in [1 : U], \max_{\tilde{a}^i \in \mathcal{C}_t^i(\delta)} |\langle \tilde{a}^i - a^i, x \rangle| \leq \rho_t(x; \delta), \quad \text{and} \quad \max_{\tilde{\theta} \in \mathcal{C}_t^0(\delta)} |\langle \tilde{\theta} - \theta, x \rangle| \leq \rho_t(x; \delta).$$

Proof of Lemma B.1.4. Let us take the case of θ . The bounds for the a^i 's arise in exactly the same way. Under the assumption of consistency, $\theta \in \mathcal{C}_t^0$. Therefore

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq |\langle \tilde{\theta} - \hat{\theta}, x \rangle| + |\langle \theta - \hat{\theta}, x \rangle|.$$

By exploiting the positive definiteness of V_{t-1} and the Cauchy-Schwarz inequality, we can further observe that

$$|\langle \theta - \hat{\theta}, x \rangle| = |\langle (\theta - \hat{\theta})V_{t-1}^{1/2}, V_{t-1}^{-1/2}x \rangle| \leq \|\theta - \hat{\theta}\|_{V_{t-1}} \cdot \|x\|_{V_{t-1}^{-1}}.$$

Running the same calculation of $\tilde{\theta}$ and adding the bounds, we conclude that

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq (\|\theta - \hat{\theta}\|_{V_{t-1}} + \|\tilde{\theta} - \hat{\theta}\|_{V_{t-1}}) \|x\|_{V_{t-1}^{-1}}$$

But both $\theta, \tilde{\theta} \in \mathcal{C}_t^0$, which by definitions means that their V_{t-1} -norm distance from $\hat{\theta}$ is bounded by $\sqrt{\omega_{t-1}(\delta)}$. The claim is immediate upon recalling that $\rho_t(x; \delta) := 2\sqrt{\omega_{t-1}(\delta)} \|x\|_{V_{t-1}^{-1}}$. \square

B.2 Appendix on the Structural Behavior of DOSLB

This section aims to demonstrate the essential structural characteristics of the behavior of DOSLB as discussed in §4.5. Specifically, we present the main outcome of §4.5, affirming that any point played by DOSLB must activate some BIS in a noisy manner. To achieve this, we initially define the behavior of DOSLB concerning polytopes contained within the permissible set. Before delving into this, it is important to note that an extreme point of a polytope (or any closed convex set) refers to a point not situated on a line connecting two other points within the polytope. Additionally, each extreme point of a polytope in \mathbb{R}^d must satisfy at least d constraints with equality. For a polytope \mathcal{P} , its extreme points are denoted as $\mathcal{E}_{\mathcal{P}}$.

Lemma B.2.1. *Assume \mathcal{P} is a polytope such that $\mathcal{P} \subset \tilde{\mathcal{S}}_t$. If DOSLB plays within \mathcal{P} , then x_t must be an extreme point of \mathcal{P} , implying $x_t \in \mathcal{P} \implies x_t \in \mathcal{E}_{\mathcal{P}}$.*

Now, we can assert that Proposition 4.4.5 logically follows from the aforementioned Lemma.

Proof of Proposition 4.4.5. Consider a chosen $\tilde{A} \in \mathcal{C}_t$ and define the polytope

$$\mathcal{P}(\tilde{A}) = \{x : \tilde{A}x \leq \alpha, Bx \leq \beta\}.$$

Observe that

$$\tilde{\mathcal{S}}_t = \bigcup_{\{\tilde{A} \in \mathcal{C}_t\}} \{x : \tilde{A}x \leq \alpha, Bx \leq \beta\} = \bigcup_{\tilde{A} \in \mathcal{C}_t} \mathcal{P}(\tilde{A}),$$

indicating that $\tilde{\mathcal{S}}_t$ can be expressed as the union of polytopes. Thus, the chosen point x_t must reside in one of these polytopes, denoted as \mathcal{P}^* .

Now, as $\mathcal{P}^* \subset \tilde{\mathcal{S}}_t$, and $x_t \in \mathcal{P}^*$, according to Lemma B.2.1, x_t must be an extreme point of \mathcal{P}^* . This implies the activation of at least d total constraints between $\tilde{A}x \leq \alpha$ and $Bx \leq \beta$ by x_t , meaning there exist sets $\mathfrak{U} \subset [1 : U], \mathfrak{K} \subset [1 : K]$ such that $|\mathfrak{U}| + |\mathfrak{K}| = d$, and $\tilde{A}(\mathfrak{U})x_t = \alpha(\mathfrak{U}), B(\mathfrak{K})x_t = \beta(\mathfrak{K})$. As per the definition, $x_t \in \tilde{X}_t^I$, establishing the claim. \square

To complete the proof, we need to validate the preceding Lemma. It's worth noting that the statement above, while intuitively clear, seems somewhat intricate to prove, as suggested by the following argument. Although the statement extends to the OFUL algorithm, to the best of our knowledge, a direct argument for this has not been presented previously. Typically, when dealing with polytopal domains, it is directly asserted that playing on the extreme points of the polytope is sufficient.

Proof of Lemma B.2.1. Suppose $x_t \in \mathcal{P}$. Due to the optimistic choice, there exists $\tilde{\theta}_t \in \mathcal{C}_t^0$ such that

$$(\tilde{\theta}_t, x_t) \in \arg \max_{\tilde{\theta} \in \mathcal{C}_t^0, x \in \mathcal{P}} \langle \tilde{\theta}, x \rangle.$$

It's important to note that x_t is also a solution to the linear program $\max_{x \in \mathcal{P}} \langle \tilde{\theta}_t, x \rangle$, placing it on the boundary of \mathcal{P} . Similarly, $\tilde{\theta}_t$ lies on the boundary of \mathcal{C}_t^0 . It remains to argue that x_t must be an extreme point of \mathcal{P} ; in other words, it should not lie in the interior of any face of dimension ≥ 1 of \mathcal{P} .

For this, assume for the sake of contradiction that x_t lies in the interior of some 1-dimensional face of \mathcal{P} , denoted as \mathcal{F} . Let u represent the direction of variation of \mathcal{F} . Then, it must satisfy $\langle \tilde{\theta}_t, u \rangle = 0$, or else $\langle \tilde{\theta}_t, x_t + \varepsilon u \rangle$ would exceed $\langle \tilde{\theta}_t, x_t \rangle$ for some small choice of ε . Now, let's rotate the domain such that u aligns with one coordinate axis and project onto the 2D subspace spanned by the orthogonal directions u and $\tilde{\theta}_t$. After rescaling both u and $\tilde{\theta}_t$ to have a norm of 1, and translating the polytope so that the u th component of x_t is 0, the projection of an ellipsoid results in an ellipsoid. Performing the same transformations to \mathcal{C}_{t-1}^0 produces a 2-dimensional convex confidence ellipsoid D .

Let's relabel the axes as u_1 and u_2 . In this coordinate system, $\tilde{\theta} = (0, 1)$, and \mathcal{F} is a line segment of the form $\{u_1 \in [p, q], u_2 = r\}$, where $p < 0 < q, r = \langle \tilde{\theta}_t, x_t \rangle / \|\tilde{\theta}_t\|$,

and $x_t = (0, r)$. Observe that $\tilde{\theta}_t$ must lie on the boundary of D . We aim to show that there exists another $z \in \mathcal{F}$ and another $\phi \in D$ such that $\langle z, \phi \rangle > r$, leading to a contradiction.

First, consider the case where $r > 0$. If any point of D has a u_2 coordinate greater than 1, it leads to a contradiction. This is because for such a point ϕ , $\langle \phi, x_t \rangle > \langle \tilde{\theta}_t, x_t \rangle$. As $\tilde{\theta}_t = (0, 1) \in D$, the ellipse D is tangent to $u_2 = 1$. Thus, for small ε , D must contain points $\phi_\varepsilon = (\varepsilon, 1 - f(\varepsilon))$ where $0 \leq f(\varepsilon) = O(\varepsilon^2)$. However, this implies a contradiction - for any $\varepsilon > 0$, consider $z_\varepsilon = (\varepsilon^{1/2}, r)$. Then $z_\varepsilon \in \mathcal{F}$ for sufficiently small ε , and $\langle z_\varepsilon, \phi_\varepsilon \rangle - r = \varepsilon^{3/2} - rf(\varepsilon)$. Since $f(\varepsilon) = O(\varepsilon^2)$, this is positive for small enough ε , demonstrating a contradiction.

If $r < 0$, the same argument can be applied mutatis mutandis. Now, D must lie above the line $u_2 = 1$, and it must be tangent to it. Points of the form $(\varepsilon, 1 + f(\varepsilon))$ for $0 \leq f = O(\varepsilon^2)$ in D can be obtained, and the analogous inner product $\langle z_\varepsilon, \phi_\varepsilon \rangle - r = \varepsilon + rf(\varepsilon)$ is again positive for small enough ε . This leads to a contradiction.

Finally, the case $r = 0$, wherein x_t lies at the origin, is considered. But in this case, any point in D with a non-zero u_1 coordinate serves as a contradiction (since either $(p, 0)$ or $(0, q)$ will yield a positive inner product).

The combination of the preceding cases implies a crucial point: that the action x_t cannot be situated within the interior of an edge of the set \mathcal{P} . However, this argument extends its applicability to the interior of any non-trivial face within \mathcal{P} . Given that $\tilde{\theta}_t$ must be orthogonal to the affine subspace formed by such a face, we can deduce that there exists a point within the interior of a 1-D face, which constitutes a boundary of the larger face. This point within the 1-D face must also achieve the optimal value for $\langle \tilde{\theta}, x \rangle$. We can then apply the same argument as before to this point. Consequently, we can conclude that x_t cannot be located within the interior of any non-trivial face of the set \mathcal{P} . \square

The aforementioned reasoning is not confined to confidence ellipsoids as presented in §4.2.1; rather, it extends to any \mathcal{C}_t featuring a smooth and convex boundary. Furthermore, this extension applies to convex \mathcal{C}_t with continuous boundaries, except in instances where $\tilde{\theta}_t$ itself constitutes the extreme point of a polytope, characterized by substantial curvature at $\tilde{\theta}_t$. In such situations, the property $f(\varepsilon) = O(\varepsilon^2)$ does not hold, necessitating a more comprehensive argument. One possible approach could

involve employing continuous noise, wherein the confidence sets are almost surely unlikely to generate extreme points orthogonal to the faces of a polytope. This is because such directions lie in a union of a finite number of dimension $d - 1$ affine subspaces, which, in turn, is Lebesgue null. Consequently, we can almost surely avoid this disadvantageous scenario.

Additionally, Lemma B.2.1 leads to an intriguing observation that further characterizes the behavior of doubly-optimistic play.

Lemma B.2.2. *Assuming the validity of all confidence sets, there exists at least one BIS $I = (\mathfrak{U}, \mathfrak{R})$ that x_t noisily activates, satisfying $\begin{pmatrix} A(\mathfrak{U}) \\ B(\mathfrak{R}) \end{pmatrix} x_t \geq \begin{pmatrix} \alpha(\mathfrak{U}) \\ \beta(\mathfrak{R}) \end{pmatrix}$.*

In essence, for at least one BIS, the action x_t not only noisily activates it but also either activates it or violates all of its true constraints. It's worth noting that if the identified BIS contains at least one unknown constraint, this implies that DOSLB must potentially violate safety, as satisfying this constraint with equality would be infrequent.

Proof of Lemma B.2.2. Consider a fixed x_t . We refer to $\tilde{A} \in \mathcal{C}_t$ as a witness for x_t if $\tilde{A}x_t \leq \alpha$, signifying that \tilde{A} attests to the presence of x_t in $\tilde{\mathcal{S}}_t$. Since x_t represents the optimistic optimum across $\tilde{\mathcal{S}}_t$, it implies that for every witness \tilde{A} of x_t , it maximizes $\max_{\tilde{\theta} \in \mathcal{C}_t^0} \langle \tilde{\theta}, x \rangle : \tilde{A}x \leq \alpha, Bx \leq \beta$.

Now, let $\mathfrak{U}_0 \subset [1 : U]$ denote all unknown constraints that x_t noisily activates, and $\mathfrak{R}_0 \subset [1 : K]$ encompass all known constraints activated by x_t . Let $k_0 = |\mathfrak{R}_0|$. Additionally, define $\mathfrak{U}_{\geq} := \{i \in \mathfrak{U}_0 : \langle a^i, x_t \rangle \geq \alpha^i\}$. We assert that $|\mathfrak{U}_{\geq}| \geq d - k_0$, which is sufficient to establish the claim.

For the sake of contradiction, assume $|\mathfrak{U}_{\geq}| \leq d - k_0 - 1$. For each $i \in \mathfrak{U}_0 \setminus \mathfrak{U}_{\geq}$, we have $\langle a^i, x_t \rangle < \alpha^i$. However, note that the matrix $\tilde{A}_{<}$ —formed by replacing the \tilde{a}^i in each row corresponding to $i \in \mathfrak{U}_0 \setminus \mathfrak{U}_{\geq}$ with a^i —remains a witness for x_t . This replacement operation is valid due to the consistency of the confidence sets, ensuring $a^i \in \mathcal{C}_t^i$ for each i .

Moreover, x_t lies in the interior of the polytope $\mathcal{P}_{<} := \{x : \tilde{A}_{<}x \leq \alpha, Bx \leq \beta\}$: it activates precisely k_0 known constraints and at most $|\mathfrak{U}_{\geq}| \leq d - k_0 + 1$ unknown constraints. In total, it activates at most $d - 1 < d$ constraints of $\mathcal{P}_{<}$. Given that

$\tilde{A}_< \in \mathcal{C}_t$ and $\mathcal{P}_< \subset \tilde{\mathcal{S}}_t$, the algorithm plays x_t in the interior point of a polytope contained in the permissible set, contradicting Lemma B.2.1. Therefore, the hypothesis is untenable, and $|\mathfrak{U}_\geq| + |\mathfrak{K}_0| \geq d$. \square

B.3 Controlling the Play of Suboptimal BISs

Here, we establish the lower bound on noise scale and subsequently demonstrate control over the play of suboptimal BISs, as outlined in §4.6.

B.3.1 Localizing Actions when a BIS is Activated

We prove Lemma 4.5.1 as a straightforward consequence of consistency and optimism.

Proof of Lemma 4.5.1. Suppose the confidence sets are consistent, and x_t noisily activates the BIS I . Given that x_t is the action of DOSLB, it is also permissible. These properties collectively imply the existence of $\tilde{A} \in \mathcal{C}_t$ satisfying the following inequalities:

$$\begin{aligned} \tilde{A}x_t &\leq \alpha, & Bx_t &\leq \beta, \\ \tilde{A}(\mathfrak{U})x_t &= \alpha(\mathfrak{U}), & B(\mathfrak{K})x_t &= \beta(\mathfrak{K}). \end{aligned}$$

As \mathcal{C}_t is consistent, applying Lemma 4.5.2 in matrix form yields:

$$Ax_t - \rho_t \mathbf{1} \leq \tilde{A}x_t \leq Ax_t + \rho_t \mathbf{1}.$$

The claim follows directly:

$$\alpha \geq \tilde{A}x_t \geq Ax_t - \rho_t \mathbf{1} \implies Ax_t \leq \alpha + \rho_t \mathbf{1},$$

and

$$\alpha(\mathfrak{U}) = \tilde{A}(\mathfrak{U})x_t \leq A(\mathfrak{U})x_t + \rho_t \mathbf{1}(\mathfrak{U}) \implies A(\mathfrak{U})x_t \geq \alpha(\mathfrak{U}) - \rho_t \mathbf{1}(\mathfrak{U}).$$

Moreover, owing to the optimism of x_t , it acts as a maximizer among the permissible set in $\max_{\tilde{\theta} \in \mathcal{C}_t^0} \langle \tilde{\theta}, x \rangle$. Under consistency, $\theta \in \mathcal{C}_t^0$, and $x^* \in \tilde{\mathcal{S}}_t$. Consequently, if $\tilde{\theta}$ is the

optimal choice in the aforementioned program, then

$$\langle \tilde{\theta}, x_t \rangle \geq \langle \theta, x^* \rangle.$$

Utilizing consistency and Lemma 4.5.2 once again, we have $\langle \tilde{\theta}, x_t \rangle \leq \langle \theta, x_t \rangle + \rho_t$, from which the claim is derived. \square

B.3.2 Proof of Noise Scale Lower Bound and the Finiteness of Spread

We start from proving Lemma 4.6.1, essentially reiterating the argument from Example 4.5.3, accounting for the extended definition of the feasibility gap.

Proof of Lemma 4.6.1. Observe that under the consistency assumption,

$$\langle \theta, x_t \rangle \leq P(\rho_t; I),$$

since $x_t \in \mathcal{T}(\rho_t; I)$, and that

$$\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \rho_t,$$

since x_t is optimistic.

Given that $x_t \in \mathcal{T}(\rho_t; I)$, this set is nonempty, implying $\rho_t \geq \zeta_*(I)$. If $\zeta_*(I) = \infty$, the claim naturally follows. Moreover, if $\zeta_*(I) < \infty$, according to the spread definition, we have

$$P(\rho_t; I) \leq P(\zeta_*(I); I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)) = \langle \theta, x^* \rangle - \xi(I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)).$$

Consequently,

$$-\rho_t \leq -\xi(I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)) \iff \rho_t(\mathfrak{s}(I) + 1) \geq \xi(I) + \mathfrak{s}(I)\zeta_*(I) \iff \rho_t \geq \eta_*(I).$$

\square

Next, we establish Proposition 4.5.7 using an argument based on linear programming duality.

Proof of Proposition 4.5.7. Consider the fixed index set $I = (\mathfrak{U}, \mathfrak{R})$, and assume

$\zeta_*(I) < \infty$. By expanding the definition of $\mathcal{T}(\zeta; I)$, the program P becomes

$$\begin{aligned} P(\zeta; I) &= \max_x \langle \theta, x \rangle \\ \text{s.t. } Ax &\leq \alpha + \zeta \mathbf{1} \\ -A(\mathfrak{U})x &\leq -\alpha(\mathfrak{U}) + \zeta \mathbf{1}(\mathfrak{U}) \\ B(\mathfrak{K})x &= \beta(\mathfrak{K}) \\ B(\mathfrak{K}^c)x &\leq \beta(\mathfrak{K}^c). \end{aligned}$$

Here, $\mathfrak{K}^c = [1 : K] \setminus \mathfrak{K}$, and the known constraints in \mathfrak{K} are already satisfied with equality, hence excluded from the final line. This forms a linear program.

Since $\zeta_*(I) < \infty$, the above program is feasible for $\zeta \geq \zeta_*(I)$. Furthermore, as $\mathcal{X} = \{x : Bx \leq \beta\} \supset \mathcal{T}(\zeta; I)$ for any ζ , the program is finite. Thus, strong duality applies.

Now, introduce dual variables $(\lambda_+, \lambda_-, \mu, \nu)$ for the four block constraints. Using standard techniques, the dual program is given by

$$\begin{aligned} D(\zeta; I) &= \min_{\lambda_+, \lambda_-, \mu, \nu} \langle \lambda_+, \alpha + \zeta \mathbf{1} \rangle + \langle \lambda_-, -\alpha(\mathfrak{U}) + \zeta \mathbf{1}(\mathfrak{U}) \rangle + \langle \beta(\mathfrak{K}), \mu \rangle + \langle \beta(\mathfrak{K}^c), \nu \rangle \\ \text{s.t. } A^\top \lambda_+ - A(\mathfrak{U})^\top \lambda_- + B(\mathfrak{K})^\top \mu + B(\mathfrak{K}^c)^\top \nu &= \theta, \\ \lambda_+ \geq 0, \lambda_- \geq 0, \nu \geq 0. \end{aligned}$$

For brevity, take the following notations: $f(\lambda_+, \lambda_-, \mu, \nu) := \langle \lambda_+, \mathbf{1} \rangle + \langle \lambda_-, \mathbf{1}(\mathfrak{U}) \rangle$, $g(\lambda_+, \lambda_-, \mu, \nu) := \langle \lambda_+, \alpha + \zeta_*(I) \mathbf{1} \rangle + \langle \lambda_-, -\alpha(\mathfrak{U}) + \zeta_*(I) \mathbf{1}(\mathfrak{U}) \rangle + \langle \beta(\mathfrak{K}), \mu \rangle + \langle \beta(\mathfrak{K}^c), \nu \rangle$, and $h(\lambda_+, \lambda_-, \mu, \nu) := A^\top \lambda_+ - A(\mathfrak{U})^\top \lambda_- + B(\mathfrak{K})^\top \mu + B(\mathfrak{K}^c)^\top \nu - \theta$.

Define $\boldsymbol{\lambda} = (\lambda_+^\top, \lambda_-^\top, \mu^\top, \nu^\top)^\top$. We can succinctly express the dual as

$$D(\zeta; I) = \min_{\boldsymbol{\lambda}} (\zeta - \zeta_*(I)) f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : h(\boldsymbol{\lambda}) = 0, \lambda_+ \geq 0, \lambda_- \geq 0, \nu \geq 0.$$

Note that since the primal is bounded and feasible for $\zeta \geq \zeta_*(I)$, so is the dual, and by strong duality $D(\zeta_*(I); I) = P(\zeta_*(I); I)$. But

$$D(\zeta_*(I); I) = \min_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) : h(\boldsymbol{\lambda}) = 0, \lambda_+ \geq 0, \lambda_- \geq 0, \nu \geq 0.$$

It follows that the set

$$\mathcal{F} := \{\boldsymbol{\lambda} : g(\boldsymbol{\lambda}) \leq P(\zeta_*(I); I), h(\boldsymbol{\lambda}) = 0, \lambda_+ \geq 0, \lambda_- \geq 0, \nu \geq 0\}$$

is nonempty. Observe that ζ does not appear anywhere in the definition of \mathcal{F} .

Let us define the two programs

$$\begin{aligned} D'(\zeta; I) &:= \min_{\boldsymbol{\lambda}} (\zeta - \zeta_*(I))f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}, \\ E(I) &:= \min_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F} \end{aligned}$$

Note that both of the above programs are feasible. As a feasible minimisation program, we also have that $E(I) < \infty$. Further, since introducing extra constraints cannot decrease the value of a program, we note that $D(\zeta; I) \leq D'(\zeta; I)$. But observe that since the constraints of $D'(\zeta; I)$ include the requirement that $g(\boldsymbol{\lambda}) \leq P(\zeta_*(I); I)$, we have for every $\zeta \geq \zeta_*(I)$ that

$$\begin{aligned} D'(\zeta; I) &\leq P(\zeta_*(I); I) + \min\{(\zeta - \zeta_*(I))f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}\} \\ &= P(\zeta_*(I); I) + (\zeta - \zeta_*(I)) \cdot \min\{f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}\} \\ &= P(\zeta_*(I); I) + (\zeta - \zeta_*(I))E(I). \end{aligned}$$

Then by strong duality,

$$P(\zeta; I) = D(\zeta; I) \leq P(\zeta_*(I); I) + (\zeta - \zeta_*(I))E(I),$$

and we conclude that $\mathfrak{s}(I) \leq \max(0, E(I)) < \infty$.

Now, since $\mathfrak{s}(I)$ is finite, in order to show that $\max(\zeta_*(I), \eta_*(I)) > 0$, it suffices to argue that for any suboptimal BIS, $\max(\zeta_*(I), \xi(I)) > 0$. But observe that if $\zeta_*(I) = 0$, then $\lim_{\zeta \searrow 0} P(\zeta; I) > -\infty$, and due to the right-continuity of P , this implies that $P(0; I) > -\infty \implies \mathcal{X}^I \neq \emptyset$, in other words, I is a feasible BIS. But if a BIS I is both feasible and suboptimal, then for every $x \in I$, it must hold that $\langle \theta, x \rangle < \langle \theta, x^* \rangle$, since otherwise I would be optimal. But, since $\mathcal{X}^I = \mathcal{T}(0; I)$ is a compact set, this means that $P(\zeta_*(I); I) = P(0; I) < \langle \theta, x^* \rangle \iff \xi(I) > 0$. \square

B.3.3 Limiting the Occurrence of Suboptimal BISs

Having established the necessary components, we proceed to prove the central result of §4.5, 4.6.

Proof of Theorem 4.6.2. Let us once again denote $\rho_t(x_t; \delta)$ as ρ_t . According to Lemma 4.6.1, playing a suboptimal BIS I ensures that $\rho_t \geq \max(\eta_*(I), \zeta_*(I))$. Conse-

quently, whenever a suboptimal BIS is played, we have $\rho_t \geq \min\{\max(\eta_*(I), \zeta_*(I)) : I \text{ is a suboptimal BIS}\}$, denoted as $\rho_t \geq \Xi$.

Now observe that:

$$\begin{aligned} \sum_{t=1}^T \mathbb{1}\{\exists \text{ suboptimal BIS } I : x_t \in \tilde{\mathcal{X}}_t^I\} &\leq \sum_{t=1}^T \mathbb{1}\{\rho_t \geq \Xi\} \\ &\leq \sum_{t=1}^T \frac{\rho_t^2}{\Xi^2} \mathbb{1}\{\rho_t \geq \Xi\} \\ &\leq \Xi^{-2} \sum_{t \leq T} \rho_t^2, \end{aligned}$$

where the second inequality is based on the fact that if $\rho_t \geq \Xi$, then $\rho_t/\Xi \geq 1$. As long as $\lambda = \Theta(1)$, applying Lemma B.1.2 allows us to bound the above expression as $O\left(\frac{d^2 \log^2 T + d \log(T) \log(U/\delta)}{\Xi^2}\right)$. \square

B.4 Proofs of Bounds on Efficiency Regret and Safety Violations

Before starting the proofs, it is essential to address the non-degeneracy assumption 4.6.3 initially. This assumption provides ample leeway for degeneracy, particularly at x^* , where numerous constraints might intersect, yet no other point in \mathcal{S} does. While such non-degeneracy conditions are common in linear programming, they can be mitigated by slightly perturbing the constraint matrix. Similarly, the noise genericity condition is standard and can be fulfilled by introducing a small independent noise to the feedback. The primary utility of this assumption lies in the subsequent result, the first part of which is elucidated in §B.4.1 through a meticulous analysis of the optimistic selection rule (4.2). This rule characterizes the structure of the noisy matrix \tilde{A} implicitly chosen by the algorithm when selecting x_t .

Now, let's proceed the proofs of the results from §4.6.

B.4.1 The Efficiency of the Actions of DOSLB when Activating Optimal BISs

Our initial objective is to demonstrate that exclusively playing optimal BISs results in actions x_t that are excessively efficient, satisfying $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$. The subsequent fundamental result is instrumental in our argument.

Lemma B.4.1. *Let $I = (\mathcal{U}, \mathcal{R})$ be any BIS such that the matrix $B(\mathcal{R})$ is full row rank. Then under the genericity of noise, for any $t \geq d$, it holds that the matrix $\begin{pmatrix} \hat{A}_t(\mathcal{U}) \\ B(\mathcal{R}) \end{pmatrix}$ is almost surely full rank.*

Proof of Lemma B.4.1. Notice that since, for any i , the noise in the feedback S_t^i is generic, it does not concentrate in any low-dimensional subspace of \mathbb{R}^d . This, in turn, means that the probability that any \hat{a}_t^i lies in a low-dimensional subspace of \mathbb{R}^d is exactly zero. The claim follows immediately: since $|\mathcal{U}| \leq d$, each \hat{a}_t^i with probability one does not lie in the span of $\{\hat{a}_t^j\}_{j \in \mathcal{U} \setminus \{i\}}$, and also does not lie in the span of $\{b^j\}_{j \in \mathcal{R}}$. Further, by assumption, the $\{b^j\}$ are linearly independent. \square

With this in hand, we argue Lemma 4.6.4 by exploiting the weak-nondegeneracy condition of Assumption 4.6.3.

Proof of Lemma 4.6.4. Proof of near-safety with small noise scale. Let us begin with the second part, i.e., if $\rho_t(x_t, \delta) < \varepsilon$, then x_t is at most ε -unsafe. This follows directly from the noise scale bound of Lemma 4.5.2: indeed, under consistency of the confidence sets, we observe that for every $i \in [1 : U]$ and every x , $\max_i \max_{\tilde{a}^i \in \mathcal{C}_t^i} |\langle a^i, x \rangle - \langle \tilde{a}^i, x \rangle| \leq \rho_t(x; \delta)$. But, if x_t is played, then $x_t \in \tilde{\mathcal{S}}_t$ and thus for each i , there exists a choice of \tilde{a}^i such that $\langle \tilde{a}^i, x_t \rangle \leq \alpha^i \implies \langle a^i, x_t \rangle \leq \alpha^i + \rho_t(x_t; \delta)$.

Proof of the ‘over-efficiency’ of x_t under optimal BIS activation. We now come to the first part, and argue that if *all* of the BISs x_t noisily activates are optimal, then $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$, which comprises the bulk of this proof. To this end, let us fix one such BIS, $I = (\mathcal{U}, \mathcal{R})$.

By Assumption 4.6.3, we know that $\{x^*\} = \mathcal{X}^I$, and that I is full-rank. Notice that as a result, we may write

$$\langle \theta, x^* \rangle = \max \langle \theta, x \rangle : A(\mathcal{U})x = \alpha(\mathcal{U}), B(\mathcal{R})x = \beta(\mathcal{R}).$$

Indeed, due to the fact that I is full rank, the latter equality constraints already enforce that x^* is the sole feasible point. Further, by strong duality, there exists a choice of vectors μ, ν such that

$$\mu^\top A(\mathfrak{U}) + \nu^\top B(\mathfrak{R}) = \theta^\top.$$

Due to the optimistic selection rule and the fact that x_t noisily saturates I , it must satisfy the following optimization problem:

$$\max_{\tilde{\theta} \in \mathcal{C}_t^0, \tilde{A} \in \mathcal{C}_t} \max_x \langle \tilde{\theta}, x \rangle : \tilde{A}(\mathfrak{U})x = \alpha(\mathfrak{U}), B(\mathfrak{R})x = \beta(\mathfrak{R}), \tilde{A}x \leq \alpha, Bx \leq \beta.$$

Now, observe that in the optimization above, we may restrict attention to \tilde{A} such that $\tilde{M}(I; \tilde{A}) := \begin{pmatrix} \tilde{A}(\mathfrak{U}) \\ B(\mathfrak{R}) \end{pmatrix}$ is full rank. If the optimal choice were rank-deficient, there must exist other constraints among \tilde{A}, B besides those in I that are activated by x_t (violating Lemma B.2.1). By dropping some linearly dependent rows, this would yield a different index set I' that x_t activates, which is not rank-deficient. By the hypothesis, this index set must also be optimal, and we can run the argument for I' instead. Then, x_t is exactly characterized by the equality conditions imposed by noisily activating the BIS I , meaning that x_t is the optimizer of:

$$\max_{\substack{\tilde{\theta} \in \mathcal{C}_t^0, \tilde{A} \in \mathcal{C}_t, \\ \tilde{M}(I, \tilde{A}) \text{ is full-rank}}} \max_x \langle \tilde{\theta}, x \rangle : \tilde{A}(\mathfrak{U})x = \alpha(\mathfrak{U}), B(\mathfrak{R})x = \beta(\mathfrak{R}).$$

Now, let us write $\tilde{A} = A + \delta A, \tilde{\theta} = \theta + \delta\theta, x = x^* + \delta x$. Further denote the optima as $\delta\theta_t, \delta A_t, \delta x_t$. With this notation, our goal is to show that $\langle \theta, \delta x_t \rangle \geq 0$. To this end, observe that since the program above has the constraint $\tilde{A}(\mathfrak{U})x = \alpha(\mathfrak{U}) = A(\mathfrak{U})x^*, B(\mathfrak{R})x = \beta = B(\mathfrak{R})x^*$, we find that:

$$\begin{aligned} \tilde{A}(\mathfrak{U})x &= A(\mathfrak{U})x^* + \delta A(\mathfrak{U})x + A(\mathfrak{U})\delta x = \alpha \iff A(\mathfrak{U})\delta x = -\delta A(\mathfrak{U})x, \\ B(\mathfrak{R})x &= B(\mathfrak{R})x^* + B(\mathfrak{R})\delta x = B(\mathfrak{R})x^* \iff B(\mathfrak{R})\delta x = 0. \end{aligned}$$

This implies:

$$\begin{aligned}
\langle \theta, \delta x \rangle &= \langle A^\top \mu + B^\top \nu, \delta x \rangle = \langle \mu, A \delta x \rangle + \langle \nu, B \delta x \rangle = -\langle \mu, \delta A x \rangle + 0 \\
\iff \langle \theta, \delta x \rangle &= \sum_{i \in \mathfrak{U}} -\mu^i \langle \delta A^i, x \rangle
\end{aligned} \tag{B.3}$$

Thus, we can rewrite the program as:

$$\max_x \max_{\delta \theta, \delta A} \langle \theta, x^* \rangle + \langle \delta \theta, x \rangle - \langle \mu, \delta A(\mathfrak{U})x \rangle : \tilde{A}(\mathfrak{U})x = \alpha(\mathfrak{U}), B(\mathfrak{K})x = \beta(\mathfrak{K}).$$

Now, recall that the confidence sets are constructed around the RLS estimates \hat{a}_t^i and $\hat{\theta}_t$, i.e.,

$$\mathcal{C}_t^0 = \{\tilde{\theta} : \|\tilde{\theta} - \hat{\theta}_t\|_{V_t} \leq \omega_t\}, \mathcal{C}_t^i = \{\tilde{a} : \|\tilde{a} - \hat{a}_t^i\|_{V_t} \leq \omega_t\}.$$

To clearly express the choice of $\delta \theta, \delta A$, we define:

$$\begin{aligned}
\Delta \theta_t &= \hat{\theta}_t - \theta, \Delta a_t^i = \hat{a}_t^i - a^i, \Delta A = \hat{A}_t - A \\
\partial \theta &= \tilde{\theta} - \hat{\theta}_t, \partial a^i = \tilde{a}^i - \hat{a}_t^i, \partial A = \tilde{A} - \hat{A}_t.
\end{aligned}$$

Observe then that:

$$\delta \theta = \Delta \theta_t + \partial \theta; \delta a^i = \Delta a_t^i + \partial a^i.$$

Further, the decision variables of the program are only the $\partial \theta$ and ∂a^i s, which lie in the set $\|\partial * \|_{V_t} \leq \omega_t$. Incorporating this structure, we can write the program as:

$$\begin{aligned}
&\langle \theta, x^* \rangle + \max_x \max_{\sigma^i} \langle \Delta \theta_t, x \rangle - \sum_{i \in \mathfrak{U}} \mu^i \langle \Delta a_t^i, x \rangle + \omega_t \|x\|_{V_t^{-1}} - \sum_{i \in \mathfrak{U}} \mu^i \sigma^i \omega_t \|x\|_{V_t^{-1}}^2. \\
&\text{s.t.} \quad \langle a^i + \Delta a_t^i, x \rangle = \alpha^i - \omega_t \sigma^i \|x\|_{V_t^{-1}}^2 \quad \forall i \in \mathfrak{U}, \\
&\quad B(\mathfrak{K})x = \beta, \\
&\quad (\sigma^i)^2 \|x\|_{V_t^{-1}}^2 \leq 1 \quad \forall i \in \mathfrak{U}.
\end{aligned}$$

But now observe that the optimal choice of $\partial \theta$ in the above is exactly $\omega_t / \|x\|_{V_t^{-1}} V_t^{-1} x$. Indeed, recall that $\|u\|_{V_t} = \sqrt{u^\top V_t u} = \|V_t^{1/2} u\|$, and similarly $\|u\|_{V_t^{-1}} = \|V_t^{-1/2} u\|$. By the Cauchy-Schwarz inequality, $\langle \partial \theta, x \rangle =$

$\langle V_t^{1/2} \partial \theta, V_t^{-1/2} x \rangle \leq \|\partial \theta\|_{V_t} \|x\|_{V_t^{-1}}$, and this is extremized when $V_t^{1/2} \partial \theta \propto V_t^{-1/2} x \iff \partial \theta \propto V_t^{-1} x$.

Further notice that the optimal choice of ∂a^i must similarly be aligned with $V_t^{-1} x$. Write $V_t^{1/2} \partial a^i = \omega_t \sigma^i V_t^{-1/2} x + \psi^i$, where σ^i is a scalar, and ψ^i is a vector such that $\langle \psi^i, V_t^{-1/2} x \rangle = 0$. Then observe that due to the orthogonality:

$$\begin{aligned} \|\partial a^i\|_{V_t}^2 &= \langle V_t^{1/2} \partial a^i, V_t^{1/2} \partial a^i \rangle = \langle \omega_t \sigma^i V_t^{-1/2} x + \psi^i, \omega_t \sigma^i V_t^{-1/2} x + \psi^i \rangle \\ &= \omega_t^2 (\sigma^i)^2 \|x\|_{V_t^{-1}}^2 + \|\psi^i\|^2, \end{aligned}$$

This means that dumping any energy into ψ^i affects neither the first constraint on $\langle a^i + \Delta a_t^i + \partial a^i, x \rangle$, nor the objective, since:

$$\langle \partial a^i, x \rangle = \langle V_t^{1/2} \partial a^i, V_t^{-1/2} x \rangle = \langle \omega_t \sigma^i V_t^{-1/2} x, V_t^{-1/2} x \rangle + \langle \psi^i, x \rangle = \sigma^i \|x\|_{V_t^{-1}}^2.$$

This means that dumping any energy into ψ^i affects neither the constraints nor the objective, so we can safely set it to zero in the following (in fact, as we shall see below, it must be zero since σ^i must saturate). This allows us to considerably simplify the above program: introducing σ^i as above, the program can be rewritten as:

$$\begin{aligned} \langle \theta, x^* \rangle + \max_x \max_{\{\sigma^i\}} \langle \Delta \theta_t, x \rangle - \sum_{i \in \mathfrak{U}} \mu^i \langle \Delta a_t^i, x \rangle + \omega_t \|x\|_{V_t^{-1}} - \sum_{i \in \mathfrak{U}} \mu^i \sigma^i \omega_t \|x\|_{V_t^{-1}}^2 \\ \text{s.t.} \quad \langle a^i + \Delta a_t^i, x \rangle = \alpha(\mathfrak{U}) - \omega_t \sigma^i \|x\|_{V_t^{-1}}^2 \quad \forall i \in \mathfrak{U}, \\ B(\mathfrak{R})x = \beta, \\ (\sigma^i)^2 \|x\|_{V_t^{-1}}^2 \leq 1 \quad \forall i \in \mathfrak{U}. \end{aligned}$$

Now, observe that the first constraint can be succinctly written in terms of $\hat{A}_t(\mathfrak{U})$, giving us the following restatement, where σ is the vector formed by stacking the σ^i s:

$$\begin{aligned} \langle \theta, x^* \rangle + \max_x \max_{\sigma} \langle \Delta \theta_t, x \rangle - \sum_{i \in \mathfrak{U}} \mu^i \langle \Delta a_t^i, x \rangle + \omega_t \|x\|_{V_t^{-1}} - \langle \mu, \sigma \rangle \omega_t \|x\|_{V_t^{-1}}^2 \\ \text{s.t.} \quad \hat{A}_t(\mathfrak{U})x = \alpha(\mathfrak{U}) - \omega_t \sigma \|x\|_{V_t^{-1}}^2 \\ B(\mathfrak{R})x = \beta, \\ (\sigma^i)^2 \|x\|_{V_t^{-1}}^2 \leq 1 \quad \forall i \in \mathfrak{U}. \end{aligned}$$

However, observe that $B(\mathfrak{R})$ is assumed to be full row rank. Consequently, applying

Lemma B.4.1, it follows with probability one that $\begin{pmatrix} \hat{A}_t(\mathfrak{U}) \\ B(\mathfrak{R}) \end{pmatrix}$ is full-rank. This implies that every value of σ satisfying the final constraint is feasible for the above program. Clearly, the optimal choice for σ^i is $-\text{sign}(\mu^i)/\|x\|_{V_t^{-1}}$, indicating that for each $i \in \mathfrak{U}$, the optimal ∂a^i at time t is:

$$\partial a_t^i = -\text{sign}(\mu^i)\omega_t V_t^{-1}x/\|x\|_{V_t^{-1}} \implies \mu^i \langle \partial a_t^i, x \rangle = \omega_t |\mu^i| \cdot \|x\|_{V_t^{-1}}.$$

Finally, we observe that for each $i \in \mathfrak{U}$, and every x , $\omega_t |\mu^i| \|x\|_{V_t^{-1}} - \mu^i \langle \Delta a_t^i, x \rangle \geq 0$. Given that the confidence sets are consistent ($a^i \in \mathcal{C}_t^i \iff \|\Delta a_t^i\|_{V_t} \leq \omega_t$), we have:

$$|\mu^i \langle \Delta a_t^i, x \rangle| = |\mu^i| \left\| \left\langle V_t^{1/2} \Delta a_t^i, V_t^{-1/2} x \right\rangle \right\| \leq |\mu^i| \|\Delta a_t^i\|_{V_t} \|x\|_{V_t^{-1}} \leq |\mu^i| \omega_t \|x\|_{V_t^{-1}}.$$

Using (B.3), we can now show:

$$\begin{aligned} \langle \theta, \delta x_t \rangle &= \sum_{i \in \mathfrak{U}} -\mu^i \langle \delta a_t^i, x_t \rangle \\ &= \sum_{i \in \mathfrak{U}} -\mu^i \langle \partial a_t^i, x_t \rangle - \mu^i \langle \Delta a_t^i, x_t \rangle \\ &= \sum_{i \in \mathfrak{U}} |\mu^i| \omega_t \|x_t\|_{V_t^{-1}} - \mu^i \langle \Delta a_t^i, x_t \rangle \geq 0. \end{aligned}$$

Thus, the proof is complete. \square

The non-degeneracy condition Assumption 4.6.3 plays a relatively weak role in the above argument. Essentially, we require that x_t noisily activates some index set such that the true θ can be expressed as a linear combination of the true constraint vectors of the index set. Without this condition, the proof does not hold as stated, since it could be the case that some constraints needed to express θ are not noisily activated by x_t , even if they are activated by x^* . This leads to a loss of equality among the various programs we defined, resulting only in a lower bound. It remains an open problem to precisely capture the freedom of optimistic play when it extends beyond the safe set, where it can activate any noisy constraints, complicating the construction of a $\delta\theta$ and δA that make the point suboptimal.

B.4.2 Proof of the Main Theorem

With all the components in position, we now proceed to establish our main claim.

Proof of Theorem 4.6.5. Assuming that with probability at least $1 - \delta$, all confidence sets are consistent, we proceed to argue the claim under this event.

Firstly, we divide the time horizon into two categories based on whether x_t noisily activates suboptimal BISs or not, defining

$$\mathfrak{T}_1 := \{t \in [d + 1 : T] : \exists \text{ a suboptimal BIS } I \text{ such that } x_t \in \tilde{\mathcal{X}}_t^I\}.$$

For $t \in [d + 1 : T] \setminus \mathfrak{T}_1$, x_t exclusively activates optimal BISs. Further categorization is done based on whether x_t is overly unsafe or not, via

$$\mathfrak{T}_2^\varepsilon := \{t \in [d + 1 : T] \setminus \mathfrak{T}_1 : \exists i : \langle a^i, x_t \rangle \geq \alpha^i + \varepsilon\},$$

where $\varepsilon > 0$ is arbitrary.

By Lemma 4.6.1, for all $t \in \mathfrak{T}_1$, $\rho_t(x_t; \delta) \geq \Xi$. According to the first part of Lemma 4.6.4, for every $t \in [d + 1 : T] \setminus \mathfrak{T}_1$, it holds that $\langle \theta, x^* - x_t \rangle \leq 0$. Further, by the second part of Lemma 4.6.4, for all $t \in \mathfrak{T}_2^\varepsilon$, $\rho_t(x_t; \delta) \geq \varepsilon$. Finally, for $t \in [d + 1 : T] \setminus (\mathfrak{T}_1 \cup \mathfrak{T}_2^\varepsilon)$, the actions x_t achieve both $\langle \theta, x^* - x_t \rangle \leq 0$, and $\max_i \langle a^i, x_t \rangle - \alpha^i - \varepsilon \leq 0$.

Now, it must hold that for all times

$$\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \rho_t(x_t; \delta).$$

This is evident since both θ and x^* are feasible choices for the actions of DOSLB. If some $\tilde{\theta}, \tilde{x}_t$ are chosen instead, then $\langle \tilde{\theta}, \tilde{x}_t \rangle \geq \langle \theta, x^* \rangle$. By Lemma 4.5.2, under consistency, $\langle \theta, x_t \rangle \geq \langle \tilde{\theta}, \tilde{x}_t \rangle - \rho_t(x_t; \delta)$, establishing the aforementioned claim.

Hence, we have the efficiency control:

$$\begin{aligned}
\mathcal{E}_T &= \sum_t \langle \theta, x^* - x_t \rangle_+ = \sum_{t \leq d} \langle \theta, x^* - x_t \rangle_+ + \sum_{t \in \mathfrak{I}_1} \langle \theta, x^* - x_t \rangle_+ + \sum_{t \notin \mathfrak{I}_1} \langle \theta, x^* - x_t \rangle_+ \\
&\leq d + \sum_{t \in \mathfrak{I}_1} \rho_t(x_t; \delta) + 0 \\
&\leq d + \sum_t \rho_t(x_t; \delta) \mathbb{1}\{\rho_t(x_t; \delta) \geq \Xi\} \\
&\leq d + \sum_t \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\Xi} \\
&= d + \frac{1}{\Xi} \sum_t \rho_t(x_t; \delta)^2,
\end{aligned}$$

leading to the claimed bound through the use of Lemma B.1.2.

The argument for controlling $\mathcal{S}_T^\varepsilon$ is nearly identical. We have:

$$\begin{aligned}
\mathcal{S}_T^\varepsilon &= \sum_{t \in \mathfrak{I}_1} \max_i (\langle a^i, x_t \rangle - \alpha^i - \varepsilon)_+ + \sum_{t \in \mathfrak{I}_2^\varepsilon} (\langle a^i, x_t \rangle - \alpha^i - \varepsilon)_+ \\
&\quad + \sum_{t \notin (\mathfrak{I}_1 \cup \mathfrak{I}_2^\varepsilon)} (\langle a^i, x_t \rangle - \alpha^i - \varepsilon)_+ \\
&\leq \sum_{t \in \mathfrak{I}_1} \rho_t(x_t; \delta) + \sum_{t \in \mathfrak{I}_2^\varepsilon} \rho_t(x_t; \delta) + 0 \\
&\leq \sum_t \rho_t(x_t; \delta) \mathbb{1}\{\rho_t(x_t; \delta) \geq \Xi\} + \sum_t \rho_t(x_t; \delta) \mathbb{1}\{\rho_t(x_t; \delta) \geq \varepsilon\} \\
&\leq \sum_t \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\Xi} + \sum_t \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\varepsilon} \\
&= \left(\frac{1}{\Xi} + \frac{1}{\varepsilon} \right) \sum_t \rho_t(x_t; \delta)^2,
\end{aligned}$$

and we are done again by invoking Lemma B.1.2. Note the simultaneous result over ε , which follows from the fact that $\varepsilon > 0$ was arbitrarily chosen in the above. \square

B.4.3 Proofs of Subsidiary Claims from §4.6

Finally, we present the proof of the subsidiary observation from §4.6.

Proof of Polynomial Bound on Violations.

It's noteworthy that since the result of Theorem 4.6.5 holds regardless of ε , we can express, for any $\varepsilon > 0$, that

$$\mathcal{S}_T \leq \varepsilon T + O\left(\frac{d^2 \log^2(T)}{\varepsilon}\right) + O\left(\frac{d^2 \log^2(T)}{\Xi}\right),$$

where we utilize the fact that $(u + v)_+ \leq u_+ + v_+$ to infer

$$\mathcal{S}_T = \sum \max_i (\langle a^i, x_t \rangle - \alpha^i)_+ = \sum \max_i (\langle a^i, x_t \rangle - \alpha^i - \varepsilon + \varepsilon)_+ \leq \mathcal{S}_T^\varepsilon + \varepsilon T.$$

The claim follows by choosing $\varepsilon = \Theta(d \log(T)/\sqrt{T})$.

It's worth noting that this claim can also be demonstrated directly. Indeed, leveraging the observation that x_t is permissible, it follows that in any round t , $\langle a^i, x_t \rangle - \alpha^i \leq \rho_t(x_t; \delta)$. Thus, we find that

$$\mathcal{S}_T \leq \sum_t \rho_t(x_t; \delta),$$

and the claim follows directly from Lemma B.1.2.

DOSLB Achieves at most Square-Root Regret in General.

This property is derived from the last observation in the previous section. Indeed, due to the optimism, it generally holds that $\mathcal{E}_T \leq \sum \rho_t(x_t; \delta)$ and $\mathcal{S}_T \leq \sum \rho_t(x_t; \delta)$. Note that these bounds apply more generally than our setting: for instance, instead of requiring the known constraints to form a finite polytope, the same result holds as long as the known constraints form a convex set. For our particular case, the significance of this result lies in the fact that we do not need any degeneracy condition for this to hold, making the lower bound of Theorem 4.6.7 effective for DOSLB.

Finite Action Setting.

A commonly assumed condition in linear bandits is that the learner is given a *finite set* of actions \mathcal{A} and must ensure $x_t \in \mathcal{A}$ (Abbasi-Yadkori et al., 2011, e.g.). This scenario considerably simplifies our analysis because resolving the optimal point is no longer a concern, enabling the entire analysis to be performed in the primal space. Specifically, let us modify DOSLB so that it makes its optimistic choice from $\tilde{\mathcal{S}}_t \cap \mathcal{A}$, and define the *finite-arm efficiency gap* of $x \in \mathcal{A}$ as $\Delta_x := \langle \theta, x^* - x \rangle_+$ and the *finite-action safety gap* of $x \in \mathcal{A}$ as $\Sigma_x := \max_i (\langle a^i, x \rangle - \alpha^i)_+$, and the gap of the problem as $\Gamma := \min_{x \in \mathcal{A}} \max(\Delta_x, \Sigma_x)$.

Proposition B.4.2. *With probability at least $1 - \delta$, the modified finite-action DOSLB achieves the following bounds in the finite-armed SLB setting: $\max(\mathcal{E}_T, \mathcal{S}_T) = O(d^2 \log^2 T / \Gamma)$.*

Notice that we do not need a precision relaxation in \mathcal{S}_T above because the precision issues arising from having to locate the optimal action are not present.

Let us specify the setting in a little more detail: we are supplied with a finite set $\mathcal{A} \subset \mathbb{R}^d$, and in each round, the learner chooses one action $x_t \in \mathcal{A}$. The linear reward and constraint structures remain identical, and x^* is updated to be the best action in \mathcal{A} , i.e.,

$$x^* := \arg \max \langle \theta, x \rangle : Ax \leq \alpha, x \in \mathcal{A}.$$

Note that the known constraints are no longer necessary: if they are given, we may filter \mathcal{A} before play starts. The gap $\Gamma := \min_{x \in \mathcal{A}} (\Delta_x, \Sigma_x)$ is non-zero simply because each suboptimal arm in \mathcal{A} must have at least one of Δ_x, Σ_x positive, and the minimization is over a finite set.

The result relies on the following observation, which follows straightforwardly from Lemma 4.5.2.

Lemma B.4.3. *If the confidence sets are consistent, and the modified finite-action version of DOSLB chooses $x_t \neq x_*$ from \mathcal{A} , then $\rho_t \geq \max(\Delta_{x_t}, \Sigma_{x_t})$.*

Proof of Lemma B.4.3. Notice that the basic result Lemma 4.5.2 remains valid in this setting. As a result, if the confidence sets are consistent, then since x_t is permissible, there exists $\tilde{A} \in \mathcal{C}_t$ such that $\tilde{A}x_t \leq \alpha$. Now suppose that $\langle a^i, x_t \rangle \geq \alpha^i + \Sigma_{x_t}$. Then the condition $|\langle \tilde{a}^i - a^i, x_t \rangle| \leq \rho_t$ of Lemma 4.5.2 implies that for every $\tilde{a}^i \in \mathcal{C}_t^i$,

$$\langle \tilde{a}^i, x_t \rangle \geq \langle a^i, x_t \rangle - \rho_t \geq \alpha^i + \Sigma_{x_t} - \rho_t.$$

Putting the above inequalities together yields $\rho_t \geq \Sigma_{x_t}$.

Similarly, much as in the proof of Lemma 4.5.1, since x_t is optimistically selected, and since the confidence sets are consistent, it holds that $\exists \tilde{\theta} : \langle \tilde{\theta}, x_t \rangle \geq \langle \theta, x^* \rangle$. But due to consistency, $\langle \tilde{\theta}, x_t \rangle \leq \langle \theta, x_t \rangle \leq \langle \theta, x^* \rangle - \Delta_{x_t}$, and so $\rho_t \geq \Delta_{x_t}$.

Putting these two lower bounds together yields the result. \square

Proof of Proposition B.4.2. This bound follows similarly to the previous control on \mathcal{E}_T and \mathcal{S}_T . Indeed, using Lemma B.4.3, observe that under the consistency of the confidence sets, if $\exists i : \langle a^i, x_t \rangle \geq \alpha^i$, then $\rho_t(x_t; \delta) \geq \Gamma$ and similarly if $\langle \theta, x^* - x_t \rangle > 0$ then $\rho_t(x_t; \delta) \geq \Gamma$. Exploiting this as in the proof of Theorem 4.6.5 yields the claim.

In fact, we can give a slightly more refined result. Let $\Delta := \min_x \Delta_x$, where recall that $\Delta_x = 0$ if $\langle \theta, x^* - x \rangle \leq 0$. Similarly define $\Sigma := \min_x \Sigma_x$. Then adapting the proof of Lemma B.4.3 slightly yields that in every round where the safety is violated, ρ_t must exceed Σ and in every round where efficiency is suboptimal, ρ_t must exceed Δ . This lets us give the separate control $\mathcal{E}_T = O(\log^2 T / \Delta)$ and $\mathcal{S}_T = O(\log^2 T / \Sigma)$ (this may or may not be a better bound: if for every action but the optimal, one of Δ_x and Σ_x is very small but the other is large, then the bound in terms of $\min_x \max(\Delta_x, \Sigma_x)$ is better.)

Finally, also observe that adapting the proof of Theorem 4.6.2 also tells us that the number of rounds in which an unsafe action was played in this finite-action setting is bounded as $O(\Gamma^{-2} \log^2 T)$. \square

B.5 Proofs of Lower Bounds

We will now establish the lower bounds that were asserted in the main text.

B.5.1 Proof of Polynomial Lower Bound

We substantiate Theorem 4.3.1 by elaborating on the example introduced in § 4.3. The proof employs techniques that are commonly used in the bandit literature (Lattimore and Szepesvári, 2020, Ch. 24).

Proof of Theorem 4.3.1. The instance we consider is

$$\mathcal{X} = [0, 1], \theta^* = 1, a^1 = (1 \pm \kappa)/2, \alpha^1 = 1/4, w_t^i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1), i \in \{0, 1\}$$

for some $\kappa \in (0, 1/4)$. Note that implicitly, the above has the known constraints $-x \leq 0$ and $x \leq 1$. Of course, this one-dimensional construction can be embedded into an arbitrary dimension (for instance, by taking a very skinny box domain, and only enforcing this single unknown constraint).

In the above case, the optimal feasible solutions are $x^+ = \frac{1}{2(1+\kappa)}$, $x^- = \frac{1}{2(1-\kappa)}$ for these two instances respectively. In addition, both of these two instances are at least $1/8$ -well separated. The key observation is the indistinguishability of these two instances with $\ll 1/\kappa^2$ actions.

Indeed, let $\mathbb{P}^+, \mathbb{P}^-$ be the distributions induced by the two problem instances and the learning algorithm. Since in either case, the noise distribution is a standard Gaussian, and the reward distributions are identical, it follows that

$$D(\mathbb{P}^+(r_t, s_t^1) \| \mathbb{P}^-(r_t, s_t^1) | x_t = x) = \frac{(\kappa x)^2}{2} \leq \frac{\kappa^2}{2},$$

where we have used standard results about the KL-divergence between two Gaussians. Further, since actions must be causal, and since the noise is independent, we conclude that over the whole trajectory,

$$D(\mathbb{P}^+(\mathcal{H}_T) \| \mathbb{P}^-(\mathcal{H}_T)) \leq \frac{T\kappa^2}{2}.$$

Let $x^{\text{av}} := (x^+ + x^-)/2 = \frac{1}{2(1-\kappa^2)}$. Observe that

- if the ground truth is $a^1 = (1 + \kappa)/2$ and $x_t \geq x^{\text{av}}$, then the algorithm incurs an instantaneous safety violation of at least $(1 + \kappa)/2 \cdot x^{\text{av}} - 1/4 = \frac{1+\kappa}{2} \cdot \frac{1}{2(1-\kappa^2)} - \frac{1}{4} = \frac{\kappa}{4(1-\kappa)} \geq \frac{\kappa}{4}$;

- if the ground truth is $a^1 = (1 - \kappa)/2$ and $x_t < x^{\text{av}}$, then the algorithm incurs an instantaneous efficiency regret of at least $\frac{1}{2(1-\kappa)} - \frac{1}{2(1-\kappa^2)} \geq \frac{\kappa}{2}$

Let \mathbf{A} be the event $\{\#\{t : x_t \geq x^{\text{av}}\} \geq T/2\}$. Using the Bretagnolle-Huber inequality ([Lattimore and Szepesvári, 2020](#), Thm. 14.2),

$$\mathbb{P}^+(\mathbf{A}) + \mathbb{P}^-(\mathbf{A}^c) \geq \frac{1}{2} \exp(D(\mathbb{P}^+(\mathcal{H}_T) \parallel \mathbb{P}^-(\mathcal{H}_T))) \geq \frac{1}{2} \exp(-T\kappa^2/2).$$

Let \mathcal{E}_T^- denote the efficiency regret incurred by the learner under \mathbb{P}^- and \mathcal{S}_T^+ denote the safety violation incurred by the learner under \mathbb{P}^+ . Under the event \mathbf{A} , if the true a was $(1 + \kappa)/2$, at least $T/2$ rounds incurred a safety regret of at least $\kappa/4$, and so $\mathcal{S}_T^+ \geq \kappa T/8$. Similarly, under \mathbf{A}^c , at least $T/2$ rounds had $z_t = -1$, implying that $\mathcal{E}_T^- \geq T\kappa/8$.

This implies that

$$\max(\mathbb{E}^-(\mathcal{E}_T^-), \mathbb{E}^+(\mathcal{S}_T^+)) \geq \frac{T\kappa}{8} \max(\mathbb{P}^+(\mathbf{A}), \mathbb{P}^-(\mathbf{A}^c)) \geq \frac{T\kappa}{32} \exp(-T\kappa^2/2).$$

For $T \geq 16$, we may choose $\kappa = 1/\sqrt{T} < 1/4$ to conclude that in at least one instance, the safety or efficiency regret incurred must be at least $\sqrt{T}/(32e^{1/2}) \geq \sqrt{T}/64$. \square

B.5.2 Necessity of Dependency on Gaps.

We establish [Theorem 4.6.7](#) through a reduction to earlier lower bounds in the safe multi-armed bandit problem ([Chen et al., 2022](#)).

The safe MAB problem involves d arms with mean rewards μ_k and mean safety risks ν_k each. The optimal arm, k^* , has a reward of μ_* and a safety risk $\nu_* < \alpha$. The associated efficiency and safety gaps are denoted as $\Delta_k := (\mu_* - \mu_k)_+$ and $\Gamma_k := (\nu_k - \alpha)_+$. In each round, the learner selects one arm and observes bounded signals with the above mean for both rewards and safety. Implicitly, this can be perceived as a linear bandit setting, with the known constraints being that x lies in a simplex, the reward vector θ , and the constraint vector a . However, it is important to note that this reduction is not entirely accurate. In the safe MAB problem, actions are

required to be solely on the corner points of the simplex, and playing in the interior is not permitted. While it is common to view x as a probability of selecting each arm in a MAB instance, this reduction fails due to the nonlinearity in our metrics. The safe MAB problem considers the metrics

$$\mathcal{E}_T^{\text{MAB}} := \sum (\mu_{A_t} - \theta^*)_+, \mathcal{S}_T^{\text{MAB}} := \sum (\nu_{A_t} - \alpha)_+.$$

Consequently, if the optimum of the SLB problem lies away from the corner points, the SLB problem can incur low regret, while the corresponding MAB actions would incur linear regret. Nonetheless, we shall argue below that for carefully designed instances, low regret in the linear bandit problem does ensure nontrivial regret in the safe MAB problem.

The primary result we leverage is a slight variation of Proposition 6 of (Chen et al., 2022), and it can be demonstrated using their proof.

Lemma B.5.1. *Let $f : \mathbb{N} \rightarrow [0, \infty)$ be any fixed function such that $f(T) \leq T$ for all T . If an algorithm ensures that, for every safe MAB instance, suboptimal arms are not played more than $f(T)$ times in expectation, then for every θ, a , there exists a choice of arm distributions for the safe MAB instance for which the means are as described, and the number of times each suboptimal arm k is played is lower bounded in expectation as*

$$\mathbb{E}[N_T^k] \geq \frac{1}{d(\mu_k \| \mu_*) \mathbb{1}\{\mu_k < \mu_*\} + d(\nu_k \| \alpha) \mathbb{1}\{\nu_k > \alpha\}} \left(\left(1 - \frac{f(T)}{T}\right) \log \frac{T}{f(T)} - \log(2) \right),$$

where $d(u \| v)$ is the KL divergence between Bernoulli laws with means u and v . In particular, these distributions are simply Bernoulli laws with the above means.

Our argument for the linear bandit proceeds as follows. We carefully design a safe linear bandit instance to essentially provide multi-armed bandit feedback by using the standard reduction that each coordinate of x_t represents the probability of pulling the corresponding arm. We show that, in the selected instance, achieving low linear regret ensures that the MAB regret is controlled (though to a weaker extent). Then,

exploiting the above lower bound, we argue that the regret of the safe linear bandit cannot be too good, as it would violate the mentioned lower bound.

Proof of Theorem 4.6.7. We initiate our proof by meticulously elucidating our primary constructions for the SLB and MAB, establishing a crude bound that enables us to apply Lemma B.5.1, and subsequently refining the analysis to demonstrate effective lower bounds on the SLB regret.

SLB Instance. We consider $d = 2$ with a single unknown constraint. Let $\theta = (\theta_1, \theta_2)$ and $a = (\alpha, a_2)$ be vectors in $[0, 1]^2$ such that $\theta_2 > \theta_1 > 0$, $a_2 > \alpha > 0$, and $\theta_2\alpha < \theta_1a_2$. The safe bandit instance we design is

$$\max \langle \theta, x \rangle : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1, \langle a, x \rangle \leq \alpha,$$

where the last constraint is unknown and the rest are known. Let us designate the three known constraints as b^1, b^2, b^3 . There are 6 BISs, with the associated points and gaps shown in Table B.1 below. Note that the only point meeting the constraints b^2 and b^3 is $(0, 1)$, which is infeasible. The situation is highly degenerate as the optimal point is $x^* = (1, 0)$, and three distinct BISs activate it. Nevertheless, each of these BISs is full rank. Furthermore, since the algorithm ensures that \mathcal{S}_T and \mathcal{E}_T are both $O(\sqrt{T})$ in general, our discussion below is effective.

Table B.1: Description of BISs in our construction.

BIS	Activating Point	$\zeta_*(I)$	$\eta_*(I)$
$(\{1\}, \{1\})$	$(0, \alpha/a_2)$	0	$(\theta_1a_2 - \alpha\theta_2)/(a_2 + \theta_2)$
$(\{1\}, \{2\})$	$(1, 0)$	0	0
$(\{1\}, \{3\})$	$(1, 0)$	0	0
$(\emptyset, \{1, 2\})$	$(0, 0)$	0	θ_1
$(\emptyset, \{2, 3\})$	$(1, 0)$	0	0
$(\emptyset, \{3, 1\})$	\emptyset	$a_2 - \alpha$	0

The gap of this instance is

$$\Xi := \min \left(\theta_1, a_2 - \alpha, \frac{\theta_1a_2 - \alpha\theta_2}{a_2 + \theta_2} \right).$$

Our construction requires that this is at least $\Omega(\min(\theta_1, a_2 - \alpha))$. This can always be ensured; for example, by using the parameterization $\theta_2 = 2\theta_1$, $a_2 = 4\theta_1$, $\alpha = \theta_1/2$,

where the expressions work out to

$$a_2 - \alpha = 7\theta_1/2, \frac{\theta_1 a_2 - \alpha \theta_2}{a_2 + \theta_2} = 3\theta_1/5,$$

giving us $\Xi \geq \theta_1/2$. We further impose the condition $4\theta_1 < 1/4$. Thus, this instance lets us express every value of $\Xi < 1/32$.

Continuation:

MAB Instance. The safe MAB instance associated with this SLB instance consists of three arms. The arms are characterized by their mean rewards μ_k and mean safety risks ν_k . We have arm 1 with $\mu_1 = \theta_1$ and $\nu_1 = 0$, arm 2 with $\mu_2 = \theta_2$ and $\nu_2 = a_2 - \alpha$, and arm 3 with $\mu_3 = 0$ and $\nu_3 = 0$. The optimal arm is arm 1, denoted as $k^* = 1$, with reward $\mu_* = \theta_1$ and safety risk $\nu_* = 0$. The efficiency gap and safety gap for each arm k are defined as $\Delta_k := (\mu_* - \mu_k)_+$ and $\Gamma_k := (\nu_k - \alpha)_+$, respectively.

Our reduction relies on Lemma B.5.1, and we need to show that for any algorithm ensuring that suboptimal arms are not played more than $f(T)$ times in expectation for all safe MAB instances, we obtain a lower bound on the number of times suboptimal arms are played in our specific MAB instance.

For the given MAB instance, we have arms 2 and 3 with non-zero safety risks, making them suboptimal. Applying Lemma B.5.1, we find that the expected number of times each suboptimal arm k is played is lower bounded by

$$\mathbb{E}[N_T^k] \geq \frac{1}{d(\mu_k \parallel \mu_*) \mathbb{1}\{\mu_k < \mu_*\} + d(\nu_k \parallel \alpha) \mathbb{1}\{\nu_k > \alpha\}} \left(\left(1 - \frac{f(T)}{T}\right) \log \frac{T}{f(T)} - \log(2) \right),$$

where $d(u \parallel v)$ is the KL divergence between Bernoulli laws with means u and v .

Specifically, for arms 2 and 3, we have

$$\mathbb{E}[N_T^2] \geq \frac{1}{d(\theta_2 \parallel \theta_1) + d(a_2 - \alpha \parallel \alpha)} \cdot \left(\left(1 - f(T)/T\right) \log \frac{T}{f(T)} - \log(2) \right).$$

Now, by choosing the parameterization $\theta_2 = 2\theta_1$, $a_2 = 4\theta_1$, $\alpha = \theta_1/2$, we can further simplify this lower bound.

This completes the construction and reduction, setting the stage for the detailed analysis that follows.

Safe MAB Instance. We will now detail the description of the associated MAB instance. Consider three arms characterized by their means and risks: arm 1 with mean reward $\mu_1 = 1/2 + \theta_1$ and mean safety risk $\nu_1 = 1/2 + \alpha$, arm 2 with $\mu_2 = 1/2 + \theta_2$ and $\nu_2 = 1/2 + a_2$, and arm 3 with $\mu_3 = 1/2$ and $\nu_3 = 1/2$. These arms are assumed to follow independent Bernoulli distributions with the associated means, forming the family of instances underlying Lemma B.5.1.

The connection between this MAB instance and the linear bandit instance is established as follows: each time a point (x_1, x_2) is selected, a random variable is sampled from $\{1, 2, 3\}$ according to the probability mass function $(x_1, x_2, 1 - x_1 - x_2)$. Subsequently, the corresponding arm is pulled, and the resulting rewards and risks, with $1/2$ subtracted, are supplied to the linear bandit instance.

This process ensures unbiased measurement of the mean for the linear bandit. Specifically,

$$\mathbb{E}[R] = x_1 \cdot (1/2 + \theta_1) + x_2 \cdot (1/2 + \theta_2) + (1 - x_1 - x_2) \cdot (1/2) - 1/2 = x_1\theta_1 + x_2\theta_2,$$

and similarly for the safety risk. The subtraction of $1/2$ is introduced to ensure that the KL divergences appearing in the bound of Lemma B.5.1 are of the form $d(1/2 \| 1/2 + \theta_1)$ and $d(1/2 + a_2 \| 1/2 + \alpha)$. This ensures that the arguments are bounded away from 0 and 1, resulting in a quadratic behavior for small θ_1 , as opposed to potentially worse dependence near 0 and 1.

To guarantee this favorable behavior, the condition $a_2 < 1/4$ is imposed, which implies $a_2 + 1/2 < 13/14$ is bounded away from 1. This condition originated from the previous paragraph's requirement $\theta_1 \leq 1/16$.

The key observation is that in the safe MAB instance, the expected number of times arm 2 is played is given by $\mathbb{E}[N_T^2] = \sum x_{t,2}$, and the expected number of times arm 3 is played is given by $\mathbb{E}[N_T^3] = \sum (1 - x_{t,1} - x_{t,2})$, where $x_{t,k}$ is the k -th component of x_t .

Crude Bound. Initially, we demonstrate that as long as the algorithm guarantees $\max(\mathcal{E}_T, \mathcal{S}_T) = O(T^{1-c})$, the play of suboptimal arms in the MAB instance is at least $\Omega(\theta_1^{-2} \log T)$.

Suppose the safe linear bandit ensures $\mathcal{E}_T \leq g(T)$ and $\mathcal{S}_T \leq g(T)$ for every instance, where $g(T) \leq T$ is an arbitrary monotonic function. Let $\zeta > 0$ be a parameter. If the

linear bandit instance ever plays a point (x_1, x_2) such that

$$\langle a, x \rangle \geq \alpha + \zeta \quad \text{or} \quad \langle \theta, x \rangle \leq \theta_1 - \zeta,$$

it would incur a pointwise cost of at least ξ in the round for either \mathcal{E}_T or \mathcal{S}_T . This implies that the number of rounds in which it plays such points is bounded as $g(T)/\zeta$. Therefore, in at least $\max(T - g(T)/\zeta, 0)$ rounds, the safe linear bandit instance plays in the region

$$P_\zeta := \{\langle a, x \rangle \leq \alpha + \zeta, \langle \theta, x \rangle \geq \theta_1 - \zeta, x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}.$$

Now, notice that both x_2 and $x_1 + x_2$ are upper bounded in this region. The corner points of this region are calculated to be

$$\begin{aligned} & \left(1 - \frac{\zeta}{\theta_1}, 0\right), \left(1 - \frac{\zeta}{a_2 - \alpha}, \frac{\zeta}{a_2 - \alpha}\right), \\ & \left(1 - \frac{\zeta}{\theta_1} \left\{1 + \frac{\theta_2(\theta_1 + \alpha)}{\theta_1(\theta_1 a_2 - \alpha \theta_2)}\right\}, \frac{\theta_1 + \alpha}{\theta_1 a_2 - \alpha \theta_2} \frac{\zeta}{\theta_1}\right), (1, 0). \end{aligned}$$

Therefore, ensuring that $a_2 - \alpha, \theta_1 a_2 - \alpha \theta_2 = \Omega(\theta_1)$, we have

$$x \in P_\zeta \implies x_2 \leq \frac{\zeta}{\theta_1}, (1 - x_1 - x_2) = O\left(\frac{\zeta}{\theta_1}\right).$$

Of course, outside of P_ζ , $x_2 \leq 1, 1 - x_1 - x_2 \leq 1$. This calculation holds for any ζ as long as $\zeta \ll \theta_1$. This implies that for every $\zeta = O(\theta_1)$,

$$\begin{aligned} \mathbb{E}[N_T^2] &\leq O\left(\frac{\zeta}{\theta_1}\right) T + \frac{g(T)}{\zeta}, \\ \mathbb{E}[N_T^3] &\leq O\left(\frac{\zeta}{\theta_1}\right) T + \frac{g(T)}{\zeta}. \end{aligned}$$

Thus, the safe MAB incurs regret bounds of at most $f(T) = O(\zeta T) + g(T)\theta_1/\zeta$.

Since $g(T) \leq CT^{1-c}$ for some constants C, c , by setting $\zeta = T^{-c/2}$, for sufficiently large T , we can establish the low-regret bound $\max(\mathbb{E}[\mathcal{E}_T^{\text{MAB}}], \mathbb{E}[\mathcal{S}_T^{\text{MAB}}]) \leq CT^{1-c/2}$.

Consequently, according to Lemma B.5.1, it follows that as $T \rightarrow \infty$,

$$\begin{aligned} \mathbb{E}[N_T^2] &\geq \frac{1}{d^{1/2} + 4\theta_1\|^{1/2} + \theta_1/2} \left((1 - o(1)) \frac{c}{2} \log T - O(1) \right) = \Omega(\theta_1^{-2} \log T), \\ \text{or } \mathbb{E}[N_T^3] &\geq \frac{1}{d^{1/2}\|^{1/2} + \theta_1} \left((1 - o(1)) \frac{c}{2} \log T - O(1) \right) = \Omega(\theta_1^{-2} \log T). \end{aligned}$$

To effectively utilize these bounds, we leverage a computer algebra system to demonstrate that

$$\forall \theta_1 \leq 1/16, d^{1/2} + 4\theta_1\|^{1/2} + \theta_1/2 \leq 27\theta_1^2, \text{ and } d^{1/2}\|^{1/2} + \theta_1 \leq 27\theta_1^2.$$

Concretely, these bounds yield

$$\mathbb{E}[N_T^2 + N_T^3] \geq \frac{c}{27\theta_1^2} \left((1 - o(1)) \log T - \frac{1}{c} \log(4) \right),$$

where the $o(1)$ term is $C/T^{c/2}$.

It is noteworthy that this bound is effective in our case since the method DOSLB does achieve $\max(\mathcal{E}_T, \mathcal{S}_T) = \tilde{O}(\sqrt{T})$ with high probability. In this scenario, we can set $c = 1/2 + \xi$ for any $\xi > 0$ in the above expressions.

Lower Bounds on SLB. Now, let's demonstrate the claims. We choose the instance with $\theta_2 = 2\theta_1, a_2 = 4\theta_1, \alpha = \theta_1/2$. In this case, the gaps are $(\theta_1, 7\theta_1/2, 3\theta_1/5)$, ensuring $\Xi \geq \theta_1/2$. Moreover, $\theta_1 a_2 - \alpha \theta_2 = 3\theta_1$, validating the claim on P_ζ for all $\zeta \leq \theta_1$. Consequently, against this instance, the earlier established lower bounds on $\mathbb{E}[N_T^2] + \mathbb{E}[N_T^3]$ hold.

Now, observe that for any choice of x_1, x_2 , the instantaneous efficiency regret and safety violations are given by:

$$\begin{aligned} (\theta_1 - \theta_1 x_1 - \theta_2 x_2)_+ &= \theta_1((1 - x_1 - x_2) - x_2)_+ \\ (\alpha x_1 + a_2 x_2 - \alpha)_+ &= \frac{\theta_1}{2}(7x_2 - (1 - x_1 - x_2))_+ \end{aligned}$$

However, both quantities are zero only if $x_2 \geq (1 - x_1 - x_2) \geq 7x_2 \implies x_2 = 1 - x_1 - x_2 = 0 \iff x_1 = 1$. So, in any round where $x_1 \neq 1$, at least one of these

quantities is nonzero. More quantitatively, we have:

$$\begin{aligned} \mathbb{E}[\mathcal{E}_T] + \mathbb{E}[\mathcal{S}_T] &\geq \theta_1 \sum \frac{5}{2} \mathbb{E}[x_{t,2}] + \frac{1}{2} \mathbb{E}[(1 - x_{t,1} - x_{t,2})] \\ &\geq \frac{\theta_1}{2} (\mathbb{E}[N_T^2] + \mathbb{E}[N_T^3]) \geq \frac{c(1 - o(1))}{54\theta_1} \log(T) - O(1), \end{aligned}$$

yielding the result upon recalling that $\theta_1 \geq \Xi \geq \theta_1/2$. □

Appendix C

Supplement for § 5

The proof for the generalized linear setting resembles that of the SLB case. We thus briefly present the key steps that endure quantities due to the generalized linear model.

C.1 Auxiliary Results for the Regret Bound

In this section, we revisit several crucial and frequently employed bounds from the literature. These encompass bounds on the confidence set, self-normalized vector series, and related concepts. This review serves to bolster our proof, enabling us to closely follow the derivation of the regret bound.

Lemma C.1.1.

$$\sum_{t=m+1}^{m+n} \|X_t\|_{V_{t-1}^2} \leq 2 \log \frac{\det V_{m+n+1}}{\det V_{m+1}} \leq 2d \log \left(\frac{\mathbf{tr}(V_{m+1}) + n}{d} \right) - 2 \log \det V_{m+1}$$

This lemma is adapted from Lemma 11 in (Abbasi-Yadkori et al., 2011) (together with Lemma B.1.1; we pick the most suitable variate here) and has been re-stated multiple times in the follow-ups, see (Li et al., 2017; Chen et al., 2023). Our use is to observe that the trace of the matrix $\mathbf{tr}(V_{m+1})$ grows at most linearly in the index m , and hence the upper bound grows logarithmically with the number of time steps.

We also re-state the following lemma for the self-containedness of the appendix.

Lemma C.1.2. *If $\lambda_{\min}(V_\tau) \geq 1$, then for any $t \geq \tau$ and any $\delta \in (0, 1)$*

$$\mathbb{P} \left(\|\theta - \hat{\theta}_t\|_{V_{t-1}} \leq \frac{1}{c_\mu} \sqrt{\frac{d}{2} \log\left(1 + \frac{2t}{d}\right) + \log\left(\frac{1}{\delta}\right)} \right) \geq 1 - \delta$$

$$\mathbb{P} \left(\|a^i - \hat{a}_t^i\|_{V_{t-1}} \leq \frac{1}{c_\nu} \sqrt{\frac{d}{2} \log(1 + \frac{2t}{d}) + \log(\frac{1}{\delta})} \right) \geq 1 - \delta$$

Lemma C.1.1 leads to the following important observation that controls the noise scale, which shall be referred to frequently.

Lemma C.1.3. *For the DOSGLB(δ) algorithm, the following holds*

$$\sum_{t=d+1}^T \rho_t^2 \leq (2d \log(1 + 2T/d) + 4 \log((U + 1)/d)) (2d \log((T + 1)/d)) = O(d^2 \log^2 T)$$

$$\sum_{t=d+1}^T \rho_t \leq O(d\sqrt{T} \log T)$$

Proof.

$$\begin{aligned} \sum_{t=d+1}^T \rho_t^2 &= \sum_{t=d+1}^T 4\beta_{t-1}(\delta) \|x_{t-1}\|_{V_{t-1}^{-1}}^2 \\ &\leq 4\beta_T(\delta) \sum_{t=d+1}^T \|x_{t-1}\|_{V_{t-1}^{-1}}^2 \\ &\leq (2d \log(1 + 2T/d) + 4 \log((U + 1)/d)) \sum_{t=d+1}^T \|x_{t-1}\|_{V_{t-1}^{-1}}^2 \\ &\leq (2d \log(1 + 2T/d) + 4 \log((U + 1)/d)) \left(d \log \left(\frac{\mathbf{tr}(V_{d+1}) + T - d}{d} \right) \right) \\ &\leq (2d \log(1 + 2T/d) + 4 \log((U + 1)/d)) (2d \log((T + 1)/d)) \end{aligned}$$

where we apply the monotonicity of β , the definition of β , Lemma C.1.1, and the fact that $\mathbf{tr}(V_{d+1}) \leq d + 1$ sequentially.

For the second part, simply apply the Cauchy-Schwarz inequality

$$\sum_{t=1}^T \rho_t \leq \sqrt{T} \sqrt{\sum_{t=1}^T \rho_t^2}.$$

□

Next we introduce a lemma stating the fact that the instantaneous regret is upper bounded by the noise scale. This resembles Lemma B.1.4

Lemma C.1.4. For any $t > d$, if $A \in \mathcal{C}_t(\delta)$ and $\theta \in \mathcal{C}_t^0(\delta)$, then $\forall x \in \mathcal{X}$,

$$\forall i \in [1 : U], \max_{\tilde{a}^i \in \mathcal{C}_t^i(\delta)} |\langle \tilde{a}^i - a^i, x \rangle| \leq \frac{\rho_t(x; \delta)}{c_\nu}, \quad \text{and} \quad \max_{\tilde{\theta} \in \mathcal{C}_t^0(\delta)} |\langle \tilde{\theta} - \theta, x \rangle| \leq \frac{\rho_t(x; \delta)}{c_\mu}.$$

Proof. Let's consider the case of θ ; similar bounds apply to a^i in the same manner.

Under the consistency assumption, $\theta \in \mathcal{C}_t^0$. Consequently,

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq |\langle \tilde{\theta} - \hat{\theta}, x \rangle| + |\langle \theta - \hat{\theta}, x \rangle|.$$

By leveraging the positive definiteness of V_{t-1} and the Cauchy-Schwarz inequality, we can further deduce that:

$$|\langle \theta - \hat{\theta}, x \rangle| = |\langle (\theta - \hat{\theta})V_{t-1}^{1/2}, V_{t-1}^{-1/2}x \rangle| \leq \|\theta - \hat{\theta}\|_{V_{t-1}} \cdot \|x\|_{V_{t-1}^{-1}}.$$

Performing the same calculation for $\tilde{\theta}$ and combining these bounds, we can conclude:

$$|\langle \tilde{\theta} - \theta, x \rangle| \leq (\|\theta - \hat{\theta}\|_{V_{t-1}} + \|\tilde{\theta} - \hat{\theta}\|_{V_{t-1}}) \|x\|_{V_{t-1}^{-1}}$$

Both θ and $\tilde{\theta}$ belong to \mathcal{C}_t^0 , which, by definition, implies that their V_{t-1} -norm distance from $\hat{\theta}$ is bounded by $\frac{1}{c_\mu} \sqrt{\beta_{t-1}(\delta)}$. The claim follows immediately, recalling that $\rho_t(x; \delta) := \frac{2}{c_\mu} \sqrt{\beta_{t-1}(\delta)} \|x\|_{V_{t-1}^{-1}}$. \square

Along with this line, we also restate the following lemma and provide the proof.

This is adapted from Lemma 4.5.1.

Lemma C.1.5. Assume the confidence sets are all valid, and the action of DOSLB at time t , x_t , noisily activates the BIS $I = (\mathfrak{A}, \mathfrak{R})$. Then the following holds.

$$\begin{aligned} Ax_t &\leq \nu^{-1}(\alpha) + \frac{\rho_t}{c_\nu} \mathbf{1}, & Bx &\leq \nu^{-1}(\beta) \\ A(\mathfrak{A})x_t &\geq \nu^{-1}(\alpha(\mathfrak{A})) - \frac{\rho_t}{c_\nu} \mathbf{1}, & B(\mathfrak{R})x_t &= \nu^{-1}(\beta(\mathfrak{R})), \\ \langle \theta, x_t \rangle &\geq \langle \theta, x^* \rangle - \frac{\rho_t}{c_\mu}. \end{aligned}$$

Proof. We first provide the proof for a^i , which is a direct application of Lemma C.1.4.

Pick any $i \in [1 : U]$, the following holds

$$\begin{aligned}\nu^{-1}(\alpha^i) &\geq \langle \tilde{a}_t^i, x_t \rangle \\ &\geq \langle a^i, x_T \rangle - \frac{\rho_t}{c_\nu}\end{aligned}$$

and for any $i \in \mathfrak{U}$, the following holds:

$$\begin{aligned}\nu^{-1}(\alpha^i) &= \langle \tilde{a}_t^i, x_t \rangle \\ &\leq \langle a^i, x_t \rangle + \frac{\rho_t}{c_\nu}\end{aligned}$$

And the rest is immediate when the concept of "noisily activating point" is introduced.

For the proof of θ , we need to observe that due to optimism

$$\langle \tilde{\theta}_t, x_t \rangle \geq \langle \theta, x^* \rangle$$

Hence

$$\begin{aligned}\langle \theta, x^* \rangle - \langle \theta, x_t \rangle &\leq \langle \tilde{\theta}_t, x_t \rangle - \langle \theta, x_t \rangle \\ &\leq \rho_t\end{aligned}$$

which completes the proof. □

C.2 Proofs on the Gaps

To establish the analysis based on the BIS, we (re)-state several related results from the literature.

Firstly, we adapt Prop 4.4.5 to the generalized linear case, which enables analysis centered on BISs.

Proposition C.2.1. $\forall t, \exists$ BIS $I_t = (\mathfrak{U}_t, \mathfrak{R}_t) : x_t \in \tilde{\mathcal{X}}_t^{I_t}$.

The next lemma is adapted from Lemma B.2.2, which serves as the basis of the activating polytope.

Lemma C.2.2. *Assume all of the confidence sets are valid. For any $t \geq d$ there exists at least one BIS $I = (\mathfrak{U}, \mathfrak{K})$ that x_t noisily activates, and such that $\begin{pmatrix} A(\mathfrak{U}) \\ B(\mathfrak{K}) \end{pmatrix} x_t \geq \nu^{-1} \left(\begin{pmatrix} \alpha(\mathfrak{U}) \\ \beta(\mathfrak{K}) \end{pmatrix} \right)$.*

The following lemma guarantees the fact that for any suboptimal BIS, at least one of the associated gaps must be positive. This is an important observation that relates the infinite action bandits to the finite action bandits (MABs): together with the next lemma that establishes the relationship between noise scale and the gap, it implies that only under the case where the noise is large enough, are there the chances that our algorithm will pick the suboptimal BISs.

Lemma C.2.3. *For any suboptimal BIS I , $\max(\zeta_*(I), \eta_*(I)) > 0$.*

Proof. We follow closely to the proof of Prop 4.5.7, but replacing the essential components by the GLB quantities.

Fix any suboptimal BIS $I = (\mathfrak{U}, \mathfrak{K})$, the primal program can be written as

$$\begin{aligned} P(\zeta; I) &= \max_x c_\mu \langle \theta, x \rangle \\ \text{s.t. } Ax &\leq \nu^{-1}(\alpha) + \frac{\zeta}{c_\nu} \mathbf{1} \\ &\quad - A(\mathfrak{U})x \leq -\nu^{-1}(\alpha(\mathfrak{U})) + \frac{\zeta}{c_\nu} \mathbf{1}(\mathfrak{U}) \\ B(\mathfrak{K})x &= \nu^{-1}(\beta(\mathfrak{K})) \\ B(\mathfrak{K}^c)x &\leq \nu^{-1}(\beta(\mathfrak{K}^c)), \end{aligned}$$

where $\mathfrak{K}^C = [1 : K] \setminus \mathfrak{K}$. If $\zeta_*(I) = \infty$ the claim naturally holds and we are done. Hence we only need to consider the case where $\zeta_*(I) < \infty$. In this case, for any $\zeta \geq \zeta_*(I)$, the (linear) program is feasible; also, since the activating polytope is always a subset of the bounded domain, i.e. $\mathcal{T}(\zeta; I) \subset \mathcal{X}$, the above program is also finite. For such a feasible and finite linear program, strong duality applies, and it follows that the dual is also feasible and finite, and shares the same value as the primal. We write down

the dual as follows:

$$\begin{aligned}
D(\zeta; I) = \min_{\lambda_+, \lambda_-, \eta, \gamma} & \left\langle \lambda_+, \nu^{-1}(\alpha) + \frac{\zeta}{c_\nu} \mathbf{1} \right\rangle + \left\langle \lambda_-, -\nu^{-1}(\alpha(\mathfrak{U})) + \frac{\zeta}{c_\nu} \mathbf{1}(\mathfrak{U}) \right\rangle \\
& + \langle \eta, \nu^{-1}(\beta(\mathfrak{K})) \rangle + \langle \gamma, \nu^{-1}(\beta(\mathfrak{K}^c)) \rangle \\
\text{s.t. } & A^\top \lambda_+ - A(\mathfrak{U})^\top \lambda_- + B(\mathfrak{K})^\top \eta + B(\mathfrak{K}^c)^\top \gamma = c_\mu \theta, \\
& \lambda_+ \geq 0, \lambda_- \geq 0, \gamma \geq 0.
\end{aligned}$$

To study the value of this program, we define the following auxiliary quantities. Abbreviate $\boldsymbol{\lambda} = (\lambda_+, \lambda_-, \eta, \gamma)$,

$$f(\boldsymbol{\lambda}) = \langle \lambda_+, \mathbf{1} \rangle + \langle \lambda_-, \mathbf{1}(\mathfrak{U}) \rangle$$

$$\begin{aligned}
g(\boldsymbol{\lambda}) = & \left\langle \lambda_+, \nu^{-1}(\alpha) + \frac{\zeta_*(I)}{c_\nu} \mathbf{1} \right\rangle + \left\langle \lambda_-, -\nu^{-1}(\alpha(\mathfrak{U})) + \frac{\zeta_*(I)}{c_\nu} \mathbf{1}(\mathfrak{U}) \right\rangle + \\
& \langle \eta, \nu^{-1}(\beta(\mathfrak{K})) \rangle + \langle \gamma, \nu^{-1}(\beta(\mathfrak{K}^c)) \rangle
\end{aligned}$$

$$h(\boldsymbol{\lambda}) = A^\top \lambda_+ - A(\mathfrak{U})^\top \lambda_- + B(\mathfrak{K})^\top \eta + B(\mathfrak{K}^c)^\top \gamma - c_\mu \theta$$

Hence the dual program can be rewritten as

$$D(\zeta; I) = \min_{\boldsymbol{\lambda}} \frac{\zeta - \zeta_*(I)}{c_\nu} f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : h(\boldsymbol{\lambda}) = 0, \lambda_+ \geq 0, \lambda_- \geq 0, \nu \geq 0$$

It is immediate that the set

$$\mathcal{F} := \{\boldsymbol{\lambda} : g(\boldsymbol{\lambda}) \leq P(\zeta_*(I); I), h(\boldsymbol{\lambda}) = 0, \lambda_+ \geq 0, \lambda_- \geq 0, \gamma \geq 0\}$$

is non-empty. Based on the feasibility of \mathcal{F} , we further define the following programs

$$D'(\zeta; I) = \min_{\boldsymbol{\lambda}} \frac{\zeta - \zeta_*(I)}{c_\nu} f(\boldsymbol{\lambda}) + g(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}$$

$$E(I) = \min_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}$$

Note that $E(I) < \infty$ since it is a minimization problem on a feasible set, and $D(\zeta; I) \leq D'(\zeta; I)$ since the program D' is introducing extra constraints, which cannot decrease the value of a minimization problem. In addition, we have the following: for

any $\zeta \geq \zeta_*(I)$

$$\begin{aligned} D'(\zeta; I) &\leq P(\zeta_*(I); I) + \min \left\{ \frac{\zeta - \zeta_*(I)}{c_\nu} f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F} \right\} \\ &= P(\zeta_*(I); I) + \frac{\zeta - \zeta_*(I)}{c_\nu} \cdot \min\{f(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathcal{F}\} \\ &= P(\zeta_*(I); I) + \frac{\zeta - \zeta_*(I)}{c_\nu} E(I). \end{aligned}$$

Combine everything together: for any $\zeta \geq \zeta_*(I)$,

$$P(\zeta; I) = D(\zeta; I) \leq D'(\zeta; I) \leq P(\zeta_*(I); I) + \frac{\zeta - \zeta_*(I)}{c_\nu} E(I)$$

Hence $\mathfrak{s}(I) \leq \max\{0, \frac{E(I)}{c_\nu}\} < \infty$.

Recall the expression of the efficacy gap, we are left to show that $\max\{\zeta_*(I), \xi(I)\} > 0$. If $\zeta_*(I) > 0$ we are done. Hence we only need to worry about the case where $\zeta_*(I) = 0$. In such case, $\lim_{\zeta \searrow 0} P(\zeta; I) > -\infty$, which implies $\mathcal{X}^I \neq \emptyset$. For a feasible suboptimal set, it must be the case that any associated point must be inefficient, i.e. $\forall x \in \mathcal{X}^I, \langle \theta, x^* \rangle > \langle \theta, x \rangle$. Since $\mathcal{T}(0; I)$ is compact, $P(\zeta_*(I); I) = P(0; I) < \langle \theta, x^* \rangle$ which implies $\xi(I) > 0$. \square

Finally, we present and prove the noise scale result.

Lemma C.2.4. *If at time t , the confidence sets are consistent and the action of DOSLB noisily activates the suboptimal BIS I , then $\rho_t \geq \max(\zeta_*(I), \eta_*(I))$.*

Proof. The following holds due to $x_t \in \mathcal{T}(\rho_t; I)$

$$c_\mu \langle \theta, x_t \rangle \leq P(\rho_t; I)$$

and

$$\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle - \frac{\rho_t}{c_\mu}$$

due to the optimism in x_t .

By definition of the feasibility gap $\rho_t \geq \zeta_*(I)$ since $x_t \in P(\rho_t; I)$.

If $\zeta_*(I) = \infty$ then we are done. Hence we only need to consider the case where $\zeta_*(I) < \infty$, and we are left to prove $\rho_t \geq \eta_*(I)$. By the definition of the spread, we

have

$$P(\rho_t; I) \leq P(\zeta_*(I); I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I)) = c_\mu \langle \theta, x^* \rangle - \xi(I) + \mathfrak{s}(I)(\rho_t - \zeta_*(I))$$

which implies that

$$-\rho_t \leq -\xi(I) + \mathfrak{s}(I)(\rho_t - \zeta^*(I))$$

and so

$$\rho_t \geq \frac{\xi(I) + \mathfrak{s}(I)\zeta_*(I)}{\mathfrak{s}(I) + 1} = \eta_*(I)$$

□

C.3 Proofs for the Regret Analysis

The following lemma is adapted from Lemma 4.6.4, showing that the play of DOSGLB is over efficient, and the safety risk can be controlled to any desired precision ε .

Lemma C.3.1. *Under Assumption 5.2, if the confidence sets are consistent, $t \geq d + 1$, and the action x_t of DOSLB(δ) is that x_t only noisily activates the optimal BIS, then $\langle \theta, x_t \rangle \geq \langle \theta, x^* \rangle$. Further, for any $\varepsilon > 0$, if $\rho_t(x_t) < \varepsilon$, then for every i , $\langle a^i, x_t \rangle \leq \alpha^i + \frac{k_\nu}{c_\nu} \varepsilon$.*

With this at hand, we are ready to prove the main claim.

Theorem C.3.2. *Under Assumption 5.2, $\forall \delta > 0$, the actions of DOSLB(δ) satisfy with high probability*

$$\mathcal{E}_T = O(d^2 \log^2 T / \Xi), \quad \text{and} \quad \mathcal{S}_T = \tilde{O}(d\sqrt{T}).$$

Proof. In this whole proof, we assume that the confidence sets are all valid. Hence the following development holds with probability at least $1 - \delta$. Firstly, ignore the first d steps since purely random exploration is happening in this stage. We then group the time steps into two sets: when suboptimal BISs are activated, and when the optimal BISs are activated.

$$\mathfrak{T}_1 := \{t \in [d + 1 : T] : \exists \text{ a suboptimal BIS } I \text{ such that } x_t \in \tilde{\mathcal{X}}_t^I\}.$$

For $t \leq d$, we use the crude bound of $k_\mu d$, for $t \in \mathfrak{T}_1$, we apply Lemma C.3.1, for $t \in \mathfrak{T}_1^C$, we apply Lemma C.2.4. Formally,

$$\begin{aligned}
\mathcal{E}_T &= \sum_{t=1}^T (\mu(\langle \theta, x^* \rangle) - \mu(\langle \theta, x_t \rangle))_+ \\
&= \sum_{t \leq d} (\mu(\langle \theta, x^* \rangle) - \mu(\langle \theta, x_t \rangle))_+ + \sum_{t \in \mathfrak{T}_1} (\mu(\langle \theta, x^* \rangle) - \mu(\langle \theta, x_t \rangle))_+ \\
&\quad + \sum_{t \notin \mathfrak{T}_1} (\mu(\langle \theta, x^* \rangle) - \mu(\langle \theta, x_t \rangle))_+ \\
&\leq k_\mu \sum_{t \leq d} \langle \theta, x^* - x_t \rangle + \frac{k_\mu}{c_\mu} \sum_{t \in \mathfrak{T}_1} \rho_t(x_t; \delta) + 0 \\
&\leq 2k_\mu d + \frac{k_\mu}{c_\mu} \sum_{t=d+1}^T \rho_t(x_t; \delta) \mathbb{1}\{\rho_t(x_t; \delta) \geq \Xi\} \\
&\leq 2k_\mu d + \frac{k_\mu}{c_\mu} \sum_{t=d+1}^T \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\Xi} \\
&= 2k_\mu d + \frac{k_\mu}{c_\mu \Xi} \sum_{t=d+1}^T \rho_t(x_t; \delta)^2 \\
&\leq 2k_\mu d + \frac{k_\mu}{c_\mu \Xi} (2d \log(1 + 2T/d) + 4 \log((U + 1)/d)) (2d \log((T + 1)/d)) \\
&= O(d^2 \log^2 T / \Xi)
\end{aligned}$$

We introduce an intermediate metric

$$\mathcal{S}_T^\varepsilon := \sum \max_i (\nu(\langle a^i, x_t \rangle) - \alpha^i - \frac{k_\nu}{c_\nu} \varepsilon)_+$$

for $\varepsilon > 0$. We further split the optimal rounds into whether x_t is overly unsafe or not via

$$\mathfrak{T}_2^\varepsilon := \{t \in [d + 1 : T] \setminus \mathfrak{T}_1 : \exists i : \langle a^i, x_t \rangle \geq \alpha^i + \frac{k_\nu}{c_\nu} \varepsilon\},$$

Similarly to \mathcal{E}_T , we can bound $\mathcal{S}_T^\varepsilon$ as follows

$$\begin{aligned}
\mathcal{S}_T^\varepsilon &= 2k_\nu d + \sum_{t \in \mathfrak{I}_1} \max_i (\nu(\langle a^i, x_t \rangle) - \alpha^i - \frac{k_\nu}{c_\nu} \varepsilon)_+ + \sum_{t \in \mathfrak{I}_2^\varepsilon} (\nu(\langle a^i, x_t \rangle) - \alpha^i - \frac{k_\nu}{c_\nu} \varepsilon)_+ \\
&\quad + \sum_{t \notin (\mathfrak{I}_1 \cup \mathfrak{I}_2^\varepsilon)} (\nu(\langle a^i, x_t \rangle) - \alpha^i - \frac{k_\nu}{c_\nu} \varepsilon)_+ \\
&\leq 2k_\nu d + \frac{k_\nu}{c_\nu} \sum_{t \in \mathfrak{I}_1} \rho_t(x_t; \delta) + \frac{k_\nu}{c_\nu} \sum_{t \in \mathfrak{I}_2^\varepsilon} \rho_t(x_t; \delta) + 0 \\
&\leq 2k_\nu d + \frac{k_\nu}{c_\nu} \sum_t \rho_t(x_t; \delta) \mathbf{1}\{\rho_t(x_t; \delta) \geq \Xi\} + \frac{k_\nu}{c_\nu} \sum_t \rho_t(x_t; \delta) \mathbf{1}\{\rho_t(x_t; \delta) \geq \frac{k_\nu}{c_\nu} \varepsilon\} \\
&\leq 2k_\nu d + \frac{k_\nu}{c_\nu} \sum_t \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\Xi} + \frac{k_\nu}{c_\nu} \sum_t \rho_t(x_t; \delta) \cdot \frac{\rho_t(x_t; \delta)}{\frac{k_\nu}{c_\nu} \varepsilon} \\
&= 2k_\nu d + \frac{k_\nu}{c_\nu} \left(\frac{1}{\Xi} + \frac{c_\nu}{k_\nu \varepsilon} \right) \sum_t \rho_t(x_t; \delta)^2 \\
&\leq 2k_\nu d + \frac{k_\nu}{c_\nu} \left(\frac{1}{\Xi} + \frac{c_\nu}{k_\nu \varepsilon} \right) \left(2d \log \left(1 + \frac{2T}{d} \right) + 4 \log \frac{U+1}{d} \right) \left(2d \log \frac{T+1}{d} \right).
\end{aligned}$$

Lastly, observe that $\mathcal{S}_T \leq \mathcal{S}_T^\varepsilon + \varepsilon T$, tuning $\varepsilon = \Theta(d \log T / \sqrt{T})$ results in

$$\mathcal{S}_T = \tilde{O}(d\sqrt{T})$$

□

References

- Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320.
- Agrawal, S. and Devanur, N. (2016). Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29:3450–3458.
- Agrawal, S. and Devanur, N. R. (2014). Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006.
- Agrawal, S., Devanur, N. R., and Li, L. (2016). An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. *Proceedings of Machine Learning Research*, 49:4–18.
- Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. *Proceedings of Machine Learning Research*, 23:39.1–39.26.
- Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for Thompson sampling. *Proceedings of Machine Learning Research*, 31:99–107.
- Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, volume 32.
- Amani, S., Alizadeh, M., and Thrampoulidis, C. (2020). Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3562–3566. IEEE.
- Arel, I., Liu, C., Urbanik, T., and Kohls, A. G. (2010). Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems*, 4(2):128–135.
- Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.
- Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256.

- Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE.
- Badanidiyuru, A., Langford, J., and Slivkins, A. (2014). Resourceful contextual bandits. *Proceedings of Machine Learning Research*, 35:1109–1134.
- Bernasconi, M., Cacciamani, F., Castiglioni, M., Marchesi, A., Gatti, N., and Trovò, F. (2022). Safe learning in tree-form sequential decision making: Handling hard and soft constraints. *Proceedings of Machine Learning Research*, 162:1854–1873.
- Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to linear optimization*, volume 6. Athena scientific Belmont, MA.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.
- Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.
- Camilleri, R., Wagenmaker, A., Morgenstern, J., Jain, L., and Jamieson, K. (2022). Active learning with safety constraints. *arXiv preprint arXiv:2206.11183*.
- Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.
- Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.
- Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. *Advances in neural information processing systems*, 24:2249–2257.
- Chen, T., Gangrade, A., and Saligrama, V. (2022). Strategies for safe multi-armed bandits with logarithmic regret and risk. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3123–3148. PMLR.
- Chen, T., Gangrade, A., and Saligrama, V. (2023). Doubly-optimistic play for safe linear bandits. *arXiv <https://doi.org/10.48550/arXiv.2209.13694>*.
- Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. *Proceedings of Machine Learning Research*, 15:208–214.

- Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *21st Annual Conference Computational Learning Theory*, pages 355–366.
- Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23.
- Gales, S. B., Sethuraman, S., and Jun, K.-S. (2022). Norm-agnostic linear bandits. *Proceedings of Machine Learning Research*, 151:73–91.
- Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. (2020a). Conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1431–1441. PMLR.
- Garcelon, E., Ghavamzadeh, M., Lazaric, A., and Pirotta, M. (2020b). Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3962–3969.
- Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings.
- Garivier, A., Ménard, P., and Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399.
- Genovese, M. C., Durez, P., Richards, H. B., Supronik, J., Dokoupilova, E., Mazurov, V., Aelion, J. A., Lee, S.-H., Coddington, C. E., Kellner, H., et al. (2013). Efficacy and safety of secukinumab in patients with rheumatoid arthritis: a phase ii, dose-finding, double-blind, randomised, placebo controlled study. *Annals of the rheumatic diseases*, 72(6):863–869.
- Hu, Q., Zhang, N., Quan, X., Bai, L., Wang, Q., and Chen, X. (2021). A user selection algorithm for aggregating electric vehicle demands based on a multi-armed bandit approach. *IET Energy Systems Integration*, 3(3):295–305.
- Jeřábek, E. (2004). Dual weak pigeonhole principle, boolean complexity, and derandomization. *Annals of Pure and Applied Logic*, 129(1-3):1–37.
- Katz-Samuels, J. and Scott, C. (2018). Feasible arm identification. *Proceedings of Machine Learning Research*, 80:2535–2543.
- Kaufmann, E., Cappé, O., and Garivier, A. (2012a). On bayesian upper confidence bounds for bandit problems. *Proceedings of Machine Learning Research*, 22:592–600.

- Kaufmann, E., Korda, N., and Munos, R. (2012b). Thompson sampling: An asymptotically optimal finite-time analysis. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *Algorithmic Learning Theory*, pages 199–213, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Kazerouni, A., Ghavamzadeh, M., Abbasi Yadkori, Y., and Van Roy, B. (2017). Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30.
- Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.
- Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. *Proceedings of Machine Learning Research*, 70:2071–2080.
- Liang, L., Ye, H., Yu, G., and Li, G. Y. (2019). Deep-learning-based wireless resource allocation with application to vehicular networks. *Proceedings of the IEEE*, 108(2):341–356.
- Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. (2021). Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767.
- Olwal, T. O., Djouani, K., and Kurien, A. M. (2016). A survey of resource management toward 5g radio access networks. *IEEE Communications Surveys & Tutorials*, 18(3):1656–1686.
- Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. *Proceedings of Machine Learning Research*, 130:2827–2835.
- Radhakrishnan, M. L. and Tidor, B. (2008). Optimal drug cocktail design: methods for targeting molecular ensembles and insights from theoretical model systems. *Journal of chemical information and modeling*, 48(5):1055–1073.
- Robbins, H. E. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.
- Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.
- Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

- Salkham, A. a., Cunningham, R., Garg, A., and Cahill, V. (2008). A collaborative reinforcement learning approach to urban traffic control optimization. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 560–566. IEEE.
- Sankararaman, K. A. and Slivkins, A. (2021). Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34:23191–23204.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.
- Vaswani, S., Yang, L., and Szepesvari, C. (2022). Near-optimal sample complexity bounds for constrained MDPs. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.
- Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., and Rellermeyer, J. S. (2020). A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):1–33.
- Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.
- Wang, Z., Wagenmaker, A., and Jamieson, K. (2021). Best arm identification with safety constraints. *arXiv preprint arXiv:2111.12151*.
- Wei, X., Jiang, Y., Liu, Q., and Wang, X. (2020). Calibration of phase shifter network for hybrid beamforming in mmwave massive mimo systems. *IEEE Transactions on Signal Processing*, 68:2302–2315.
- Wei, X., Jiang, Y., and Wang, X. (2019). Online calibration of phase shifter network for mmwave massive mimo systems in multipath channels. In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE.
- Wei, X., Jiang, Y., Wang, X., and Shen, C. (2021). Tx-rx reciprocity calibration for hybrid massive mimo systems. *IEEE Wireless Communications Letters*, 11(2):431–435.
- Wei, X. and Shen, C. (2022). Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1253–1268.
- Wei, X., Shen, C., Yang, J., and Poor, H. V. (2023a). Random orthogonalization for federated learning in massive mimo systems. *IEEE Transactions on Wireless Communications*.

- Wei, X., Wang, T., Huang, R., Shen, C., Yang, J., and Poor, H. V. (2023b). Floras: Differentially private wireless federated learning using orthogonal sequences. In *ICC 2023-IEEE International Conference on Communications*, pages 3121–3126. IEEE.
- Wu, Y., Shariff, R., Lattimore, T., and Szepesvari, C. (2016). Conservative bandits. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1254–1262, New York, New York, USA. PMLR.
- Yu, H., Neely, M., and Wei, X. (2017). Online convex optimization with stochastic constraints. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

CURRICULUM VITAE

