MULTI-ARMED BANDIT ALGORITHMS FOR COMMUNICATION NETWORKS AND HEALTHCARE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
İlker Demirel
June 2022

Multi-Armed Bandit Algorithms for Communication Networks and Healthcare
By İlker Demirel
June 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Cem Tekin(Advisor)

Aykut Koç

Elif Vural

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

MULTI-ARMED BANDIT ALGORITHMS FOR COMMUNICATION NETWORKS AND HEALTHCARE

İlker Demirel
M.S. in Electrical and Electronics Engineering
Advisor: Cem Tekin
June 2022

Multi-armed bandits (MAB) is a well-established sequential decision-making framework. While the simplest MAB framework is useful in modeling a wide range of real-world applications ranging from adaptive clinical trial design to financial portfolio management, it requires further extensions for other problems. We propose three novel MAB algorithms that are useful in optimizing bolus-insulin dose recommendation in type-1 diabetes, best channel identification in cognitive radio networks, and online recommender systems. First, we introduce and study the "safe leveling" problem, where the learner's objective is to keep the arm outcomes close to a target level rather than maximize them. We propose a novel algorithm, ESCADA, with cumulative regret and safety guarantees. We demonstrate its effectiveness against the straightforward adaptations of standard MAB algorithms to the "leveling task". Next, we study the "federated multi-armed bandit" (FMAB) problem, where a cohort of clients play the same MAB game to learn the globally best arm. We consider adversarial "Byzantine" clients disturbing the learning process with false model updates and propose a robust algorithm, Fed-MoM-UCB. We provide theoretical guarantees on Fed-MoM-UCB while identifying the certain performance sacrifices that robustness requires. Finally, we study the "combinatorial multi-armed bandits with probabilistically triggered arms" (CMAB-PTA), where the learner chooses a set of arms at each round that may trigger other arms. CMAB-PTA is useful in modeling various problems such as influence maximization on graphs and online recommendation systems. We propose a Gaussian process-based algorithm, ComGP-UCB. We provide upper bounds on its regret and demonstrate its effectiveness against the state-of-the-art baselines when arm outcomes are correlated.

Keywords: Multi-armed bandits, Federated learning, Machine Learning, Communication networks, Healthcare.

ÖZET

İLETİŞİM AĞLARI VE SAĞLIK UYGULAMALARI İÇİN ÇOK KOLLU HAYDUT ALGORİTMALARI

İlker Demirel
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Cem Tekin
Haziran 2022

Çok kollu haydutlar (ÇKH) ardışık karar verme modellerinin başında gelmektedir. En temel ÇKH modeli adaptif klinik deneylerden finansal potrföy planlamasına kadar birçok gerçek dünya uygulamasını başarı ile modellerken, farklı uygulamalar için yetersiz kalabilmektedir. Bu tezde, tip-1 diyabet tedavisinde bolus-insülin doz önerisi, bilişsel radyo ağı uygulamaları için en iyi kanal tespiti, ve çevrimiçi öneri sistemleri için üç yenilikçi ÇKH algoritması tasarlanmıştır. İlk olarak, öğrenicinin aksiyonların sonuçlarını belirli hedef bir değere yakın tutmaya çalıştığı "güvenli kontrol" problemini tanıtılmış ve bu problem için güvenlik ve performans garantileri sağlanan "ESCADA" algoritması tasarlanmıştır. ESCADA algoritmasının "güvenli kontrol" probleminde başarısı ve standart ÇKH algoritmalarının yüzeysel uzantılarına üstünlüğü deneylerle desteklenmiştir. Sonrasında, birden çok öğrenicinin aynı ÇKH oyununu oynadığı "federe çok kollu haydutlar" (FÇKH) problemi çalışılmıştır. Öğrenicilerin bir kısmının performansı düşürmeye yönelik ataklar gerçekleştiren "Bizanslı" öğreniciler olduğu senaryo göz önüne alınmış, ataklara dayanıklı "Fed-MoM-UCB" algoritması geliştirilmiş, ve etkinliği deneyler ile gösterilmiştir. "Fed-MoM-UCB" algoritmasının pişmanlık analizi yapılmış ve dayanıklılık için feragat edilmesi gereken performans garantileri tespit edilmiştir. Son olarak, öğrenicinin her turda birden fazla kol (aksiyon) seçmesi gereken "olasılıksal tetiklemeli birleşimsel çok kollu haydutlar" (OT-BÇKH) problemi çalışılmıştır. OT-BÇKH problemi için kollar arası korelasyonları kullanabilen "ComGP-UCB" algoritması tasarlanmış ve pişmanlık analizi yapılmıştır. "ComGP-UCB" algoritmasının kollar arası korelasyonun yüksek olduğu senaryolarda rakip algoritmalara üstünlüğü deneyler ile desteklenmiştir.

Anahtar sözcükler: Çok kollu haydutlar, Federe öğrenme, Makine öğrenmesi, İletişim ağları, Sağlık.

Acknowledgement

I would like to extend my sincere thanks to my advisor Assoc. Prof. Cem Tekin for his invaluable support and guidance in my studies, which have been essential in completing this dissertation.

I would also like to express my gratitude to the members of the thesis committee, Asst. Prof. Aykut Koç and Assoc. Prof. Elif Vural, for their time and thoughtful comments.

I am much obliged to my excellent collaborators, Ahmet Alparslan Çelik and Yiğit Yıldırım, for making some of the work in this thesis possible.

I am grateful to Ozan Aygün and Emre Mumcuoğlu for their invaluable friendships.

Finally, I thank my family for their love and support throughout my studies.

This work is supported by Vodafone within the framework of 5G and Beyond Joint Graduate Support Programme coordinated by Information and Communication Technologies Authority, and in part by the Scientific and Technological Research Council of Turkey under Grant 215E342.

Contents

1 Introduction **1**

1.1 Multi-Armed Bandits 1

1.2 Our Contributions and Organization 4

2 Related Work **6**

2.1 Safe Leveling Bandits 6

2.1.1 The Leveling Problem 6

2.1.2 Safe Exploration 7

2.1.3 Precision Medicine and Blood Glucose Control 10

2.2 Federated Bandits 10

2.2.1 Federated Learning 11

2.2.2 Robust Federated Learning 12

2.2.3 Federated Multi-armed Bandits 13

2.2.4 Byzantine Resilient FMAB 15

2.3 Combinatorial MABs with Probabilistically Triggered Arms . . . 15

3 ESCADA: Efficient Safety and Context-Aware Dose Allocation for Precision Medicine 18

3.1 Introduction . . . 18

3.2 Problem Formulation . . . 19

3.2.1 Regularity Assumptions . . . 20

3.3 ESCADA Algorithm . . . 21

3.3.1 Acquisition Strategy . . . 21

3.3.2 Safety Awareness . . . 24

3.4 Theoretical Analyses . . . 24

3.5 Experimental Results . . . 28

3.5.1 Experimental setup . . . 28

3.5.2 Performance metrics . . . 28

3.5.3 Single meal event (SME) scenario . . . 29

3.5.4 Multiple meal events (MME) scenario . . . 30

3.5.5 Algorithms . . . 30

3.5.6 Dose calculators . . . 31

3.5.7 Discussion of results . . . 31

4 Federated Multi-Armed Bandits Under Byzantine Attacks 38

4.1 Introduction 38

4.2 Problem Formulation 39

4.3 Fed-MoM-UCB Algorithm 42

4.3.1 Client Side 42

4.3.2 Server Side 42

4.4 Theoretical Analyses 45

4.4.1 MoM Estimator Concentration 45

4.4.2 Main Results and Regret Bound 49

4.4.3 Choosing the mapping 55

4.5 Experimental Results 57

4.5.1 Scenario 1 58

4.5.2 Scenario 2 61

4.5.3 Scenario 3 63

5 Combinatorial Gaussian Process Bandits with Probabilistically Triggered Arms 65

5.1 Introduction 65

5.2 Problem Formulation 66

5.2.1 Base Arm Outcomes 66

5.2.2 Triggering Process 67

5.2.3 Reward 67

5.2.4 Observation Model 69

5.3 ComGP-UCB Algorithm 69

5.4 Theoretical Analysis 71

5.4.1 Main results 71

5.4.2 Preliminaries for Theorems 1 and 2 74

5.4.3 Supplemental Lemmas 74

5.4.4 Proof of Theorem 1 75

5.4.5 Proof of Theorem 2 77

5.5 Experimental Results 79

5.5.1 Cascading Bandit 79

5.5.2 Sparse GPs and Hyper-parameter Mismatch 81

6 Concluding Remarks and Future Directions 82

A ESCADA: Efficient Safety and Context-Aware Dose Allocation for Precision Medicine 100

A.1 Proof of Theorem 3.1 100

A.2 Proof of Theorem 3.2 102

A.3 Auxiliary Results 103

A.4 Proof of Theorem 3.3 108

B Combinatorial Gaussian Process Bandits with Probabilistically Triggered Arms **110**

B.1 Preliminaries 110

B.2 Proof of Lemma 5.1 111

B.3 Proof of Lemma 5.2 112

B.4 Proof of Lemma 5.3 113

B.5 Proof of Lemma 5.4 114

List of Figures

3.1 *ESCADA Algorithm Description (left).* Upon observing a context $\boldsymbol{z}_n$ in round n, TACO forms the set $\mathcal{C}_n(\boldsymbol{z}_n) \subseteq S_n(\boldsymbol{z}_n)$ after eliminating the doses that are suboptimal with high probability. If $\mathcal{C}_n(\boldsymbol{z}_n) \neq \emptyset$, it recommends the dose whose mean response is closest to the target T. If $\mathcal{C}_n(\boldsymbol{z}_n) = \emptyset$, it recommends the dose with the widest confidence interval in $S_n(\boldsymbol{z}_n) \cap \bar{\mathcal{D}}_{\boldsymbol{z}_n}$. *Flowchart (right).* A simple interpretation of the dose allocation process intended for domain experts. 22

3.2 33

3.3 35

4.1 An exemplary scenario with ten clients where two are Byzantine (red lines), and they perform outlier attacks. Fed-MoM-UCB first divides the clients into five groups. In the worst case, the Byzantine clients are assigned to different groups and saturate their groups' local updates. The saturation will not affect the median estimate in any scenario. 44

4.2 The effect of Byzantine clients' fraction λ and the mapping $\alpha(\lambda)$ on the unavoidable error margin ω, and the constants Γ_4 and Γ_5. The ladder structure occurs due to the dependence on the number of clients in a group, B, which is an integer. To draw the figures, we take $K = 10$, $T = 5 \times 10^4$, $\delta = 0.001$, $\sigma = 0.5$, and $\sigma_c = 0.01$. . 56

4.3 Regrets and communication costs for "Scenario 1" ($\pm$ 2 std. dev.). 60

4.4 Regrets and communication costs for "Scenario 2" ($\pm$ 2 std. dev.). 62

4.5 Regrets and communication costs for "Scenario 3" ($\pm$ 2 std. dev.). 64

5.1 *Algorithm Flow.* At each round t, the environment draws a random outcome vector $\boldsymbol{Y}(t)$ from the true outcome distribution $\mathcal{D}$. ComGP-UCB first forms an optimistic estimate for the expected base arm outcome vector, $\bar{\boldsymbol{\mu}}_t$, using the posterior GP distribution and feeds it to the oracle. The oracle knows the problem structure and it returns an optimal super-arm $S(t)$ for the given parameter set, $\bar{\boldsymbol{\mu}}_t$. Then, the learner plays the super-arm $S(t)$ and observes a feedback, $Q(t)$, which depends on $\boldsymbol{Y}(t)$. Finally, the learner updates the GP distribution based on the feedback it observes. Blue boxes are the variables that are created and updated by the learner, whereas the red boxes are external and environmental entities. . . 70

5.2 Regrets and communication costs for "Scenario 1" ($\pm$ 2 std. dev.). 80

List of Tables

1.1 True (p, unknown) and estimated ($\hat{p}$) reward yielding probabilities for Bernoulli arm rewards after one trial for each arm. 3

3.1 "-TC" indicates that tuned calculator was used. Target PPBG is $T = 112.5$ mg/dl. "PPBG" column is averaged over all recommendations. "Hyper" and "Hypo" columns denote the hyperglycemia and hypoglycemia event frequencies, respectively, averaged over all patients. Similarly, HBGI and LBGI are averaged over all patients. We report (`mean` ± `stdandard deviation`). 32

4.1 Different choices $\alpha(\lambda)$ satisfying Definition 4.1. 55

4.2 Unknown global mean arm outcomes 58

4.3 Parameters for different scenarios. The "M", "B", and "G" columns show the number of total clients recruited, the number of clients in a group, and the number of groups for the Fed-MoM-UCB algorithm, respectively. The "M(p)" and "M" columns show the number of clients recruited in phase p and the number of total clients recruited by the Fed2-UCB algorithm, respectively. (± standard deviation). 59

Chapter 1

Introduction

This thesis proposes novel multi-armed bandit algorithms for various applications, including communications, healthcare, and personalized online recommendations. Considering the practical requirements that may be of pivotal importance in different applications, we design safe, robust, and generic MAB algorithms with theoretical guarantees and corroborate our findings with experimental results.

This section introduces the multi-armed bandit framework and related literature on some active research directions to which the works in the present thesis contribute.

1.1 Multi-Armed Bandits

Multi-armed bandits (MAB) is one of the simplest reinforcement learning models where the tradeoff between *exploration* and *exploitation* can be elegantly analyzed [1]. In the simplest MAB problem, the learner interacts with its environment through a set of K arms whose outcome distributions are unknown *a priori*. At each round, the learner pulls an arm and observes an immediate stochastic *reward* (outcome). The natural performance metric in a MAB problem is the cumulative *regret*, which is the expected difference in cumulative rewards obtained by the

learner and by an omniscient player who knows the reward distributions and consistently plays the best arm. The learner's ultimate goal is to minimize the cumulative regret by sequentially learning to pull better arms based on her past observations. In order to do that, it must both *explore* different arms to have confident estimates of arms' outcomes and *exploit* by playing the arms that have yielded high outcomes to control the growth of cumulative regret. [2] shows that when the arm outcomes are independent, the cumulative regret will increase at least logarithmically over time (i.e., $\mathcal{O}(\log T)$). That is, exploration is necessary, and it is not free.

Before delving into the MAB literature, let us motivate the importance of balancing the exploration and the exploitation. Consider a simple scenario where the learner can choose one of the $K = 2$ arms at each round t. The reward of each arm k is a Bernoulli random variable with unknown success probability p_k. That is, upon playing the arm $I(t) \in \{1, 2\}$ at round t, the learner collects the following instantaneous reward,

$$X_{I(t)}(t) = \begin{cases} 1, & \text{with probability } \ p_{I(t)} \\ 0, & \text{with probability } \ 1 - p_{I(t)} \end{cases}$$

Let us denote the sample mean estimate of p_k at the end of round t by $\hat{p}_k(t)$. We have,

$$\hat{p}_k(t) = \frac{1}{N_k(t)} \sum_{\tau=1}^{t} \mathbb{I}\{I(\tau) = k\} X_{I(\tau)}(\tau) \ ,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and $N_k(t)$ denotes the number of times the learner played the arm k by the end of round t. Now, let the learner play each arm one time by the end of round $t = 2$. Table 1.1 shows the true success probabilities p_k that are unknown to the learner, and a possible set of empirical mean success probabilities $\hat{p}_k(2)$.

Based on the scenario in Table 1.1, "Arm 1" yielded a reward and "Arm 2" did not. Consider the case where the learner employs the following greedy exploitation strategy to choose an arm at round $t > 2$,

$$I(t) = \underset{k}{\operatorname{argmax}}\, \hat{p}_k(t-1) \ . \tag{1.1}$$

Table 1.1: True (p, unknown) and estimated ($\hat{p}$) reward yielding probabilities for Bernoulli arm rewards after one trial for each arm.

	Arm 1	Arm 2
p_k	**0.5**	**0.7**
$\hat{p}_k(2)$	**1**	**0**

That is, at any round $t > 2$, the learner chooses the arm that looks best *so far*. Then, it should choose "Arm 1" at round $t = 3$, since we have $\hat{p}_1(2) > 0$ and $\hat{p}_2(2) = 0$. Due to this greedy exploitation strategy, the learner will keep choosing "Arm 1", since we will always have $\hat{p}_1(t) > 0$ and $\hat{p}_2(t) = 0$ for all $t > 2$. This is a suboptimal strategy since "Arm 2" is the better arm in reality. That simple experiment reveals the importance of sufficiently *exploring* different arms, to have a better estimate of their reward distributions.

Intuitively, one can articulate that if an arm is played fewer times, then our estimates for that arm are less accurate than the others that are played more. For instance, in the previous example, the learner should come up with a different arm acquisition strategy than (1.1) to guarantee that the better arm ("Arm 2") will be selected in the long term. It can do that by incorporating an additional term that favors the arms which are played fewer times as follows,

$$I(t) = \underset{k}{\operatorname{argmax}} \left\{ \hat{p}_k(t-1) + \sqrt{\frac{2 \log t}{N_k(t-1)}} \right\} . \tag{1.2}$$

With that strategy, an arm played fewer times can continue to be chosen over time due to the additional term. The acquisition strategy in (1.2) is the famous UCB1 algorithm [3].

Different strategies have been proposed to tradeoff exploration and exploitation and to minimize the cumulative regret, mainly rooting from two lines of algorithms: Thompson sampling (TS)-based [4, 5] and upper confidence bound (UCB)-based [2, 3, 6] algorithms (e.g., (1.2)). TS-based algorithms maintain a posterior distribution over arm outcomes. At each round, the learner samples random estimates of arm outcomes from their posterior distributions and plays

an arm accordingly. Due to this random sampling procedure, TS methods intrinsically perform exploration. UCB-based algorithms (e.g., (1.2)) leverage the *optimism in the face of uncertainty* (OFU) principle to form optimistic estimates of arm outcomes and play the arm that looks best under these estimates [3, 7–10]. MAB algorithms are used to model and solve a variety of real-world applications including online recommendation systems [11, 12], influence maximization on social networks [13], financial portfolio management [14, 15], anomaly detection [16], control and robotics [17, 18], communications [19], and design of adaptive clinical trials [20, 21].

1.2 Our Contributions and Organization

We study various problems that are of both theoretical and practical interest in the following chapters. We propose novel multi-armed bandit algorithms with sound theoretical guarantees and corroborate our findings with detailed experimental results. We investigate the following problems.

- Chapter 3 studies the "safe leveling" bandit problem where the objective is to keep the outcomes close to a target rather than maximize them. Such an objective may be relevant in different domains, especially healthcare. We propose ESCADA, which uses Gaussian processes to learn the underlying function and safely and efficiently implements the safe leveling task.

- Chapter 4 analyzes the "federated multi-armed bandit" (FMAB) problem, a recently emerging framework where a cohort of clients play the same MAB game with different local models. Among other applications, FMAB is useful in modeling the cognitive radio problem where a fixed base station attempts to learn the best channel on average in a coverage area. We study the FMAB problem under "Byzantine" attacks and propose a robust algorithm, Fed-MoM-UCB.

- Chapter 5 investigates the "combinatorial multi-armed bandits with probabilistically triggered arms" (CMAB-PTA). Most real-world applications,

from online recommendation systems to influence maximization, cannot be modeled with the vanilla MAB framework. The learner needs to select a *set* of arms that can *activate* other arms as well. We propose the ComGP-UCB algorithm for the CMAB-PTA framework, which uses Gaussian processes to capture and exploit the correlations between arm outcomes.

The following chapter provides background on some of the related MAB literature from which we draw inspiration and make contributions.

Chapter 2

Related Work

2.1 Safe Leveling Bandits

This section introduces the "leveling" bandit problem with a two-sided safety constraint. This framework helps model various real-world problems, especially in healthcare, where the objective is to control a target variable while respecting a two-sided safety constraint. In Chapter 3, we provide a novel algorithm for the safe leveling bandit framework. We use Gaussian processes to model and efficiently learn the unknown objective function, which determines the arm outcomes [22].

2.1.1 The Leveling Problem

Maximizing the expected cumulative reward is equivalent to maximizing the outcomes of the arms in the classical MAB setting. However, in some problems, the goal may be to keep the outcomes close to a target level T rather than maximizing them, which we refer to as *leveling*. One such example is the blood glucose control in type 1 diabetes, where the goal is to keep the patients' blood glucose levels close to a comfortable level. Traditional UCB-based algorithms, such as UCB1 [3]

or GP-UCB [10] are inapplicable to the leveling objective. That is because UCB-based algorithms use the *optimism in the face of uncertainty* principle to generate optimistic estimates of arm outcomes and select the arm with the *maximum* estimated outcome at each round. In our case, however, *optimism* refers to how close an arm's outcome is to the target level. Because of the fundamental difference in our task, we must make radical changes to our arm selection strategy. It is possible to use "pseudo-rewards" such as $r_\alpha(n) = \log\big(1/(|\alpha(n) - T| + \epsilon)\big)$, where $\alpha(n)$ is the outcome of arm α at round n, T is the target level, and $\epsilon > 0$. Since $r_\alpha(n)$ is decreasing with $|\alpha(n) - T|$, keeping the outcome close to the target level will yield higher rewards, in which case one can use the UCB-based algorithms to maximize the pseudo-rewards. We avoid defining pseudo-rewards and use readily available bandit algorithms as different reasonable choices for the pseudo-reward will cause the algorithm to behave differently in practice. Therefore, we keep the objective (i.e., minimize $|\alpha(n) - T|$) in the most generic form and propose a suitable novel acquisition strategy with theoretical guarantees in Chapter 3. We compare our algorithm to simple adaptations of UCB-based algorithms for the "leveling" task and demonstrate its effectiveness with experimental results.

2.1.2 Safe Exploration

While our goal is to keep the arm outcomes close to a target variable T, we also consider a two-sided safety requirement. For example, in the previous blood glucose control problem, we had to ensure patient safety during the "exploration" phase. Hypoglycemia (hyperglycemia) occurs when the patient's blood glucose level falls below 70 mg/dl (or exceeds 180 mg/dl). Both events have the potential to result in life-threatening conditions [23]. As a result, we create algorithms that guarantee (with high probability) that the outcomes of the dose recommendations (i.e., arms) will always remain within a previously determined safe range $[T\text{min}, T\text{max}]$ during exploration.

2.1.2.1 Risk-averse and Conservative Bandits

There is some related literature on "risk-averse" MABs, where the learner is concerned not only with maximizing long-term earnings but also with reducing a certain measure of *risk* [24, 25]. [25–27] investigate the MAB problem using two risk measures, Mean-Variance and Value-at-Risk. The Mean-Variance measure attempts to strike a balance between an arm's expected outcome and its risk (variance), which is remarkably important in financial portfolio management [25, 27, 28]. In a similar fashion, the Value-at-Risk measure, which is widely used in financial risk management, is concerned with the minimum return of an arm at a confidence level $p \in (0, 1)$ [26]. [29–31] study the Conditional-Value-at-Risk measure, which is discussed to capture the tail-risk better compared to the Value-at-Risk measure. Another area of study is the "conservative" bandit setting, in which the learner's cumulative reward must exceed a predetermined fraction of a baseline's at any given time [32, 33]. These works, however, do not address *stagewise* safety constraints on instantaneous arm outcomes, which must be explicitly satisfied at any given time.

2.1.2.2 Stagewise safety

Because many real-world problems require strict stagewise safety guarantees during exploration, there is a surge of interest in safe exploration for Bayesian optimization (BO), Markov decision processes, MABs, and reinforcement learning in general [34–37]. Clinical trials are an inspiring example of such applications, where the goal is to identify the best therapy while ensuring patient safety at all times [20, 38]. [39] studies a safe variant of the UCB1 algorithm in [3] (safe linear UCB) preceded by a pure exploration phase. [40] proposes a safe algorithm for the linear stochastic bandit setting with a near-optimal expected regret, which forms and employs lower confidence bounds of arm outcomes to ensure the safety of the exploration and greedily exploits when it is safe to do so. [41] extends the linear TS algorithm in [42] to a safety constrained setting with certain adjustments on the randomization routine to leverage the "optimism in the face of uncertainty"

principle under safety constraints. [43] investigates a safe variant of the Frank-Wolfe algorithm for solving a smooth optimization problem with unknown linear safety constraints and provides convergence guarantees but no upper bound on the regret. [36, 37] consider Bayesian optimization with safety constraints. However, they seek to find optimal safe solutions and allow unsafe evaluations during the exploration. [44] propose a safe variant of GP-UCB, which employs a pure exploration phase at the beginning and provides upper bounds on its cumulative regret. [45, 46] investigate safety for a broad class of smooth reward functions using Gaussian processes and propose the SafeOPT and StageOPT algorithms, which guarantee the safety of the exploration process. However, they model safe set exploration as a proxy objective, resulting in unnecessary suboptimal evaluations at the safe set's boundaries [47]. Furthermore, despite the convergence guarantees, they do not provide formal upper bounds on the cumulative regret. Goal-oriented Safe Expansion (GoOSE) algorithm, on the other hand, works as a *plug-in* safety mechanism with any acquisition function and encourages safe set expansion only when necessary [47]. When the query is not guaranteed to be safe, GoOSE attempts to expand the safe set by iteratively evaluating the function at safe points to learn more about the safety of the initial query, preserving the performance of the original acquisition strategy within safety constraints. However, in many practical problems, such re-evaluations may be impossible (e.g., dynamic treatment assistance).

All of the works presented above consider the classical MAB setting, where the goal is to maximize the outcomes of the actions while taking into account a one-sided safety constraint. However, we are interested in a different problem in which the goal is to keep the outcomes of the actions close to a target level, which we call *leveling*, while respecting a *two-sided* safety constraint at each round. In Chapter 3, we formalize this problem from a MAB perspective by using a suitable definition of the *regret* as a proxy performance metric to capture our goal.

2.1.3 Precision Medicine and Blood Glucose Control

Precision medicine aims to provide the best possible treatment on an individual level by taking into account the variability of patient characteristics [48, 49]. Many healthcare problems demand maintaining a physiological variable (e.g., blood glucose level) in a safe range, preferably close to a target level, which can be perfectly modeled as the "safe leveling bandit" problem we described. Electrolyte disorders, common among intensive care unit patients, are one such example. When the blood sodium level falls below 135 milliequivalents per liter (mEq/L) or rises above 145 mEq/L, the patient suffers from hypo-/hyper-natremia, which has negative health consequences [50]. As a result, correct electrolyte dosing is critical to patient safety, and there is no agreement on how to assess the correct dosage for different patient characteristics. Another serious issue is high blood pressure. These are hypo-/hypertension events where the blood pressure deviates from the normal range and must be corrected. Patient characteristics are important in determining the blood pressure response to the therapeutic agent and should be considered during the dosing process [51].

To evaluate the performance of our algorithms, we run extensive *in silico* simulations on type 1 diabetes mellitus (T1DM) disease in Chapter 3. T1DM is characterized by insulin deficiency as a result of pancreatic *beta*-cell loss, and it can have serious adverse effects due to hypo-/hyper-glycemia (low and high blood glucose levels), which can result in hospitalization and even death [23]. As a result, T1DM patients must control their blood glucose levels by administering bolus insulin doses before meals. We attempt to optimize the dose recommendation process by learning to recommend better doses in a safe and efficient manner.

2.2 Federated Bandits

In Chapter 4, we study the federated multi-armed bandit (FMAB) problem introduced in [52]. In the FMAB problem, a group of clients plays the same MAB

game with different local models, which are *perturbed* versions of the global feedback model. The clients communicate with a parameter server over phases. At the end of each phase, the clients communicate their feedback regarding their local observations on local arms to a parameter server. The parameter server then aggregates the local feedback from clients and performs an arm elimination routine. The ultimate goal is to eliminate all suboptimal arms and find the globally optimal arm. Federated learning models are vulnerable to adversarial attacks such as model-update attacks, or data poisoning [53–55]. We study the FMAB problem in the presence of *Byzantine* clients who can send *arbitrarily* corrupted feedback to the parameter server and threaten the model performance and reliability [56]. Since the communication cost is a primary bottleneck of the federated learning system, we also incorporate it into the definition of cumulative regret. We borrow tools from robust statistics and propose a median-of-means-based estimator: Fed-MoM-UCB, to cope with the Byzantine clients. We show that if the Byzantine clients constitute at most half the cohort, it is possible to incur a cumulative regret on the order of $\mathcal{O}(\log T)$ with respect to an unavoidable error margin, including the communication cost between the clients and the parameter server. We analyze the interplay between the algorithm parameters, unavoidable error margin, regret, communication cost, and the arms' suboptimality gaps. We demonstrate Fed-MoM-UCB's effectiveness against the baselines in the presence of Byzantine attacks via experiments. In the remaining of this section, we provide background on different topics closely related to our work.

2.2.1 Federated Learning

Federated learning (FL) is a prominent paradigm aiming to alleviate some of the prevalent challenges of distributed machine learning. Among the most notable challenges FL seeks to address are *privacy* and *communication cost* [57]. Increasing volumes of data are being generated at edge devices such as mobile phones and tablets, which can be leveraged by machine learning algorithms. However, the sensitive nature of such data makes it infeasible to be stored in a centralized server [57]. FL removes this barrier by asking edge devices to train local models

using their data and only share local model updates with the parameter server instead of raw data.

The communication cost is a challenging bottleneck of the federated systems, especially when the number of clients is large [58]. Therefore, one needs to design an efficient communication scheme rather than asking the clients to communicate their updates to the parameter server after every iteration. Another challenge is the non-independent and identically distributed (non-IID) local datasets. Different clients' local datasets are likely to be drawn from non-IID distributions, thus not representative of the global population on their own. We refer the interested reader to [59] for a comprehensive survey on the advances and challenges in FL.

2.2.2 Robust Federated Learning

Different clients in FL make the learning process vulnerable for various reasons. The threats include adversarial attacks, communication line failures, and individual sensor malfunctions [60, 61], among others. Adversarial attacks can be in the form of data poisoning [53–55], model evasion attacks [62, 63], or model update poisoning [64] (e.g., Byzantine attacks [56, 65]). Among the most important properties of an FL framework is robustness against Byzantine attacks, where the attack is completely arbitrary, and there is no structure that can be leveraged in defense [56, 66]. It is known that even a *single* Byzantine client can make the model unreliable [55, 64, 66, 67]. Therefore, the study of robustness in FL systems has received considerable interest recently [59, 64, 68]. Next, we mention some of the works in robust FL that consider a supervised FL framework where the learning model is a neural network. [65, 69] propose geometric *median* estimators to aggregate the individual stochastic gradient descent (SGD) updates. [66] proposes a Byzantine-robust SGD algorithm called *Krum*, which uses the pairwise distances between individual updates and selects the update that is closest to its neighbors. [70] obtains optimal error rates for loss functions with different convexity properties under Byzantine attacks. It also proposes coordinate-wise median and trimmed-mean algorithms to combine individual gradient updates,

which are order-optimal for strongly convex loss functions. [71] shows that a Byzantine client can exploit the non-convex shape of the loss function to make Byzantine-resilient algorithms converge to local minimums, especially in high-dimensional settings. It proposes the *Bulyan* algorithm, which works with any Byzantine-resilient SGD algorithm by iteratively calling it to obtain a set of local gradients, which are then aggregated by a variant of trimmed mean [71, 72]. [73] attempts to identify malicious training data by investigating its effect on the performance of a reliable model trained on a clean dataset. [74] considers targeted poisoning attacks and proposes the *Auror* algorithm, which uses the masked features without accessing the individual training data to cluster and identify the malicious clients in an indirect collaborative learning setting. Federated learning with communication over noisy uplink and downlink channels and with imperfect channel state information have also been studied [75–77]. We refer the interested reader to [68] for an in-depth survey on the privacy and robustness in FL.

2.2.3 Federated Multi-armed Bandits

While recent studies on FL generally consider supervised learning problems, various distributed learning scenarios emerge naturally in different frameworks such as MABs. [78–80] propose decentralized multi-player MAB algorithms with collisions for opportunistic spectrum access in cognitive radio networks. [81–83] investigate a distributed MAB problem in which multiple agents play the same MAB game and collaborate to estimate arm outcomes across a communication network with delays. [84] analyzes a decentralized MAB problem in which local learners playing the same MAB game have potentially biased local reward distributions, resulting in suboptimal decisions. It proposes a UCB-based gossiping algorithm that allows a learner and its neighbors to communicate in a differentially private manner while ensuring sublinear regret [85]. [86] proposes a privacy-preserving and communication efficient algorithm for both the centralized and decentralized MAB framework, improving the privacy and regret-related costs in [84]. [87] studies the federated linear contextual bandit problem focusing

on differential privacy and proposes a LinUCB [11] based algorithm. [88] studies the same problem, and it proposes the Fed-PE algorithm, which incurs lower regret absent differential privacy guarantees.

In Chapter 4, we consider the FMAB framework introduced in [52]. We assume that there is an *unknown* global feedback model (for arm outcomes) and each client's local feedback model is a *perturbed* version of that global model. After playing an arm, each client makes a local observation for that arm generated by their local feedback models. The non-IID-ness challenge in FL surfaces as heterogeneous local models in the FMAB problem, which [52] refers to as *client sampling*. The client sampling adds an extra layer of uncertainty on top of the one arising from the *arm sampling*, which refers to the stochasticity of the observations.

[52] proposes Fed2-UCB for the FMAB problem which incurs order optimal regret ($\mathcal{O}(\log T)$). Each client pulls every *active* arm several times and sends only the *mean* outcomes to the parameter server as updates for better privacy preservation, which concludes a *phase*. At the end of a phase, the parameter server aggregates the updates from clients, and it eliminates the arms that are suboptimal with high probability. Then, until only one arm remains, it sends the clients an updated active arm set for the next phase. The FMAB framework with heterogeneous local models is effective at modeling a wide range of real-world problems, and [52] points out two exciting applications. The first is *cognitive radio*, in which a base station attempts to identify the best available channel on average for a given coverage area. The base station is fixed in place and cannot directly sample channel availabilities at different locations. However, it can use feedback from various devices spread across the coverage area to determine the best channel. Another application is online recommender systems [11]. The central server has no prior knowledge of global item popularity, but it can interact with different clients and gather information. Client interests vary and learning global item popularity while maintaining user privacy results in a federated bandit problem.

2.2.4 Byzantine Resilient FMAB

The study of the adversarial MAB problem dates back to 1995 [89], and it has garnered increasing interest recently. [90–97] study the standard and linear MAB problems for regret minimization and best arm identification settings under adversarial corruptions where the attack is structured to some extent (e.g., ϵ-contamination model [97]). [98] studies a setting where the reward distributions are heavy-tailed rather than being subgaussian and propose a "Robust UCB" algorithm. We study the FMAB problem [52] under Byzantine attacks [56, 59], where Byzantine clients in the cohort send arbitrarily corrupted model updates to the parameter server.

[99] studies a closely related problem to ours, where the objective is to identify the best arm in a federated setting. It proposes the "Federated Successive Elimination" (Fed-SEL) algorithm. [99] also investigates the scenario where some clients are Byzantine, and it proposes the robust version of the Fed-SEL algorithm. [100] analyzes a multi-agent MAB problem where the agents collaborate to minimize their cumulative regrets, with a subset of agents being malicious. However, both of the studies above assume that the arm outcomes are independent and identically distributed across different clients; that is, they do not consider the possible model heterogeneity arising from the *client sampling* discussed in Section 2.2.3.

2.3 Combinatorial MABs with Probabilistically Triggered Arms

Unlike the classical MAB framework, many sequential decision-making problems, ranging from online influence maximization to bundle recommendation, necessitate the simultaneous selection of multiple arms at each round. These and many others can be modeled as a combinatorial multi-armed bandit (CMAB) problem,

which is an extension of the original MAB problem. For instance, CMAB framework is useful in modeling organized and energy efficient activation of small cells in 5G networks [101]. At each round in the CMAB framework, the learner selects a *super-arm* (a subset of base arms) to observe the overall reward for that super-arm and individual outcomes for each base arm within [102–105]. This feedback is known as the *semi-bandit feedback*

Even the CMAB framework may fall short in modeling the problem in many real-world applications. The CMAB framework, for example, does not represent the probabilistic nature of the item list recommendation problem or influence maximization in a graph [106]. As a result, we are interested in a broader framework known as combinatorial multi-armed bandit with probabilistically triggered arms (CMAB-PTA), in which each base arm in a super-arm may trigger other arms based on its outcome [107]. If a user does not click on a recommended webpage, she may click on the one after that. At the end of the round, the reward is determined by all triggered base arms. For the CMAB-PTA framework, cumulative regret of order $\mathcal{O}(\log T)$ is achievable when the expected reward function has ℓ_∞ bounded smoothness property [107]. A particular setback of that bound is that it depends on $1/p^*$, where p^* is the minimum non-zero triggering probability over all base arms. [108] derives a bound free of $1/p^*$ by imposing stricter assumptions on the expected reward function, called triggering probability modulated bounded smoothness. Moreover, [108] shows that the dependence on $1/p^*$ is unavoidable in general.

In Chapter 5, we propose the "Combinatorial Gaussian Process Upper-Confidence Bound" algorithm (ComGP-UCB) for the CMAB-PTA framework and analyze its cumulative regret. Gaussian processes (GP) provide a strong and convenient framework for CMAB problems. The immediate benefit is that GPs can model the dependencies between the expected outcomes of different base arms, which are normally ignored in the classical MAB setting framework. Let $x_k \in \mathbb{R}^d$ denote the d-dimensional context (side-information) of an arm k. The additional context of the arm may assist the learner in making decisions. In an online recommendation problem, the context x_k may encode information on a news article's genre, length, and publishment date. We may infer whether a

user will click on a particular webpage recommendation based on her previous actions on similar recommendations ($x \in \mathbb{R}^d$). In contrast to other algorithms that assume independence between different base arm outcomes, a GP bandit updates its body of information about the other base arms via its kernel function, which models the similarity between different base arms. Furthermore, due to the analytical tractability and simple update equations of GPs, such models are simple to implement and analyze. We demonstrate through simulations that when the correlations between base arm outcomes are strong, ComGP-UCB outperforms the state-of-the-art methods Combinatorial TS (CTS) and Combinatorial UCB (CUCB).

Chapter 3

ESCADA: Efficient Safety and Context-Aware Dose Allocation for Precision Medicine

3.1 Introduction

In this chapter, we study the "safe leveling" problem discussed in Section 2.1. We propose a novel MAB algorithm using Gaussian processes and test its performance against the baselines through experiments. Our main contributions are as follows.

- We study an important problem in medicine that is rather overlooked and may be relevant in other domains.
- We formalize the "safe leveling" problem from a contextual MAB perspective via a suitable *regret* definition as the proxy performance metric. (in Section 3.2)
- Since our objective is keeping the outcomes close to a target rather than maximizing contrary to the usual MAB setting, we propose a novel acquisition method. (in Section 3.3)

- We design a safe exploration scheme for our acquisition function in Section 3.3 and derive high probability upper bounds on its regret along with safety guarantees. (in Section 3.4)
- We make extensive *in silico* simulations on type 1 diabetes mellitus (T1DM) disease. (in Section 3.5)

3.2 Problem Formulation

We denote by $[N]$ the set $\{1, \dots, N\}$, $\boldsymbol{z} \in \mathcal{Z}$ a context, and $d \in \mathcal{D}$ a dose, where both $\mathcal{Z}$ and $\mathcal{D}$ are compact and convex, and $\mathcal{D} = [0, \overline{D}]$. Here, $\overline{D} \in \mathbb{R}_+$ denotes the maximum possible value. We denote by $f : \mathcal{Z} \times \mathcal{D} \to \Omega$ the unknown function that maps context-dose pairs to the physiological variable of interest, where $\Omega = [0, \overline{T}]$. At each round $n \in [N]$, the learner first observes a context, $\boldsymbol{z}_n$, and recommends a dose, d_n, depending on the context. Then, it makes a noisy observation for the unknown function f as $y_n = f(\boldsymbol{z}_n, d_n) + \nu_n$, where ν_n are zero-mean i.i.d. Gaussian with known variance σ^2. The learner's task is to keep $f(\boldsymbol{z}_n, d_n)$ in a predetermined safe range and close to a target level, which we refer to as *safe leveling*. Through the lens of MABs, we use the following cumulative regret as the proxy performance metric along with a safety constraint.

$$\text{minimize} \quad R_N = \sum\nolimits_{n=1}^{N} |f(\boldsymbol{z}_n, d_n) - T| \tag{3.1}$$

$$\text{subject to} \quad T_{\min} \leq f(\boldsymbol{z}_n, d_n) \leq T_{\max} \quad \forall n \in [N] \ , \tag{3.2}$$

where $T_{\min},\ T_{\max} \in \mathbb{R}$ denote the lower and upper safety boundaries for f, respectively. $T \in (T_{\min} + \alpha,\ T_{\max} - \alpha)$ denotes the target value, where $\alpha > 0$. We have the non-zero α to ensure that the target value T is not exactly equal to the boundaries, which is leveraged in the analyses later. We assume $\forall \boldsymbol{z} \in \mathcal{Z}$, there exists $d_z^* \in \mathcal{D}$ such that $f(\boldsymbol{z}, d_z^*) = T$.

3.2.1 Regularity Assumptions

Without the existence of initial safe sets $S_0(\boldsymbol{z})$ for the context $\boldsymbol{z} \in \mathcal{Z}$ and certain smoothness properties on the function f, it is not possible satisfy the safety constraints in (3.2) [45]. Inspiring from the literature, we design a safe exploration scheme that leverages the smoothness properties of the objective function f to expand around the initial safe sets over rounds.

We denote by $\mathcal{X} = \mathcal{Z} \times \mathcal{D}$ the space of context-dose pairs. Let $k(\cdot, \cdot)$ be a positive definite kernel function on $\mathcal{X}$. We assume that f is a function from the *Reproducing Kernel Hilbert Space* (RKHS) corresponding to $k(\cdot, \cdot)$. In addition, we assume that f has bounded norm in this particular RKHS, i.e., $\|f\|_k < B_f$ [109]. This smoothness assumption makes f efficiently learnable. Precisely speaking, f is L-Lipschitz continuous w.r.t. kernel metric $q(\boldsymbol{x}, \boldsymbol{x'}) = \sqrt{k(\boldsymbol{x}, \boldsymbol{x}) - 2k(\boldsymbol{x}, \boldsymbol{x'}) + k(\boldsymbol{x'}, \boldsymbol{x'})}$, where $L = B_f$ [110]. Also, we denote by $q_{\boldsymbol{z}}(d, d') \coloneqq q((\boldsymbol{z}, d), (\boldsymbol{z}, d'))$ to keep the notation uncluttered later.

Next, we introduce a discretization of the dose space $\mathcal{D}$ for every $\boldsymbol{z} \in \mathcal{Z}$ as $\bar{\mathcal{D}}_{\boldsymbol{z}} \coloneqq \{d_i \in \mathcal{D} \mid i \in \{1, \dots, k\}\}$, where $d_1 = 0$, $d_i > d_j$ for $i > j$, $q_{\boldsymbol{z}}(d_i, d_{i+1}) = \lambda/2L$, $q_{\boldsymbol{z}}(\overline{D}, d_k) < \lambda/2L$, and $\lambda > 0$ is the discretization parameter. We assume that an initial safe set of discretized doses $S_0(\boldsymbol{z})$ is available for each context $\boldsymbol{z} \in \mathcal{Z}$. These regularity assumptions allow us to model the unknown objective function f with a Gaussian process (GP) and analyze its regret and safety guarantees [22]. A GP can be thought as a distribution over functions which is completely described by its mean, $\mu(\cdot)$, and covariance, $k(\cdot, \cdot)$, functions. Moreover, if we assume a GP prior over f, then, after observing a set of noisy evaluations $\boldsymbol{y}_N = [y_1 \dots y_N]^T$ at points $A_N = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, the posterior over f is a GP again, with the following covariance and mean functions,

$$
\begin{aligned}
k_N(\boldsymbol{x}, \boldsymbol{x'}) &= k(\boldsymbol{x},\ \boldsymbol{x'}) - \boldsymbol{k}_N(\boldsymbol{x})^T (\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_N(\boldsymbol{x'}) \\
\sigma_N^2(\boldsymbol{x}) &= k_N(\boldsymbol{x}, \boldsymbol{x}) \\
\mu_N(\boldsymbol{x}) &= \boldsymbol{k}_N(\boldsymbol{x})^T (\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_N \ ,
\end{aligned}
$$

where $\boldsymbol{k}_N(\boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}) \dots k(\boldsymbol{x}_N, \boldsymbol{x})]^T$ and $\boldsymbol{K}_N$ is the positive definite kernel

matrix $[k(\boldsymbol{x}, \boldsymbol{x}')]_{\boldsymbol{x},\boldsymbol{x}' \in A_N}$.

3.3 ESCADA Algorithm

In this section, we propose the Efficient Safe and Context Aware Dose Allocation (ESCADA) algorithm for the "safe leveling" task desribed in Sections 2.1 and 3.2. ESCADA consists of two main parts: (**i**) an acquisition function, called TACO: TArget-based COnfident-acquisition, (**ii**) a safety mechanism to render TACO safe. Algorithm 1 and Figure 3.1 describe the safe dose recommendation strategy of ESCADA. ESCADA's recommendation procedure can also be interpreted to domain experts via the flowchart in Figure 3.1 as opposed to black-box models [111].[1]

3.3.1 Acquisition Strategy

Our objective is to keep the objective function f close to a target level rather than maximizing it. Therefore, we cannot directly use the standard MAB algorithms (e.g., GP-UCB, TS). We propose TACO, a novel acquisition method tailored for the safe leveling task explained in Section 3.2. At each round n, TACO uses the confidence bounds of doses $d \in \mathcal{D}$ for z_n derived from the GP prior as,

$$l_n(\boldsymbol{z}_n, d) = \mu_{n-1}(\boldsymbol{z}_n, d) - \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d) \tag{3.3}$$

$$u_n(\boldsymbol{z}_n, d) = \mu_{n-1}(\boldsymbol{z}_n, d) + \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d) \ . \tag{3.4}$$

We define β_n later in Section 3.4 in a way that the true value of f resides in the confidence interval with high probability (see Lemma 3.1). Next, we leverage the Lipschitz continuity of f to form the final lower and upper confidence bounds for

[1] Flowchart assumes that GP-induced confidence intervals are correct, i.e., the event $\mathcal{E}$ in Lemma 3.1 holds.

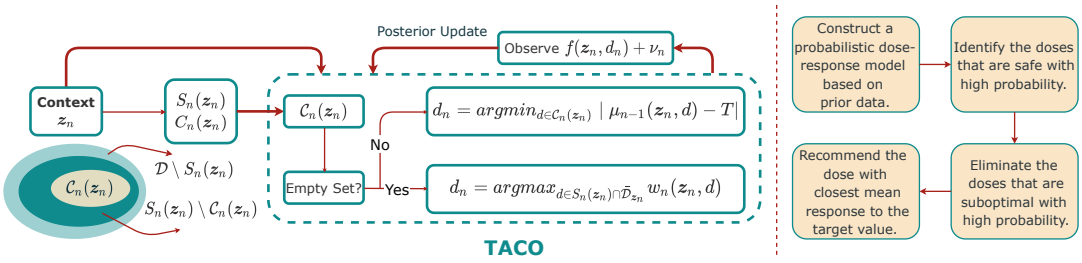


Figure 3.1: *ESCADA Algorithm Description (left).* Upon observing a context $\boldsymbol{z}_n$ in round n, TACO forms the set $\mathcal{C}_n(\boldsymbol{z}_n) \subseteq S_n(\boldsymbol{z}_n)$ after eliminating the doses that are suboptimal with high probability. If $\mathcal{C}_n(\boldsymbol{z}_n) \neq \emptyset$, it recommends the dose whose mean response is closest to the target T. If $\mathcal{C}_n(\boldsymbol{z}_n) = \emptyset$, it recommends the dose with the widest confidence interval in $S_n(\boldsymbol{z}_n) \cap \bar{\mathcal{D}}_{\boldsymbol{z}_n}$. *Flowchart (right).* A simple interpretation of the dose allocation process intended for domain experts.

every $d \in \mathcal{D}$ as,

$$\bar{l}_n(\boldsymbol{z}_n, d) \coloneqq \max\{l_n(\boldsymbol{z}_n, d), l_n(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d, d')\}$$
$$\bar{u}_n(\boldsymbol{z}_n, d) \coloneqq \min\{u_n(\boldsymbol{z}_n, d), u_n(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d, d')\} \ ,$$

where $d' = \text{argmin}_{d^* \in \bar{\mathcal{D}}_{\boldsymbol{z}_n}} q_{\boldsymbol{z}_n}(d, d^*)$. We denote by $C_n(\boldsymbol{z}_n, d) = [\bar{l}_n(\boldsymbol{z}_n, d), \bar{u}_n(\boldsymbol{z}_n, d)]$ the confidence interval of a dose $d \in \mathcal{D}$ in round n, and by $C_n(\boldsymbol{z}_n) = \{C_n(\boldsymbol{z}_n, d) \mid d \in \mathcal{D}\}$ the set of all confidence intervals. Finally, we form the confidence widths for each dose $d \in \mathcal{D}$ as $w_n(\boldsymbol{z}_n, d) = \bar{u}_n(\boldsymbol{z}_n, d) - \bar{l}_n(\boldsymbol{z}_n, d)$.

TACO's recommendation procedure can be explained in three main parts. It takes a dose set $\mathcal{D}_n$ as input, from which the recommendation will be made from. We note that if the dose set $\mathcal{D}_n$ consists of safe doses, then TACO's recommendations will automatically be safe as well. We handle the safety in the next section separately. TACO first observes the context $\boldsymbol{z}_n$ for that round. Then, it takes the following steps,

1. Identify the dose set $\mathcal{C}_n \subseteq \mathcal{D}_n$ whose elements' confidence intervals contain the target value, T.

2. If $\mathcal{C}_n \neq \emptyset$, recommend the dose in $\mathcal{C}_n$ with closest mean response to the target value T.

Algorithm 1 ESCADA algorithm

Inputs: $\mathcal{X} = \mathcal{Z} \times \mathcal{D}$; GP prior (μ_0, k, σ);
$T_{\min}$; $T_{\max}$; T; L; $S_0(\boldsymbol{z}),\ \forall \boldsymbol{z} \in \mathcal{Z}$
for $n = 1, 2, \ldots$ **do**
  Observe context $\boldsymbol{z}_n$; form $C_n(\boldsymbol{z}_n)$
  Update $S_n(\boldsymbol{z}_n)$ via (3.5)
  $d_n \leftarrow$ TACO$(C_n(\boldsymbol{z}_n), S_n(\boldsymbol{z}_n))$
  Observe $y_n = f(\boldsymbol{z}_n, d_n) + \nu_n$
  Update GP posterior
end for

Subroutine: TACO
Inputs: $C_n(\boldsymbol{z}_n); \mathcal{D}_n$
$\mathcal{C}_n = \{d \in \mathcal{D}_n \mid T \in C_n(\boldsymbol{z}_n, d)\}$
if $\mathcal{C}_n \neq \emptyset$ **then**
  $d \leftarrow \underset{d' \in \mathcal{C}_n}{\arg\min}\ |\mu_{n-1}(\boldsymbol{z}_n, d') - T|$
else
  $d \leftarrow \underset{d' \in \mathcal{D}_n \cap \bar{\mathcal{D}}_{\boldsymbol{z}_n}}{\arg\max}\ w_n(\boldsymbol{z}_n, d)$
end if
return d

3. If $\mathcal{C}_n = \emptyset$, recommend the dose in $\mathcal{D}_n \cap \bar{\mathcal{D}}_{\boldsymbol{z}_n}$ with the widest confidence interval.

TACO first eliminates the doses which are "suboptimal" with high probability. First step includes the elements of *exploration* and *exploitation* together. A dose with a *mean* response close to the target value T can be selected (exploitation) since its confidence interval is likely to contain the target value T. At the same time, under-explored doses with wide confidence intervals that may contain the target can still be selected (exploration). TACO executes the final step if there is no dose in $\mathcal{D}_n$ whose confidence interval contains the target value. In that case, TACO employs a pure exploration strategy to identify the optimal doses. TACO is *efficient* in the sense that it treats exploration as a proxy objective –in the third step– only when all the feasible doses $(\mathcal{D}_n)$ are suboptimal with high probability.

3.3.2 Safety Awareness

Inspiring from the previous works on safe exploration with Gaussian processes [45, 46], we design a safe exploration scheme to tend to the two-sided safety constraint in (3.2). We denote by $S_n(\boldsymbol{z}_n)$ the safe set for the context $\boldsymbol{z}_n$ and implement the following safe expansion rule to derive $S_n(\boldsymbol{z}_n)$ each round,

$$S_n(\boldsymbol{z}_n) = S_{n-1}(\boldsymbol{z}_n) \cup \left\{ \bigcup_{d \in S_{n-1}(\boldsymbol{z}_n)} \{d' \in \mathcal{D} \mid \bar{l}_n(\boldsymbol{z}_n, d) - Lq_{\boldsymbol{z}_n}(d, d') \geq T_{\min} \right. \\ \left. \wedge\ \bar{u}_n(\boldsymbol{z}_n, d) + Lq_{\boldsymbol{z}_n}(d, d') \leq T_{\max}\} \right\} . \tag{3.5}$$

At each round n, TACO recommends a dose from $\mathcal{D}_n = S_n(\boldsymbol{z}_n)$. $S_n(\boldsymbol{z}_n)$ only contains the doses for which f resides in the target interval almost certainly (see Theorem 3.1).

Next, we define the ϵ-reachability operator R_ϵ and the ϵ-reachable safe set for $\boldsymbol{z} \in \mathcal{Z}$, $\bar{R}_\epsilon(S_0(\boldsymbol{z}))$ (where $\epsilon > 0$ is the error margin accounting for the uncertainty in measurements) as in [45],

$$R_\epsilon(S_0(\boldsymbol{z})) \coloneqq S_0(\boldsymbol{z}) \cup \Big\{ d \in \mathcal{D} \mid \exists d' \in S_0(\boldsymbol{z}),\ f(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d') - \epsilon \geq T_{\min} \\ \wedge\ f(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') + \epsilon \leq T_{\max} \Big\} , \tag{3.6}$$

and $\bar{R}_\epsilon(S_0(\boldsymbol{z})) \coloneqq \lim_{n \to \infty} R^n_\epsilon(S_0(\boldsymbol{z}))$. R^n_ϵ denotes the n-time reachability operator, which calls R_ϵ n times using the previous step's output. $\bar{R}_\epsilon(S_0(\boldsymbol{z}))$ represents the subset of $\mathcal{D}$ that can be identified as safe for the context $\boldsymbol{z}$ by observing f up to a statistical certainty restricted by ϵ.

3.4 Theoretical Analyses

Consider a sequence of patient contexts $\bar{\boldsymbol{z}} = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N]$. Let $\mathbb{X}_N = X_1 \times \dots X_N$ denote the space of all context-admissible recommendation pairs, where $X_n =$

$\boldsymbol{z}_n \times \mathbb{D}_n$, and $\mathbb{D}_n \subseteq \mathcal{D}$ is the admissible dose space for $\boldsymbol{z}_n$. For a given sequence of context-recommendation set A, let $\boldsymbol{y}_A$ denote the $|A|$-dimensional vector containing corresponding noisy evaluations of f. The quantity governing our regret bounds after N rounds in this scenario is a volatility-adapted maximum information gain term, $\gamma_N^{vol} = \max_{A \subset \mathbb{X}_N} I(\boldsymbol{y}_A; \boldsymbol{f}_A)$, where $\boldsymbol{f}_A = [f(\boldsymbol{x})]_{\boldsymbol{x} \in A}$ and $I(\boldsymbol{y}_A; \boldsymbol{f}_A)$ is the mutual information between f and observations at points in A. In the general setting where there is not a fixed context sequence, we have $\gamma_N = \max_{A \subset \mathcal{X}^N} I(\boldsymbol{y}_A; \boldsymbol{f}_A)$. Note that since $\mathbb{X}_N \subseteq \mathcal{X}^N$, we have $\gamma_N^{vol} \leq \gamma_N$. Explicit bounds on γ_N depending on N are available in the literature [10, 112]. In this section, we first derive a high probability upper bound on the cumulative regret of TACO for a fixed context sequence without safety constraints. Then, we bound the regret of ESCADA in a single context scenario with safety constraints. For the former, we have $\bar{\boldsymbol{z}}_1 = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N]$, and $\mathbb{D}_n = \mathcal{D}$, and we denote the upper bound on the information gain term (see Lemma 3.2) by γ_N^{vol1}. For the latter, we have $\bar{\boldsymbol{z}}_2 = [\boldsymbol{z} \dots \boldsymbol{z}]$, $\mathbb{D}_n = S_n(\boldsymbol{z})$, and we denote the upper bound on the information gain term by γ_N^{vol2}. We also prove that every dose recommended by ESCADA is safe with high probability (w.h.p.). Detailed derivations for the proofs can be found in Appendix A.

First, we mention two standard results used in the analysis. Lemma 3.1 shows that f is contained in the GP-induced confidence intervals w.h.p. and Lemma 3.2 expresses the information gain in terms of predictive variances.

Lemma 3.1. *(Theorem 1 in [113]) Pick $\delta \in (0, 1)$, and define $\beta_n = 2L^2 + 300\gamma_n \log^3(n/\delta)$, where L is the Lipschitz constant. Let $\mathcal{E} = \{|\mu_{n-1}(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{x}), \forall n \in \mathbb{N}, \forall \boldsymbol{x} \in \mathcal{X}\}$. We have $\mathbb{P}\{\mathcal{E}\} \geq 1 - \delta$.*

Lemma 3.2. *(Lemma 5.3 in [10]) The information gain for the points selected can be expressed in terms of the predictive variances. If $\boldsymbol{f}_N = (f(\boldsymbol{x}_n))$, $I(\boldsymbol{y}_N; \boldsymbol{f}_N) = \frac{1}{2} \sum_{n=1}^{N} \log(1 + \sigma^{-2} \sigma_{n-1}^2(\boldsymbol{x}_n))$.*

The following theorem provides a safety guarantee for ESCADA under the high probability event $\mathcal{E}$ in Lemma 3.1. The proof depends on an inductive argument on the safe sets constructed by ESCADA, starting from the initial safe sets.

Theorem 3.1. *All doses recommended by ESCADA are safe, that is,* $T_{\min} \leq f(\boldsymbol{z}_n, d_n) \leq T_{\max}\ \forall n \in [N]$*, with at least* $1-\delta$ *probability.*

We proposed a novel acquisition function in Section 3.3.1, TACO, for the generic problem structure described in Sections 2.1 and 3.2. Theorem 3.2 provides an upper-bound on the regret of TACO, without any safety constraints in place.

Theorem 3.2. *Define* β_n *as in Lemma 3.1 and let* $C \coloneqq 8/\log(1+\sigma^{-2})$*. Cumulative regret of TACO for a fixed context sequence is upper-bounded as follows,*

$$\mathbb{P}\big\{R_N \leq \sqrt{CN\beta_N\gamma_N^{vol1}}\big\} \geq 1-\delta\ .$$

Next, we introduce a new concept, *safe path*.

Definition 3.1. (Safe Path) For a fixed context $\boldsymbol{z} \in \mathcal{Z}$, we say that there exists a safe path between two doses $d_1, d_2 \in \mathcal{D}$ if the following is satisfied,

$$\eta(d_1,d_2) = \min\left(\min_{d\in[d_1,d_2]}\big(T_{\max}-\epsilon-f(\boldsymbol{z},d)\big), \min_{d\in[d_1,d_2]}\big(f(\boldsymbol{z},d)-T_{\min}-\epsilon\big)\right) > 0\ . \tag{3.7}$$

Definition 3.1 simply states that if there exists a safe path between two doses d_1 and d_2, then there is no dose violating or exactly at the safety constraints between them. That is, $f(d) \in (T_{\min}+\epsilon+\eta(d_1,d_2), T_{\max}-\epsilon-\eta(d_1,d_2))$ for all $d \in [d_1,d_2]$. Next, we give the regret bound for ESCADA, which uses TACO as its acquisition function and employs the safety mechanism in Section 3.3.2. We assume a fixed context scenario and show that the safety constraints result in at most a constant addition to the regret.

Theorem 3.3. *If there exists a safe path between at least one dose* $d \in S_0(\boldsymbol{z})$ *and* $d^*_{\boldsymbol{z}}$*, and we have* $q_{\boldsymbol{z}}(d_1,d_2) = K(|d_1-d_2|)$ *for some monotonically increasing mapping* $K : \mathbb{R}^+ \to \mathbb{R}^+$ *and for all* $d_1, d_2 \in \mathcal{D}$*, then the cumulative regret of ESCADA in a safety constrained single context (*$\boldsymbol{z}$*) scenario can be upper-bounded by setting the discretization parameter* $\lambda < \epsilon$ *as follows,*

$$\mathbb{P}\big\{R_N \leq \sqrt{CN\beta_N\gamma_N^{vol2}} + \overline{T}N_z\big\} \geq 1-\delta\ ,$$

where $N_z \in \mathbb{N}$ *is a constant independent of* N*.*

Note that since $f(\boldsymbol{z}, d_{\boldsymbol{z}}^*) = T$ and $T \in (T_{\min} + \alpha, T_{\max} - \alpha)$, one must ensure that $\alpha > \epsilon$ for the possibility of a safe path between some $d \in S_0(\boldsymbol{z})$ and $d_{\boldsymbol{z}}^*$ at the first place.

The assumption that $q_{\boldsymbol{z}}(d_1, d_2) = K(|d_1 - d_2|)$ for a monotonically increasing mapping K holds in our working example where the blood glucose response to insulin dose can be characterized by the carbohydrate factor (CF). That is, if we let $L \gg \text{CF}$, then we have $f(\boldsymbol{z}, d_1) - f(\boldsymbol{z}, d_2) \leq L|d_1 - d_2|$ for $d_1, d_2 \in \mathcal{D}$ and $q_{\boldsymbol{z}}(d_1, d_2) = |d_1 - d_2|$. Moreover, this is the case for a variety of widely used kernel induced distance metrics. For the squared exponential kernel $k(\alpha, \beta) = \exp\left(-\|\alpha - \beta\|^2 / 2\sigma^2\right)$, we have (see Section 3.2),

$$q_{\boldsymbol{z}}(d_1, d_2) = \sqrt{2 - 2\exp\left(-|d_1 - d_2|^2/\sigma^2\right)}\,. \tag{3.8}$$

Similarly, the Laplacian kernel (also a radial-basis function) $k(\alpha, \beta) = \exp\left(-\|\alpha - \beta\| / \sigma\right)$ induces the metric,

$$q_{\boldsymbol{z}}(d_1, d_2) = \sqrt{2 - 2\exp\left(-|d_1 - d_2|/\sigma\right)}\,, \tag{3.9}$$

where both (3.8) and (3.9) are monotonically increasing in $|d_1 - d_2|$. Theorems 3.2 and 3.3 constitute the non-incremental parts in our analysis as they provide explicit regret guarantees for a novel problem structure and acquisition strategy, both with and without safety constraints for a *compact* and *convex* action set. To generalize the bound in Theorem 3.3 to mixed context scenarios, one needs to impose further assumptions on the regularity of context arrivals over time. We provide experimental results on mixed context scenarios in Section 3.5 and show that the inter-contextual information transfer actually improves the performance as expected.

3.5 Experimental Results

3.5.1 Experimental setup

Real-world online experimentation in clinical settings is risky, and it has ethical challenges [114–117]. When the patient group is homogeneous, some of the previous studies on dose-finding clinical trials validate their methods either through experiments in synthetic environments or by using external algorithms to fit a dose-response model to real-world data [21, 118, 119]. Such algorithms are not applicable in our case as they assume a shared dose-response model among patients, whereas we aim to learn *individualized* representations. We make *in silico* experiments using the open-source implementation [120] of the U.S. FDA approved University of Virginia (UVa)/PADOVA T1DM simulator [121], which is the most frequently used framework in blood glucose control studies [122–129]. It includes 30 virtual patients, each with their own metabolic parameters: 10 adults, 10 adolescents, and 10 children. The simulator uses a complicated model with differential equations and patient characteristics to calculate a patient's postprandial blood glucose (PPBG) response for (meal event, bolus insulin dose) combinations [121]. We experiment with all 30 virtual patients that come with the simulator. We use separate models for each patient, and model the different meal events as contexts (i.e., $\boldsymbol{z} \in \mathcal{Z}$ in Section 3.2). We also assess ESCADA's performance against a clinician for five virtual adult patients to evaluate the success and potential of ESCADA as a supplemental tool in the clinical setting and to provide some external validation.

3.5.2 Performance metrics

When the PPBG level of the patient drops below 70 mg/dl (or exceeds 180 mg/dl), hypoglycemia (hyperglycemia) events occur, which may lead to life-threatening conditions [23]. We set the target blood glucose (BG) level to 112.5 mg/dl [130]. Our main goal is to recommend insulin doses that minimize the cumulative *regret*

while ensuring patient *safety* at all times (with high probability). By (3.1), minimizing the cumulative regret is equivalent to recommending doses that keep the PPBG close to the target BG. Ensuring safety requires not recommending doses that trigger hypoglycemia and hyperglycemia events (see (3.2)). We evaluate an algorithm's performance by combining its regret, hypoglycemia, hyperglycemia frequencies (error frequencies), and glycemic risk indices. Glycemic risk indices are low blood glycemic index (LBGI), high blood glycemic index (HBGI), and risk index (RI $\coloneqq$ LBGI + HBGI). LBGI (HBGI) characterizes the risk of hypoglycemia (hyperglycemia) events for a patient in the long term [130]. A successful algorithm should incur low cumulative regret along with small risk index values by *safely* and *efficiently* learning to suggest better insulin doses. We also discuss the competing algorithms' consistency as inexplicable variations in treatment are undesirable [131]. To be more specific, a consistent algorithm should recommend the same insulin dose to a patient for the same meal event for a given history. A meal event consists of two components: (carbohydrate intake and fasting blood glucose). We create different meal events through uniform sampling to create an ensemble of different scenarios. We sample carbohydrate intake for each meal event from $[20,\ 80]$ g and fasting blood glucose from $[100,\ 150]$ mg/dl.

3.5.3 Single meal event (SME) scenario

In this section, we recommend insulin doses to a patient for the same meal event, assuming the insulin dose is taken directly before the meal. This setup is useful for two reasons: (**i**) it evaluates the algorithms' performance in the standard *non-contextual* MAB setting, (**ii**) it provides a simple benchmark to understand performance metrics and later compare them to the contextual setup. After recommending a dose, we observe the patient's PPBG level and update the algorithms' parameters before the next recommendation. The aim is to optimize the PPBG 150 minutes after the meal. In a single run, we recommend 15 consecutive doses for a meal event. This experiment is repeated 30 times with 30 different meal events for each patient. We then report the mean and the standard deviation results from all $(30 \times 30 = 900)$ runs.

3.5.4 Multiple meal events (MME) scenario

In this section, we recommend insulin doses to a single patient for a series of meal events. This scenario demonstrates how learning from different meal events can help adjust recommendations for each other. We run simulations for all 30 patients using the same 30 meal events created in the SME scenario. In a round-robin fashion, we make consecutive recommendations for different meal events and recommend a total of 15 doses for each meal event. To be more specific, after recommending a dose for a meal event, we recommend doses for the remaining 29 meal events and measure the PPBGs before returning to the first meal event. We then average the results from all 30 runs (one per patient). The difference in this part from the SME scenario is that we make insulin dose recommendations to a patient for different meal events in a *single run* (i.e., the setting here is *contextual*), whereas, in the SME scenario, we recommend doses to patients for the same meal event in a single run (i.e., *non-contextual*). This setup demonstrates how information gained from a context can aid in decision-making in various other contexts. Because of this contextual knowledge transfer, our algorithm can adapt to intra- and inter-daily variability in meal events.

3.5.5 Algorithms

We simulate ESCADA and TACO (i.e., without the safety mechanism). Besides, we propose a Thompson sampling (TS)-based algorithm and its safe version (STS), which operate as follows: TS samples a PPBG function from the posterior GP in each round and recommends the dose that achieves the PPBG closest to the target BG. STS implements the safe exploration strategy in Section 3.3 and uses TS as the acquisition function. In the final part, we implement the GP-UCB algorithm in [10] using three different "reward" functions in Figure 3.3a and compare it to our acquisition functions TACO and TS. We use two versions of dose calculators as baselines, whose details are given below.

3.5.6 Dose calculators

Because they are transparent and interpretable, rule-based bolus insulin dose calculators are widely used in diabetes care [132]. We use them to initialize the safe dose set (i.e., $S_0(\boldsymbol{z})$) for patient and meal event pairs. A rule based-calculator recommends a bolus insulin dose via a simple equation using carbohydrate intake, fasting blood glucose, and patient-specific parameters. We use the one from [133], which is given below.

$$\text{Bolus-insulin Dose} = \Big(\frac{\text{CHO}}{\text{ICR}} + \frac{\text{G}_M - \text{G}_T}{\text{CF}}\Big)^+ \tag{3.10}$$

where $a^+ \coloneqq \max\{0, a\}$, CHO (g) is the carbohydrate intake, ICR (g/U) is the insulin-to-carbohydrate ratio, and CF (mg/dl/U) is the insulin correction factor. G_M and G_T (mg/dl) denote the preprandial BG level and the postprandial target BG levels, respectively. To avoid hypoglycemia and hyperglycemia events, ICR and CF values must be precisely tuned for each patient, which can be difficult in practice. Even if ICR and CF values are tuned correctly, rule-based calculators (e.g., (3.10)) discount other patient characteristics which may affect PPBG. Research shows that the correction doses constitute 9% of the patients' total daily insulin dose intake due to the calculator's failure [132].

We consider two scenarios. In the first, we employ a calculator setting that occasionally fails to provide safe dose recommendations and forego the assumption that an initial safe set, S_0, is always available. We use tuned calculators for each patient in the second setup to ensure that S_0 is always available.

3.5.7 Discussion of results

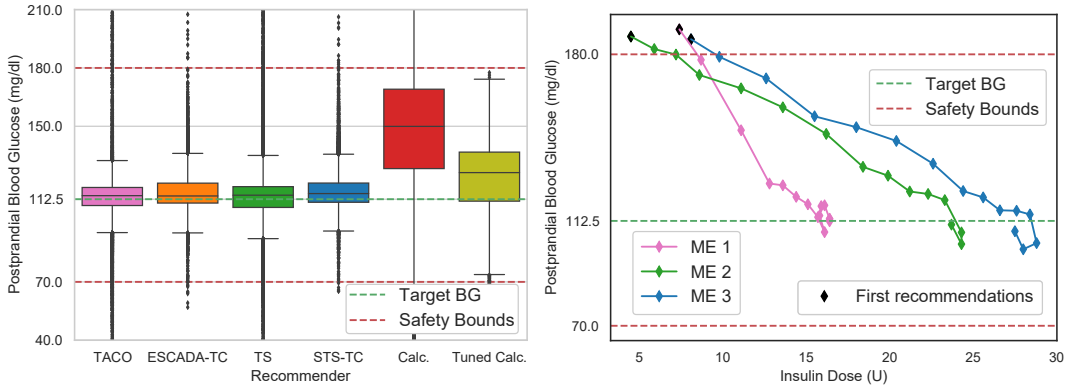
3.5.7.1 Safety

Ensuring patient safety is pivotal. Theorem 3.1 shows that ESCADA recommends safe doses with high probability when an initial safe dose set is available.

Table 3.1: "-TC" indicates that tuned calculator was used. Target PPBG is $T = 112.5$ mg/dl. "PPBG" column is averaged over all recommendations. "Hyper" and "Hypo" columns denote the hyperglycemia and hypoglycemia event frequencies, respectively, averaged over all patients. Similarly, HBGI and LBGI are averaged over all patients. We report (`mean` ± `stdandard deviation`).

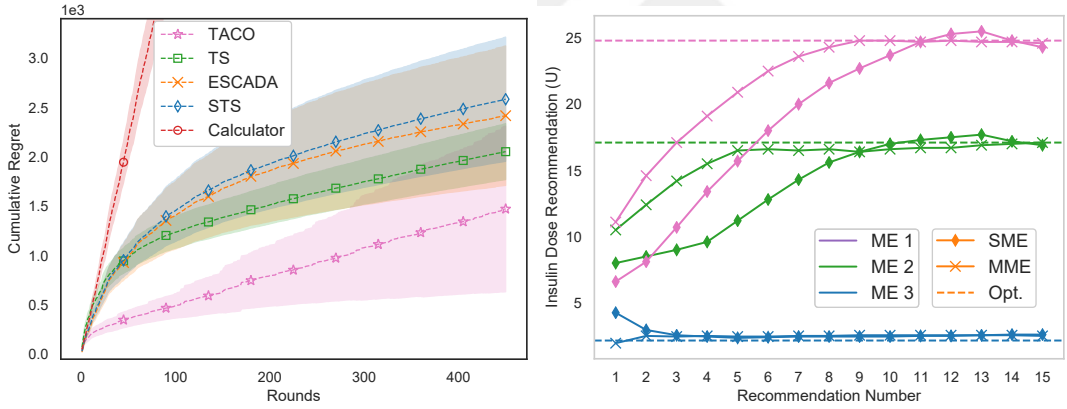
	Algorithm	PPBG	Hyper	Hypo	HBGI	LBGI
	Calc.	144.0 ± 39.5	$.143 \pm .217$	$.0614 \pm .189$	3.84 ± 3.41	1.37 ± 3.95
	Tuned Calc.	123.7 ± 18.1	0	$.0021 \pm .010$	0.83 ± 0.66	0.24 ± 0.47
SME	TS	119.8 ± 42.2	$.046 \pm .029$	$.0216 \pm .028$	1.52 ± 1.25	1.01 ± 2.31
	TACO	121.7 ± 50.4	$.049 \pm .032$	$.0175 \pm .019$	1.89 ± 1.63	0.52 ± 0.46
	STS	121.6 ± 24.8	$.031 \pm .063$	$.0029 \pm .010$	1.07 ± 1.33	0.15 ± 0.23
	ESCADA	122.2 ± 20.0	$.015 \pm .030$	$.0031 \pm .008$	0.77 ± 0.82	0.11 ± 0.24
	STS-TC	**117.1 ± 11.9**	**.002 ± .004**	**.0004 ± .001**	**0.28 ± 0.24**	**0.05 ± 0.05**
	ESCADA-TC	**116.1 ± 12.5**	**.002 ± .004**	**.0007 ± .003**	**0.26 ± 0.21**	**0.07 ± 0.09**
MME	GP-UCB-1	124.1 ± 87.0	$.179 \pm .050$	$.2618 \pm .202$	5.43 ± 2.26	15.4 ± 22.5
	GP-UCB-2	103.7 ± 59.4	$.080 \pm .060$	$.2873 \pm .254$	2.21 ± 1.58	16.0 ± 27.0
	GP-UCB-3	111.0 ± 32.6	$.022 \pm .010$	$.0648 \pm .076$	0.73 ± 0.33	3.45 ± 5.17
	TS	**112.4 ± 14.4**	**.003 ± .003**	$.0107 \pm .011$	**0.16 ± 0.18**	0.53 ± 0.95
	TACO	**113.7 ± 19.2**	**.006 ± .031**	**.0010 ± .002**	**0.29 ± 1.18**	**0.07 ± 0.04**
	STS	116.5 ± 12.5	**.004 ± .015**	**.0007 ± .002**	**0.32 ± 0.49**	**0.05 ± 0.05**
	ESCADA	116.9 ± 13.1	**.006 ± .017**	**.0005 ± .002**	**0.34 ± 0.55**	**0.04 ± 0.04**

However, the initially provided set may not always be safe in reality due to calculator or clinician mistakes. We simulate two scenarios when an initial safe set is almost always available and not. For the latter, Table 3.1 shows that the error frequencies of ESCADA are not zero. We expect that error since the calculator fails to consistently provide safe doses in the beginning. However, ESCADA yields significantly lower error frequencies and risk index values than the calculator. That improvement stems from ESCADA's ability to gradually identify and recommend safe doses, even when initially misdirected. We plot consecutive dose recommendations by ESCADA in SME scenario for three different meal events in Figure 3.2b. For each of these meal events, rule-based calculator fails to provide safe doses in the beginning. Notwithstanding, ESCADA expands its safe set in the right direction and eventually recommends safe doses. Figure 3.2a and Table 3.1 confirm the safety mechanism's effectiveness as ESCADA and STS yield significantly better safety metrics than the unsafe algorithms, TACO and TS, especially for hypoglycemia. Next, we tune the calculator parameters for each patient so that it provides an initial safe set almost always (Tuned Calc., Table 3.1). Figure 3.2a and Table 3.1 show that ESCADA-TC and STS-TC yield remarkably lower error frequencies and risk indices with better PPBG distributions.



(a) PPBG distribution boxplots in SME scenario. "-TC" suffix indicates that the tuned calculator is used.

(b) Consecutive recommendations to three meal events with *unsafe* $S_0(\boldsymbol{z})$ in SME scenario.



(c) The cumulative regrets averaged over all 30 patients in MME scenario ($\pm$ 0.25 std. dev.).

(d) Consecutive recommendations for three meal events (ME) in SME and MME scenarios.

Figure 3.2

3.5.7.2 Regret

Minimizing the regret is equivalent to recommending doses that lead to PPBG values close to the target BG by (3.1). We observe from Figures 3.2a and 3.2c, and Table 3.1 that ESCADA(-TC) and STS(-TC) significantly outperforms the (tuned) calculator. Figure 3.2c shows that TACO and TS incur lower cumulative regrets than ESCADA and STS in the MME scenario. That is a natural trade-off between safety and regret since the safety mechanism restricts the allocation of a dose before it is identified as safe. Therefore, a safe algorithm yields higher regret

when the initial safe set is far from the optimal dose.

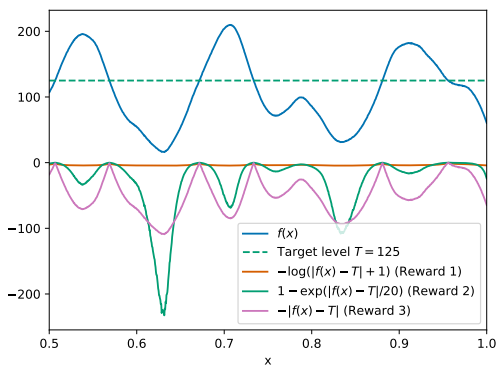
3.5.7.3 Inter-contextual information transfer

We investigate the efficiency of GP-induced smoothness in *transferring* information between different contexts. We mark an evident advancement in PPBG distributions and safety metrics in the MME scenario compared to the SME scenario in Table 3.1. Examining Figure 3.2d, we observe that ESCADA expands the safe dose set and identifies the optimal dose faster in the MME scenario. Remember that ESCADA recommends doses for different meal events between two consecutive recommendations for the same meal event in the MME scenario. That is, the information gained from a context improves the performance for other contexts. Besides, we observe significant advances in the safety metrics of TACO and TS in the MME scenario as well.

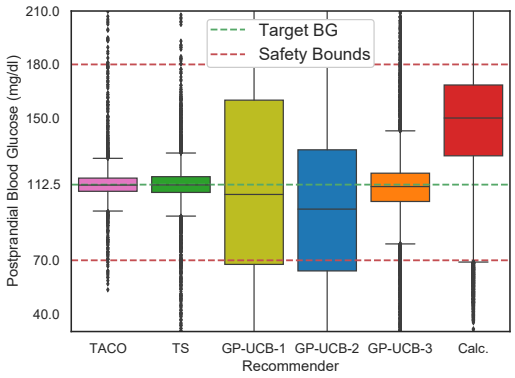
3.5.7.4 Comparison with GP-UCB

Our objective is to keep $f(x)$ close to a target level T while satisfying safety ($f(x) \in [T_{\min}, T_{\max}]$). In Section 3.2.1, we discuss the conditions for $f(x)$ to be efficiently learnable by a GP and support our modeling choice with discussions after Theorem 3.3. If the goal was to maximize $f(x)$, we could use the GP-UCB algorithm in [10]. However, we have to define *pseudo-rewards* to maximize such as $-\log(|f(x) - T| + 1)$ that are monotonically decreasing with $|f(x) - T|$. We present three such reward functions in Figure 3.3a. We compare our acquisition functions, TACO and TS, against the adaptations of the GP-UCB algorithm for three different reward functions in Figure 3.3a. "GP-UCB-X" uses "Reward X" in Figure 3.3a, which are defined as follows at each round n,

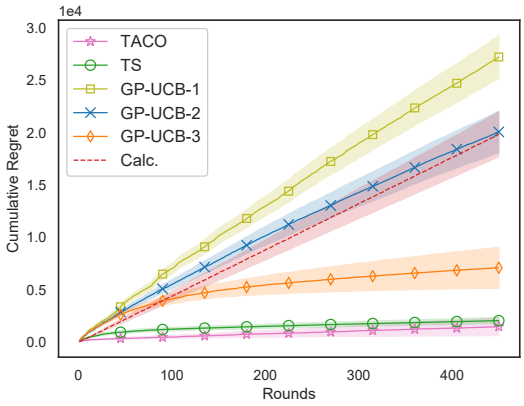
$$
\begin{aligned}
r_1(n) &= -\log(|f(\boldsymbol{z}_n, d_n) - T| + 1) \\
r_2(n) &= 1 - \exp(|f(\boldsymbol{z}_n, d_n) - T|/20) \\
r_3(n) &= -|f(\boldsymbol{z}_n, d_n) - T|
\end{aligned}
$$



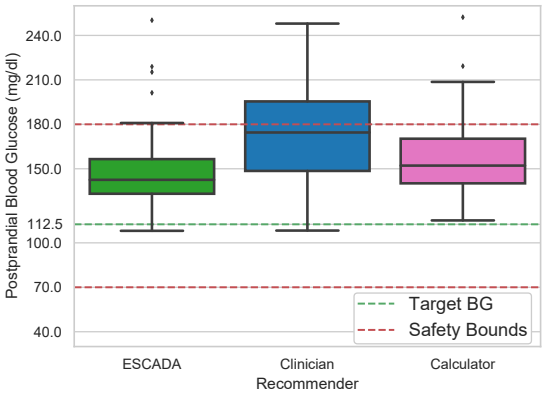
(a) An examplary objective function $f(x)$, and three candidate reward functions.



(b) PPBG distribution boxplots for TACO, TS, GP-UCB, and the calculator in MME scenario.



(c) Cumulative regrets for TACO, TS, GP-UCB, and calculator in MME scenario (± 0.25 std. dev.).



(d) PPBG distribution boxplots for five virtual patients in the clinician comparison experiment.

Figure 3.3

However, $f(x)$ being smooth and efficiently learnable by a GP does not imply that a reward functional defined from $f(x)$ will be too. Figure 3.3a shows that different "plausible" reward functions can have significantly different landscapes. For instance, it is almost impossible to achieve our task efficiently by using the "Reward 1" function in Figure 3.3a. Figures 3.3b and 3.3c show that different reward functions lead to significant differences in the performance of GP-UCB. The practitioner needs to decide on a "good" reward function for each problem in practice. Our algorithms do not require that. Moreover, TACO and TS outperform **all** reward-based GP-UCB models by a wide margin. Also, when

$T \neq (T_{\min} + T_{\max})/2$ (which may well be the case, as in our experiments since $112.5 \neq (70 + 180)/2$), the reward-based UCB methods need another GP to model $f(x)$ to efficiently satisfy safety requirements, doubling the computational complexity compared to our algorithms which use a single GP for everything. Finally, by learning $f(x)$ with a GP, we can provide interpretations for our model's recommendations (see Figure 3.1).

3.5.7.5 Consistency

Figures 3.2a and 3.2c, and Table 3.1 reveal that ESCADA and STS yield similar results. Both algorithms use GPs and have $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory complexities where n is the number of observations. The key difference between them is that STS strikes the balance between exploration and exploitation through intrinsic randomization. That is, for a fixed patient history, STS can make different recommendations for the same meal event in test time, damaging its interpretability and leading to undesired inexplicable variations in the treatment [131]. On the other hand, ESCADA trades-off the exploration and exploitation through the explicit and deterministic machinery described in Section 3.3 which makes it a fairly interpretable model. Moreover, even though we design and test STS, we do not provide an upper bound on its regret as opposed to ESCADA, which is an interesting future work.

3.5.7.6 Comparison with a clinician

We compare ESCADA's performance against a clinician's for five virtual patients. For each patient, we provided the clinician with 20 samples in the form of (meal event, insulin dose, PPBG) and asked her to make recommendations for 20 *unseen* meal events. We provided ESCADA with the same 20 samples for each patient and queried recommendations for the same 20 test meal events. Figure 3.3d shows that the clinician performs slightly worse than the calculator, and ESCADA outperforms both significantly. These results suggest that making inferences about a

patient's dose response is not trivial, and ESCADA is promising as a supplementary tool in clinical setting. Moreover, ESCADA can provide the clinicians with various useful statistics regarding dose responses, such as the confidence region of the response, hypo-/hyper-glycemia probabilities, or probability of response residing in a specific interval for a given patient, meal event, and insulin dose.

Chapter 4

Federated Multi-Armed Bandits Under Byzantine Attacks

4.1 Introduction

In this chapter, we study the FMAB problem discussed in Section 2.2 under Byzantine attacks. We propose a novel median-of-means based algorithm, Fed-MoM-UCB, and demonstrate its effectiveness under Byzantine attacks while also comparing it against the baseline algorithm Fed2-UCB in [52]. Our main contributions are as follows.

- We formalize the FMAB framework with Byzantine clients. (in Section 4.2)
- We propose a median-of-means (MoM) based algorithm, Fed-MoM-UCB, to maintain robustness against the Byzantine clients. (in Section 4.3)
- We show that when an MoM estimator is used, it is not possible to eliminate the arms with suboptimality gaps smaller than an *unavoidable* error margin. (in Section 4.4)
- We derive a high probability $\mathcal{O}(\log T)$ upper bound on the cumulative regret. (in Section 4.4)

- We provide insights towards the parameter selection for the MoM estimator considering the trade-offs between various performance metrics (in Section 4.4)
- We conduct detailed experiments to compare Fed-MoM-UCB's performance against the Fed2-UCB's in [52] under Byzantine attacks and to investigate the effects of different parameters on the performance. (in Section 4.5)

4.2 Problem Formulation

We study the FMAB problem in [52]. There is a *global* feedback model with K arms indexed by the set $[K] \coloneqq \{1, 2, ..., K\}$ and new clients can be enrolled by the parameter server over *phases*. When pulled, the k-th arm returns a *global* reward X_k sampled from a σ-subgaussian distribution with *unknown* mean μ_k and *known* σ. We denote by $k^* \coloneqq \text{argmax}_{k\in[K]}\, \mu_k$ the *globally optimal* arm and assume that it is unique.

Clients do not interact directly with the global model but rather with their *local* models, which are stochastic realizations of the global model. Precisely, the mean outcome of arm k for client m, denoted by $\mu_{m,k}$, is sampled from a σ_c-subgaussian distribution with unknown mean μ_k and known σ_c. Then, client m plays a K-armed bandit problem, where playing an arm k at round t yields the following local stochastic reward $X_{m,k}(t)$,

$$X_{m,k}(t) = \underbrace{\mu_k + \xi_{m,k}}_{\mu_{m,k}} + \zeta_t \ , \tag{4.1}$$

where $\xi_{m,k}$ and ζ_t are sampled from zero-mean σ_c-subgaussian and σ-subgaussian distributions, respectively. While ζ_t are sampled each time an arm is pulled, $\xi_{m,k}$ are sampled once at the beginning, and $\mu_{m,k}$ are fixed after that. Note that for different clients m and n, we do not have $\mu_{m,k} = \mu_{n,k}$ in general. Since we also do not have $\mu_{m,k} = \mu_k$, a group of M clients may not agree with the true globally optimal arm, i.e., $\text{argmax}_{k\in[K]}\, \frac{1}{M}\sum_{m=1}^{M} \mu_{m,k} \neq k^*$. Then, trying to identify the globally optimal arm based on local observations is a futile effort as the global

and local models are inconsistent. [52] demonstrates as the intuition suggests that this problem can be handled by recruiting a sufficient number of clients over phases and exploiting the subgaussian tail behavior of the "client noise." However, when there is even a single Byzantine client, identifying the optimal arm using the Fed2-UCB algorithm in [52] which solely relies on sample means, maybe impossible, regardless of how many clients the server recruits. Consider a Byzantine client m who sends the following updates as sample means for arm outcomes to the server at the end of a phase p where Fed2-UCB is used as the learning protocol:

$$U_{m,k}(p) = \begin{cases} \widehat{\mu}_{m,k}(p) & \text{if } k \neq k^* \\ -\infty & \text{if } k = k^* \end{cases} \tag{4.2}$$

where $\widehat{\mu}_{m,k}(p)$ is the *true* sample mean of observations for arm k collected by client m by the end of phase p. Since Fed2-UCB uses the mean over the clients' individual updates for an arm by the end of each phase, (4.2) will cause the global estimate for the optimal arm, $\overline{\mu}_{k^*}(p)$, to go to $-\infty$, hence making it impossible for Fed2-UCB to identify the globally optimal arm. Similarly, a Byzantine client can favor *any* suboptimal arm to make it appear like the optimal arm.

We show that using an MoM estimator; it is possible to guarantee robustness against the Byzantine clients (with high probability) at the expense of partial identifiability. We cannot guarantee that an arm with a suboptimality gap smaller than $\omega + \epsilon > 0$ will be eliminated. Similar to [134], we call such arms $(\omega + \epsilon)$-optimal arms and all the other arms $(\omega + \epsilon)$-suboptimal arms, and Fed-MoM-UCB a $(\omega + \epsilon,\ \delta)$-PAC (probably approximately correct) algorithm, with a catch. Even though we can control the confidence level $\delta \in (0, 1)$ and $\epsilon > 0$, we cannot control $\omega > 0$. We characterize that unavoidable error margin ω in Section 4.4 (see (4.21)). Moreover, in Sections 4.3 and 4.4, we show that enrolling more clients does not reduce the uncertainty emerging from client sampling, which in turn leads to the unavoidable ω-suboptimality gap. Therefore, we recruit all the clients in the beginning in a way that strikes the optimal balance between the cumulative regret and the high probability guarantees (see Section 4.4.1). An immediate question following such an observation is: "Why do we have phases if all the clients are recruited at once?". As a matter of fact, Fed-MoM-UCB can

eliminate all $(\omega+\epsilon)$-suboptimal arms after a single phase. In that case, however, we incur additional regret because all arms are played until the "best $(\omega+\epsilon)$-suboptimal arm" is eliminated. As a result, we divide the time horizon into phases and eliminate the $(\omega+\epsilon)$-suboptimal arms as soon as we have enough confidence to incur less regret. It is worth noting that we incur additional communication cost in the latter case, which is included in the regret definition. To keep the notation uncluttered later, we define and use $\beta \coloneqq \omega+\epsilon$. We make the following assumption on the ratio of the Byzantine clients.

Assumption 4.1. At most $0 \leq \lambda < 0.5$ fraction of the clients recruited by the server are Byzantine clients where the parameter server knows λ beforehand. Neither the server nor any honest client knows whether a client is Byzantine.

We denote by $\mathcal{M}$ the set of clients recruited by the parameter server at the beginning with cardinality $|\mathcal{M}| = M$, and by $\mathcal{M}^{\text{mal}} \subset \mathcal{M}$ the set of Byzantine clients with cardinality $|\mathcal{M}^{\text{mal}}| = M^{\text{mal}}$. Let $\mathcal{K}_\beta$ denote the set of β-optimal arms, that is, $\mu_{k^*} - \mu_k \leq \beta$, for every $k \in \mathcal{K}_\beta$, where k^* is the globally optimal arm. Also, let $\neg\mathcal{K}_\beta \coloneqq [K] \setminus \mathcal{K}_\beta$ denote the set of β-suboptimal arms. Using an MoM estimator, our goal is to guarantee finding an arm $k \in \mathcal{K}_\beta$. We denote by $\mathcal{M}_t^\beta \subseteq \mathcal{M}$ the set of clients who plays an arm $k \in \neg\mathcal{K}_\beta$ at round t. Let I_t^m denote the arm selected by client m at round t, and Δ_k the suboptimality gap for arm k, i.e., $\Delta_k = \mu_{k^*} - \mu_k$. We define the following *pseudo-regret* as the proxy performance metric which also considers the communication cost between the server and clients,

$$R_T^\beta \coloneqq \sum_{t=1}^{T} \sum_{m \in \mathcal{M}_t^\beta} (\Delta_{I_t^m} - \beta) + CMP_T \ , \tag{4.3}$$

where the constant $C > 0$ is the communication cost of a single interaction between a client and the parameter server. P_T denotes the number of communication rounds by the end of round T.

4.3 Fed-MoM-UCB Algorithm

We propose Fed-MoM-UCB: Federated Median-of-Means Upper Confidence Bound algorithm to cope with the Byzantine attacks. While the clients follow the same arm sampling procedure as in Fed2-UCB in [52], the parameter server runs an entirely new MoM-based robust estimator for global arm outcomes, which is the critical component of Fed-MoM-UCB. At the end of each phase, the parameter server communicates with the clients to gather local observations, perform arm elimination, and send back the updated set of active arms. Each phase lasts for a set number of rounds, which the server communicates to the clients before the phase begins.

4.3.1 Client Side

The server sends the clients the set of active arms $\mathcal{A}_p \subseteq [K]$ at the start of a phase p. Each client $m \in \mathcal{M}$ samples each arm in $\mathcal{A}_p$ to accumulate a total of $s(p)$ samples from each arm by the end phase p, where the server communicates $s(p) \in \mathbb{Z}_+$, which increases in p, to the clients. When clients complete their sampling, we mark the end of a phase. At the end of a phase p, an *honest* client $m \in \mathcal{M} \setminus \mathcal{M}^{\text{mal}}$ forms its local update vector as,

$$\boldsymbol{U}_m(p) = \{\widehat{\mu}_{m,k}(p)\}_{k \in \mathcal{A}_p} \ .$$

where $\widehat{\mu}_{m,k}(p)$ is the sample mean of the observations for arm k collected by client m by the end of phase p. However, a Byzantine client $m \in \mathcal{M}_p^{\text{mal}}$ can send an arbitrary $\boldsymbol{U}_m(p)$ to the parameter server (see (4.2)).

4.3.2 Server Side

The parameter server recruits M clients only once at the beginning where the choice of M is discussed in Section 4.4.1 (see (4.13)). After a phase p is done, the parameter server executes an arm elimination protocol. First, the clients are

Algorithm 2 Fed-MoM-UCB Parameter Server Side

Inputs: T; K; δ; ϵ; λ; $\alpha(\lambda)$; σ; σ_c $s(p)$

Calculate the number of clients to recruit (M) as in (4.13)
Calculate the number of groups $G = \lceil \alpha(\lambda) M \rceil$
Calculate the number of clients in a group $B = \lfloor M/G \rfloor$
Calculate the unavoidable error margin ω as in (4.21)
Form $\beta = \omega + \epsilon$
Recruit M clients to constitute the client set $\mathcal{M}$
Set $t = 0$, $p = 1$, $\mathcal{A}_1 = [K]$, $E_0 = +\infty$
while $E_{p-1} > \beta/4$ and $|\mathcal{A}_p| > 1$ **do**
    Ask each client $m \in \mathcal{M}$ to collect a total of $s(p)$ observations for each active arm $k \in \mathcal{A}_p$
    Query the local updates $\boldsymbol{U}_m(p)$ (i.e., mean arm outcomes) for each client $m \in \mathcal{M}$
    Divide the clients into G groups
    Calculate group means $U_k^i(p)$ for $i \in [G]$ as in (4.4)
    Form the median-of-means (MoM) estimates $\overline{U}_k(p)$ for global means as: $\overline{U}_k(p) = \text{med}\big(\{U_k^i(p)\}_{i \in [G]}\big)$
    Form the active arm set $\mathcal{A}_{p+1}$ for next phase as in (4.6)
    $t \leftarrow t + \big(s(p) - s(p-1)\big)|\mathcal{A}_p|$
    $p \leftarrow p + 1$
end while
Have the clients pull the arms in $\mathcal{A}_p$ uniformly until $t = T$.

divided into $G = \lceil \alpha(\lambda) M \rceil$ groups randomly for some $\alpha(\lambda) \in (0, 1]$, where each group consists of at least $B = \lfloor M/G \rfloor$ clients.[1] We provide guidelines and insights towards the selection of the mapping $\alpha : \ [0, 1/2] \to [0, 1]$ to guarantee that at least half of the groups are populated by honest clients only in Section 4.4.3. Let $\mathcal{B}_{i,p}$ represent the set of clients in group i in phase p, where $i \in [G]$. The *intermediary* mean outcome of arm $k \in \mathcal{A}_p$ for group i is calculated as,

$$U_k^i(p) = \frac{1}{B} \sum_{m \in \mathcal{B}_{i,p}} \widehat{\mu}_{m,k}(p) \ . \tag{4.4}$$

[1]Without loss of generality, B is assumed to be the same among all groups in any phase p. When M is not exactly divisible with G, a group will have more than B clients. That does not change our analysis since estimations will improve when more clients are in a group.

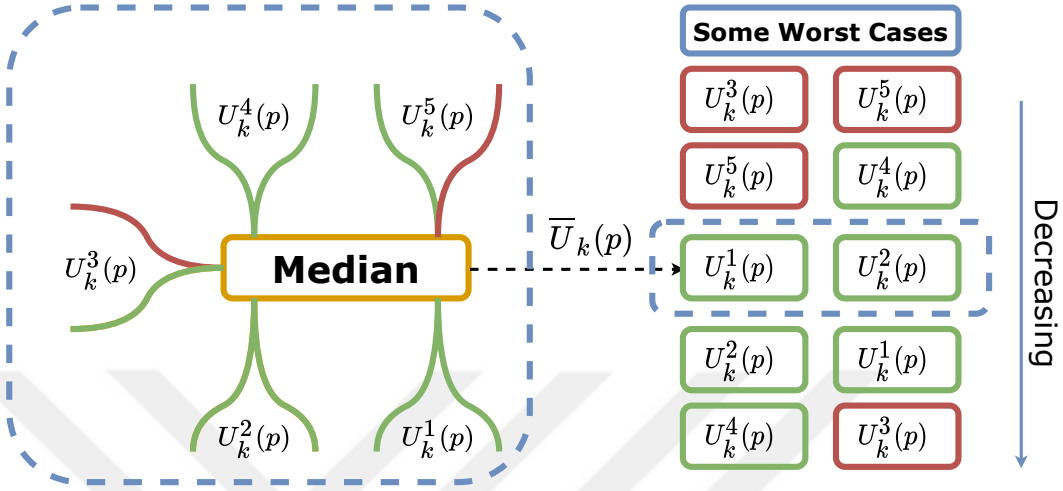


Figure 4.1: An exemplary scenario with ten clients where two are Byzantine (red lines), and they perform outlier attacks. Fed-MoM-UCB first divides the clients into five groups. In the worst case, the Byzantine clients are assigned to different groups and saturate their groups' local updates. The saturation will not affect the median estimate in any scenario.

Then, the server merges the group means for arm outcomes by using a median-of-means (MoM) estimator to form estimates for the global arm rewards as,

$$\begin{aligned} \overline{\boldsymbol{U}}(p) &= \{\overline{U}_k(p)\}_{k\in\mathcal{A}_p} \\ \text{where} \quad \overline{U}_k(p) &= \text{med}\big(\{U_k^i(p)\}_{i\in[G]}\big) \ . \end{aligned} \tag{4.5}$$

Since the local updates are mean arm outcomes, the resulting estimator is a median-of-means-of-means (MoMoM) estimator. Finally, the server uses $\overline{\boldsymbol{U}}(p)$ to eliminate the suboptimal arms in $\mathcal{A}_p$ to update the active arm set for the next phase as,

$$\mathcal{A}_{p+1} = \mathcal{A}_p \setminus \{k \in \mathcal{A}_p | \overline{U}_k(p) + E_p \leq \max_{k\in\mathcal{A}_p} \overline{U}_k(p) - E_p\} \ , \tag{4.6}$$

where $E_p \in \mathbb{R}_+$ is chosen to satisfy,

$$\Pr\left\{|\overline{U}_k(p) - \mu_k| \leq E_p, \ \forall p \in [P_T], \forall k \in \mathcal{A}_p\right\} \geq 1-\delta \ , \tag{4.7}$$

for a given $\delta \in (0,1)$. We provide detailed explanations on how to choose E_p to guarantee the high probability condition in (4.7) in the next section, which is intertwined with the selection of the cohort size M. Figure 4.1 provides an exemplary scenario where the Byzantine clients perform outlier attacks, which

are discarded thanks to the grouping strategy with a sufficient number of distinct groups.

4.4 Theoretical Analyses

4.4.1 MoM Estimator Concentration

Our analyses rely on the fact that with a proper choice of $\alpha(\lambda)$, we can guarantee that at least half of the $G = \lceil \alpha(\lambda)M \rceil$ groups in any phase p will consist of honest clients only (i.e., see Figure 4.1). Simultaneously, we want each group to have as many clients as possible to reduce the effect of local model heterogeneity (i.e., client sampling noise) on the estimates. To control the group size, we propose the following generic mapping as in [135].

Definition 4.1. Let $\alpha : [0, 1/2] \to [0, 1]$ be a mapping s.t.

$$\forall \lambda \in [0, 1/2),\ 2\lambda < \alpha(\lambda) < 1 \ .$$

Remark 4.1. Notice that for a fixed mapping $\alpha(\lambda)$, the group size $B = \lfloor M/G \rfloor \leq M/(\alpha(\lambda)M) \leq 1/\alpha(\lambda)$ can not increase with p as it is strictly upper-bounded by $1/\alpha(\lambda)$. Because the number of clients in a group cannot increase with p, the uncertainty arising from the *client sampling* within a group does not decrease over phases, which carries over to the global estimates because the server selects a single group using a median-of-means estimator. This observation highlights how partial identifiability manifests in the MoMoM estimator. Furthermore, this is precisely why we recruit all the clients only once at the beginning.

The following lemma presents a concentration bound for the MoM estimate of a global arm outcome μ_k in a phase p.

Lemma 4.1. *(MoM concentration bound). Let* $G = \lceil \alpha(\lambda)M \rceil$, $\eta(\lambda) \coloneqq (\alpha(\lambda) - \lambda)/\alpha(\lambda)$, *and* $\rho := \sigma/\sqrt{s(p)} + \sigma_c$. *We have,*

$$\Pr\left\{|\overline{U}_k(p) - \mu_k| > \gamma_p\right\} \leq \exp\left(-\frac{2\eta(\lambda) - 1}{16\rho^2} M\gamma_p^2\right) \ , \tag{4.8}$$

for all,

$$\gamma_p^2 \geq 4\left(\frac{\sigma}{\sqrt{s(p)}} + \sigma_c\right)^2 \left(\frac{4\eta(\lambda)-1}{2\eta(\lambda)-1}\frac{\log 2}{B}\right) \ . \tag{4.9}$$

Proof. The proof builds upon the proof of MoM concentration bound given in Proposition 2 of [135]. Remember that we have defined in Section 4.2 $\mathcal{M}$ as the client cohort with cardinality M, and $\mathcal{M}^{\text{mal}} \subset \mathcal{M}$ as the set of Byzantine clients with cardinality M^{mal}. Let $\mathcal{G}_p$ denote the set of client groups formed by the server at the end of phase p and $\mathcal{G}_p^{\text{tru}} \subseteq \mathcal{G}_p$ denote the subset of groups with no Byzantine clients. We denote by $G_p = |\mathcal{G}_p|$ and $G_p^{\text{tru}} := |\mathcal{G}_p^{\text{tru}}|$ the cardinalities of these sets, respectively (i.e., the number of client groups within the sets). For the groups including even a single Byzantine client, we can not provide any concentration guarantee. The worst case is when each Byzantine client belongs to a different group, rendering the maximum number of groups unreliable. In such a case, at most $\lfloor \lambda M \rfloor$ of the G_p groups will be corrupted by Byzantine clients, that is $M^{\text{mal}} \leq \lambda M$. Then, we have,

$$G_p^{\text{tru}} \geq G_p - \lambda M \geq G_p - \frac{\lambda G_p}{\alpha(\lambda)} \ .$$

Fix $\gamma_p > 0$. By definition of the sample median, if at least half of the groups have a sample mean estimate within γ_p-neighborhood of μ_k, their sample median must be within γ_p-neighborhood of μ_k as well. We can then write,

$$\begin{aligned}
&\Pr(|\overline{U}_k(p) - \mu_k| > \gamma_p | \mathcal{G}_p) \\
&\leq \Pr\left(\sum_{i \in \mathcal{G}_p^{\text{tru}}} \mathbb{I}\{|U_k^i(p) - \mu_k| > \gamma_p\} \geq G_p^{\text{tru}} - G_p/2 \middle| \mathcal{G}_p\right) \\
&\leq \Pr\left(\sum_{i \in \mathcal{G}_p^{\text{tru}}} \mathbb{I}\{|U_k^i(p) - \mu_k| > \gamma_p\} \geq \frac{2\eta(\lambda)-1}{2\eta(\lambda)} G_p^{\text{tru}} \middle| \mathcal{G}_p\right) \\
&\leq \sum_{g=\lceil \frac{2\eta(\lambda)-1}{2\eta(\lambda)} G_p^{\text{tru}} \rceil}^{G_p^{\text{tru}}} \binom{G_p^{\text{tru}}}{g} q_p^g (1-q_p)^{G_p^{\text{tru}}-g} \\
&\leq q_p^{\frac{2\eta(\lambda)-1}{2} G_p} 2^{\eta(\lambda) G_p} \ ,
\end{aligned} \tag{4.10}$$

where $q_p \coloneqq \Pr(|U_k^i(p) - \mu_k| > \gamma_p | \mathcal{G}_p)$. (4.10) is derived as in the proof of Proposition 2 in [135] after observing that $\sum_{i \in \mathcal{G}_p^{\text{tru}}} \mathbb{I}\{|U_k^i(p) - \mu_k| > \gamma_p\}$ is a binomial random variable. Next, we observe the following for $i \in \mathcal{G}_p^{\text{tru}}$,

$$
\begin{aligned}
U_k^i(p) - \mu_k &= \frac{1}{B} \sum_{m \in \mathcal{B}_{i,p}} (\widehat{\mu}_{m,k}(p) - \mu_{m,k} + \mu_{m,k} - \mu_k) \\
&= \underbrace{\frac{1}{B} \sum_{m \in \mathcal{B}_{i,p}} (\widehat{\mu}_{m,k}(p) - \mu_{m,k})}_{\text{Term I}} + \underbrace{\frac{1}{B} \sum_{m \in \mathcal{B}_{i,p}} (\mu_{m,k} - \mu_k)}_{\text{Term II}} \ .
\end{aligned}
$$

By definition, Term I is the average of B independent $\sigma/\sqrt{s(p)}$-subgaussian random variables. Thus, Term 1 is $\sigma/\sqrt{s(p)B}$-subgaussian. Similarly, Term 2 is the average of B independent σ_c-subgaussian random variables. Therefore, it is $\sigma_c/\sqrt{B}$-subgaussian. As a result, Term I plus Term II is $\sigma/\sqrt{s(p)B} + \sigma_c/\sqrt{B}$-subgaussian. Let $\rho = \sigma/\sqrt{s(p)} + \sigma_c$ and $\rho' \coloneqq \frac{1}{\sqrt{B}}\rho$. Then, using the subgaussian tail inequality (Theorem 5.3 in [136]), we have $q_p \leq 2e^{-\gamma_p^2/(2(\rho')^2)} = 2e^{-B\frac{\gamma_p^2}{2\rho^2}}$. Plugging this into (4.10), we obtain,

$$
\Pr(|\overline{U}_k(p) - \mu_k| > \gamma_p) \leq \exp\left(-\frac{2\eta(\lambda) - 1}{16\rho^2} M \gamma_p^2\right) \ ,
$$

for all,

$$
\gamma_p^2 \geq \frac{4\rho^2}{B} \frac{4\eta(\lambda) - 1}{2\eta(\lambda) - 1} \log 2 \ ,
$$

which completes the proof. □

The MoM concentration bound in (4.8) conditional to (4.9) immediately provides the practitioner with insights towards the cohort size M and the high probability confidence bound E_p in (4.7). Contrary to the scenario where the final estimator for μ_k, $\overline{U}_k(p)$, has subgaussian tail behavior, we can not provide any guarantees for an arbitrarily small confidence bound (γ_p in (4.8)). The confidence bounds we can provide any guarantees for is strictly lower bounded by

(4.9). Then our best chance is to set (by (4.9)),

$$E_p = 2\left(\underbrace{\frac{\sigma}{\sqrt{s(p)}}}_{\substack{\text{arm}\\ \text{sampling}}} + \underbrace{\sigma_c}_{\substack{\text{client}\\ \text{sampling}}}\right)\left(\frac{4\eta(\lambda)-1}{2\eta(\lambda)-1}\frac{\log 2}{B}\right)^{1/2} . \tag{4.11}$$

Regardless of how many phases the algorithm operates, the confidence bound in (4.11) does not shrink beyond an unavoidable margin partially determined by the local model heterogeneity level σ_c (client sampling), which is formalized in Lemma 4.4. That happens because even though the clients collect new observations over phases and decrease the uncertainty caused by the "arm sampling" in (4.11), the number of clients within a group remains the same. Next, by (4.8), we need to have,

$$\exp\left(-\frac{2\eta(\lambda)-1}{16\rho^2}ME_p^2\right) \leq \frac{\delta}{KT} , \tag{4.12}$$

to satisfy (4.7) for an arbitrary $\delta \in (0,1)$. Note that since we have $P_T \leq T$ and $|\mathcal{A}_p| \leq K$, ensuring that (4.12) holds implies that (4.7) holds after a union bound over all $p \in [P_T]$ and $k \in \mathcal{A}_p$. In the case where $\overline{U}_k(p)$ has subgaussian tail behavior, one could leverage employing more clients over phases (as in [52]) to maintain the high probability guarantees for a given $\delta \in (0,1)$ with a tighter confidence bound E_p. However, as we established in (4.11) by Lemma 4.1, our high probability confidence bound E_p can shrink only so much in a given phase p. That is, employing more clients over phases does not provide us any advantage. Now, observe that in order to satisfy (4.12) after inserting E_p as in (4.11), we must ensure,

$$M \geq \frac{4B\log(\frac{KT}{\delta})}{(4\eta(\lambda)-1)\log 2} .$$

Since we have $B \leq 1/\alpha(\lambda)$ by Remark 4.1 and M must be an integer, it suffices to have,

$$M = \frac{4\log(\frac{KT}{\delta})}{(4\eta(\lambda)-1)\alpha(\lambda)\log 2} + \Gamma_1 , \tag{4.13}$$

for some constant $\Gamma_1 \in [0,1)$. Finally, by (4.11) and (4.13), we can ensure that (4.7) holds for the tightest possible confidence bound E_p using an MoM estimator,

with probability at least $1-\delta$, for a given $\delta \in (0,1)$. For the rest of the manuscript, we always assume that the cohort size M and the confidence bounds E_p are set as in (4.11) and (4.13). Next, we define the "good event" where the confidence bounds hold over all phases and all active arms, and summarize our findings up to now in a single lemma.

Definition 4.2. (Good event)

$$\mathcal{E} \coloneqq \left\{ |\overline{U}_k(p) - \mu_k| \leq E_p, \forall p \in [P_T], \forall k \in \mathcal{A}_p \right\} . \tag{4.14}$$

Lemma 4.2. *Pick $\delta \in (0,1)$ and a mapping $\alpha(\lambda)$ as in Definition 4.1. Set the cohort size M as in* (4.13) *and the confidence bounds E_p as in* (4.11)*. We have,*

$$\Pr\{\mathcal{E}\} \geq 1 - \delta . \tag{4.15}$$

Proof. The proof simply reiterates over the results of our choices of the cohort size M and the confidence bounds E_p. Observe that we have (4.12) by (4.11) and (4.13). Combining (4.12) with (4.8), we have,

$$\Pr\left\{ |\overline{U}_k(p) - \mu_k| > E_p \right\} \leq \delta / KT . \tag{4.16}$$

for any phase p and arm k. Then,

$$\begin{aligned} &\Pr\left\{ |\overline{U}_k(p) - \mu_k| \leq E_p,\ \forall p \in [P_T], \forall k \in \mathcal{A}_p \right\} \\ &\quad \geq 1 - \sum_{p \in [T]} \sum_{k \in [K]} \Pr\left\{ |\overline{U}_k(p) - \mu_k| > E_p \right\} \qquad (4.17) \\ &\quad \geq 1 - \delta , \qquad (4.18) \end{aligned}$$

where (4.17) follows from the union bound and the fact that $P_T \leq T$ and $|\mathcal{A}_p| \leq K$. (4.18) then follows from (4.16). □

4.4.2 Main Results and Regret Bound

By the end of each phase p, we execute the arm elimination rule in (4.6). We first show that the globally optimal arm $k^* \in [K]$ is guaranteed to always remain active.

Lemma 4.3. *Assume that the good event $\mathcal{E}$ in (4.14) holds. Then, the globally optimal arm k^* is always guaranteed to remain active.*

Proof. For any phase p, let $k' = \text{argmax}_{k \in \mathcal{A}_p}\{\overline{U}_k(p)\}$, that is, k' is *empirically* the best arm by the end of phase p. For the globally optimal k^* to be eliminated by k', we need to have $\overline{U}_{k^*}(p) + E_p \leq \overline{U}_{k'}(p) - E_p$ by (4.6). That implies,

$$\begin{aligned} \mu_{k^*} &\leq \overline{U}_{k^*}(p) + E_p && (4.19) \\ &\leq \overline{U}_{k'}(p) - E_p \\ &\leq \mu_{k'} \ , && (4.20) \end{aligned}$$

which is a contradiction since we have $\mu_{k^*} > \mu_{k'}$. Note that we have (4.19) and (4.20) since the good event $\mathcal{E}$ holds. □

The globally optimal arm k^* will never be eliminated under the good event $\mathcal{E}$, according to Lemma 4.3. This fact demonstrates the elimination of β-suboptimal arms in the following lemmas. Fed-MoM-UCB uses uniform exploration in conjunction with arm elimination. The following lemma first formalizes the unavoidable error margin ω, which has previously been discussed in Sections 4.2, 4.4.1, and Remark 4.1. It then specifies the maximum number of phases required to eliminate a β-suboptimal arm.

Lemma 4.4. *(β-suboptimal arm elimination) Following the "client sampling" term that is non-decreasing over phases in (4.11), we define the unavoidable error margin ω as,*

$$\omega \coloneqq \frac{8\sigma_c\sqrt{\log 2}\sqrt{4\eta(\lambda) - 1}}{\sqrt{B}\sqrt{2\eta(\lambda) - 1}} \ . \tag{4.21}$$

Then, under the good event $\mathcal{E}$, any arm $k \in [K]$ with suboptimality gap $\Delta_k > \omega$ will be eliminated by the end of phase p_k, where p_k is the smallest integer satisfying

$$s(p_k) \geq \frac{64\sigma^2 \log 2(4\eta(\lambda) - 1)}{(\Delta_k - \omega)^2 B(2\eta(\lambda) - 1)} \ . \tag{4.22}$$

Since $\beta = \omega + \epsilon$, (4.22) holds for any β-suboptimal arm as well.

Proof. Consider an ω-suboptimal arm k, that is, $\Delta_k > \omega$. Furthermore, let p_k be the smallest integer satisfying (4.22), which can be checked to imply $4E_{p_k} \leq \Delta_k$. Then, we have,

$$\begin{aligned}
\overline{U}_k(p_k) + E_{p_k} &\leq \mu_k + 2E_{p_k} && (4.23)\\
&\leq \mu_k + 2E_{p_k} + \overline{U}_{k^*}(p_k) - \mu_{k^*} + E_{p_k} && (4.24)\\
&= \overline{U}_{k^*}(p_k) + 3E_{p_k} - \Delta_k && (4.25)\\
&\leq \overline{U}_{k^*}(p_k) - E_{p_k} \ , && (4.26)
\end{aligned}$$

where (4.23) and (4.24) hold under the good event $\mathcal{E}$. (4.25) follows from the definition of the suboptimality gap $\Delta_k = \mu_{k^*} - \mu_k$, and (4.26) holds since we have $4E_{p_k} \leq \Delta_k$. Finally, by the arm elimination rule in (4.6), (4.26) then implies that an ω-suboptimal arm k will be eliminated at the end of phase p_k. □

Note that p_k is the smallest *integer* satisfying (4.22). Then, we have,

$$s(p_k) = \frac{64\sigma^2 \log 2(4\eta(\lambda) - 1)}{(\Delta_k - \omega)^2 B(2\eta(\lambda) - 1)} + \Gamma_2^k \ , \tag{4.27}$$

for some $\Gamma_2^k \geq 0$.

Remark 4.2. Because p_k must be an integer, we have the additional constant $\Gamma_2^k \geq 0$, which is determined by the suboptimality gap Δ_k of arm $k \in [K]$ and the mapping $s(p)$. For example, if we choose $s(p) = a^p$ for some $a \in \{2, 3, \ldots\}$, the residual constant Γ_2^k may be larger than if we choose $s(p) = ap$. The number of phases (p_k) required to eliminate a β-suboptimal arm is reduced in the former case because clients collect more samples between phases, increasing the number of β-suboptimal arm pulls before they are eliminated (i.e., larger Γ_2^k) and the cumulative regret. In the latter case, a β-suboptimal arm (i.e., a smaller Γ_2^k) will be eliminated sooner. However, this comes at the expense of more frequent communication with the parameter server, which raises the total regret due to the communication cost. One extreme case is $s(p) = p$, in which a β-suboptimal arm k is eliminated "as soon as possible," that is, $\Gamma_2^k = 0$. However, the clients communicate with the parameter server every time after pulling all of the arms *once*. Another edge case is eliminating all the β-suboptimal arms in a single communication round, which we investigate in Corollary 4.1.

By providing closed-form expressions for the phases where β-suboptimal arms will be eliminated, Lemma 4.4 forms the basis of the cumulative regret incurred by Fed-MoM-UCB. The following lemma ensures that all β-suboptimal arms are eliminated.

Lemma 4.5. *(Termination) By the time Fed-MoM-UCB terminates, all the β-suboptimal arms are eliminated under the good event $\mathcal{E}$.*

Proof. Let us denote by $p_{\max}$ the phase after which Fed-MoM-UCB terminates, that is, $E_{p_{\max}} \leq \beta/4$ or $|\mathcal{A}_{p_{\max}}| = 1$ by Algorithm 2. If the latter is the case, we have $\mathcal{A}_{p_{\max}} = \{k^*\}$ by Lemma 4.3 and we are done. For the former case where $E_{p_{\max}} \leq \beta/4$, remember that we have $\Delta_k > \beta$ for a β-suboptimal arm $k \in \neg\mathcal{K}_\beta$. Then, similar to the proof of Lemma 4.4, it immediately follows that,

$$\begin{aligned} \overline{U}_k(p_{\max}) + E_{p_{\max}} &\leq \overline{U}_{k^*}(p_{\max}) + 3E_{p_{\max}} - \Delta_k && (4.28) \\ &\leq \overline{U}_{k^*}(p_{\max}) - E_{p_{\max}} \ , && (4.29) \end{aligned}$$

where (4.28) follows from (4.25), and (4.29) follows from $\Delta_k > \beta \geq 4E_{p_{\max}}$, suggesting that the β-suboptimal is bound to be eliminated by the end of phase $p_{\max}$. □

Before moving on to bounding the cumulative regret of Fed-MoM-UCB, we will derive an expression for $p_{\max}$ in Lemma 4.5, which will directly determine the upper bound on the total communication cost.

Lemma 4.6. *(Number of communication phases) Pick some $\epsilon > 0$ to form $\beta = \omega + \epsilon$. Then, we have,*

$$s(p_{\max}) \leq \left\lceil \frac{64\sigma^2}{\epsilon^2} \left(\frac{4\eta(\lambda) - 1}{2\eta(\lambda) - 1} \frac{\log 2}{B} \right) \right\rceil \ , \tag{4.30}$$

where $\lceil x \rceil$ is the smallest integer greater than x.

Proof. By our definitions of the unavoidable error margin ω (see (4.21)) and the confidence bounds E_p (see (4.11)), we have,

$$E_p = \frac{\omega}{4} + \frac{2\sigma}{\sqrt{s(p)}} \left(\frac{4\eta(\lambda) - 1}{2\eta(\lambda) - 1} \frac{\log 2}{B} \right)^{1/2} \ . \tag{4.31}$$

Remember that $\beta = \omega + \epsilon$, where $\epsilon > 0$ can be chosen arbitrarily small. Fed-MoM-UCB stops the communication and arm elimination protocol when $E_p \leq \beta/4 = (\omega + \epsilon)/4$. Let p' be the smallest integer satisfying,

$$\frac{2\sigma}{\sqrt{s(p')}} \left(\frac{4\eta(\lambda) - 1}{2\eta(\lambda) - 1} \frac{\log 2}{B} \right)^{1/2} \leq \frac{\epsilon}{4} \ . \tag{4.32}$$

Then, Fed-MoM-UCB is guaranteed to terminate at the end of phase p' by (4.31). Combining (4.32) with the fact that $p_{\max} \leq p'$ and $s(p)$ is increasing in p, (4.30) follows. □

As one expects, the total number of phases increases for smaller ϵ. Similar discussions to Remark 4.2 regarding effect of $s(p)$ on $p_{\max}$ follow as well. We also note that $p_{\max} \geq P_T$ in (4.3). We are now ready to bound the cumulative regret of Fed-MoM-UCB in (4.3). First, let us define the following constants to keep the notation uncluttered later.

Definition 4.3. (Auxiliary constants)

$$\Gamma_3 \coloneqq \frac{64\sigma^2(4\eta(\lambda) - 1)\log 2}{(2\eta(\lambda) - 1)B} \ . \tag{4.33}$$

$$\Gamma_4 \coloneqq \frac{4}{(4\eta(\lambda) - 1)\alpha(\lambda)\log 2} \ . \tag{4.34}$$

$$\Gamma_5 \coloneqq \Gamma_3\Gamma_4 \ . \tag{4.35}$$

Theorem 4.1. *Pick $\delta \in (0, 1)$ and $\epsilon > 0$. Let $\lambda < 0.5$ be the fraction of Byzantine clients and choose a mapping $\alpha(\lambda)$ satisfying Definition 4.1. Set $\beta = \omega + \epsilon$, the cohort size M as in* (4.13)*, and the confidence bounds E_p as in* (4.11)*. The cumulative regret of Fed-MoM-UCB can be upper-bounded as,*

$$\begin{aligned} R_T^\beta < \sum_{k \in \neg\mathcal{K}_\beta} \Bigg(\Bigg(& \frac{\Gamma_5}{\Delta_k - \omega} + \Gamma_4\Gamma_2^k(\Delta_k - \beta) \Bigg) \log(\frac{KT}{\delta}) \\ & + \frac{\Gamma_3}{\Delta_k - \omega} + \Gamma_2^k(\Delta_k - \beta) \Bigg) \\ & + C p_{\max} \left(\Gamma_4 \log(\frac{KT}{\delta}) + 1 \right) \ , \end{aligned} \tag{4.36}$$

with probability at least $1 - \delta$, where $p_{\max}$ satisfies (4.30)*.*

Proof. Given that a β-suboptimal arm $k \in \neg\mathcal{K}_\beta$ is bound to be eliminated by the end of the phase p_k, we have,

$$R_T^\beta \leq \sum_{k \in \neg\mathcal{K}_\beta} M(\Delta_k - \beta)s(p_k) + CMp_{\max} \ , \tag{4.37}$$

By (4.27) and (4.33) we have,

$$(\Delta_k - \beta)s(p_k) = (\Delta_k - \beta)\left(\frac{\Gamma_3}{(\Delta_k - \omega)^2} + \Gamma_k^2\right) \tag{4.38}$$

$$< \frac{\Gamma_3}{\Delta_k - \omega} + \Gamma_2^k(\Delta_k - \beta) \ , \tag{4.39}$$

where (4.39) holds since $\beta > \omega$. Let us restate the expression for M in (4.13) using (4.34),

$$M = \Gamma_4 \log(\frac{KT}{\delta}) + \Gamma_1 \ , \tag{4.40}$$

where $\Gamma_1 \in [0, 1)$ is a constant. Theorem 4.1 then follows after combining (4.35), (4.37), (4.39), (4.40), and the fact that $\Gamma_1 < 1$. □

In the regret bound (4.36), the Γ_2^k term multiplies the $\Delta_k - \beta$ term (i.e., instantaneous regret). This is consistent with our findings in Remark 4.2, namely that a larger residual Γ_2^k increases regret. The $\Gamma_5/(\Delta_k - \omega)$ term implies that a smaller suboptimality gap Δ_k will increase regret because the suboptimal arm k will be active for a longer period of time. Finally, communication regret increases directly with decreasing ϵ while also being affected by the choice of $s(p)$ through $p_{\max}$. It is worth noting that for specific $s(p)$ values, one can explicitly find an upper bound on $p_{\max}$ using (4.30).

Corollary 4.1. *It is possible to eliminate all β-suboptimal arms in a single communication round by choosing,*

$$s(p) = \left\lceil \frac{\Gamma_3}{\epsilon^2} \right\rceil \times p \ , \tag{4.41}$$

in which case the cumulative regret can be upper bounded as,

$$R_T^\beta \leq \sum_{k \in \neg\mathcal{K}_\beta} (\Delta_k - \beta)\left(\Gamma_4 \log(\frac{KT}{\delta}) + 1\right)\left(\frac{\Gamma_3}{\epsilon^2} + 1\right) + C\left(\Gamma_4 \log(\frac{KT}{\delta}) + 1\right) \ , \tag{4.42}$$

Proof. By (4.30), (4.33), and (4.41), we have $p_{\max} = 1$ and $s(1) = \Gamma_3/\epsilon^2 + \Gamma_6$ for some $\Gamma_6 \in [0, 1)$. Since $p_k = p_{\max} = 1$ for all $k \in \neg\mathcal{K}_\beta$, the result follows from (4.37) and (4.40). □

When we compare (4.42) to (4.37), we can see that even though a single round of communication is sufficient, the summation term on the right-hand side grows as $\Gamma_3/\epsilon^2 + 1 \geq s(p_k)$ for all $k \in \neg\mathcal{K}_\beta$. By design, the term governing the order of the cumulative regret with respect to the time horizon T (see (4.13)) is the number cohort size M. While the choice of $s(p)$ does not affect the order of the regret order, it determines the number of communication rounds before all the β-suboptimal arms are eliminated. Thus, taking into account communication costs and resources, the practitioner can select the best mapping $s(p)$ for the practical problem.

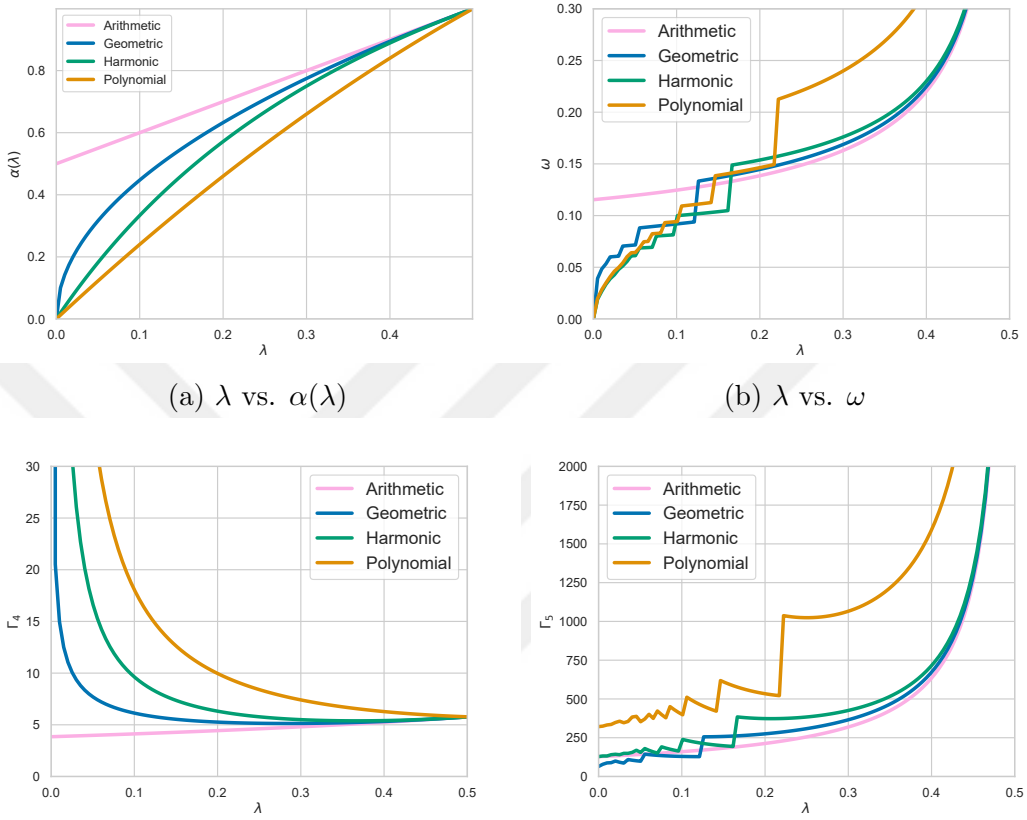
4.4.3 Choosing the mapping

Table 4.1: Different choices $\alpha(\lambda)$ satisfying Definition 4.1.

Mapping	$\rightarrow$	Arithmetic	Geometric	Harmonic	Polynomial
$\alpha(\lambda)$	$\rightarrow$	$(1 + 2\lambda)/2$	$\sqrt{2\lambda}$	$4\lambda(1 + 2\lambda)$	$\lambda(2.5 - \lambda)$

The mapping $\alpha(\lambda)$ ensures that we have enough distinct client groups to maintain robustness for a given $\lambda \in [0, 0.5)$ fraction of Byzantine clients. This section looks at how $\alpha(\lambda)$ affects various parameters. Table 4.1 offers some alternatives to the $\alpha(\lambda)$ suggested in [135]. We are primarily concerned with three key parameters directly affected by $\alpha(\lambda)$: the unavoidable error margin ω, and the constants Γ_4 and Γ_5.

The reason for the existence of the unavoidable error margin ω was briefly discussed in Remark 4.1. The more clients we recruit, the more groups we need to ensure that more than half of the groups are made up of only honest clients. Recruiting more clients does not help to increase B. However, in order to maximize the number of clients in a group, one could argue that $\alpha(\lambda)$ should be as



(a) λ vs. $\alpha(\lambda)$ (b) λ vs. ω

(c) λ vs. Γ_4 (d) λ vs. Γ_5

Figure 4.2: The effect of Byzantine clients' fraction λ and the mapping $\alpha(\lambda)$ on the unavoidable error margin ω, and the constants Γ_4 and Γ_5. The ladder structure occurs due to the dependence on the number of clients in a group, B, which is an integer. To draw the figures, we take $K = 10$, $T = 5 \times 10^4$, $\delta = 0.001$, $\sigma = 0.5$, and $\sigma_c = 0.01$.

small as possible while satisfying Definition 4.1, such as $\alpha(\lambda) = 2\lambda + \kappa$ for some arbitrarily small $\kappa > 0$. As the number of groups $G = \lceil \alpha(\lambda) M \rceil$ will decrease with $\alpha(\lambda)$, we will end up with more crowded groups as $B = \lfloor M/G \rfloor$ for the same cohort size M. The unavoidable error margin ω in (4.21) decreases for a larger B as expected. However, obtaining a larger B is not free. Observe that ω in (4.21) also accommodates the the term,

$$\left(\frac{1}{2\eta(\lambda) - 1} \right)^2 = \left(\frac{\alpha(\lambda)}{\alpha(\lambda) - 2\lambda} \right)^2 , \tag{4.43}$$

since $\eta(\lambda) = (\alpha(\lambda) - \lambda)/\alpha(\lambda)$ (see Lemma 4.1). The r.h.s. in (4.43) goes to $+\infty$ as $\alpha(\lambda) \to 2\lambda$, exploding the unavoidable error ω. That happens because

even though the groups become more crowded with a smaller $\alpha(\lambda)$, the number of groups decrease, which hurts the estimations as a median estimator is used. There is a trade-off between having more clients in a group to reduce the effect of client sampling and having more groups with fewer clients to improve precision. Figure 4.2b shows how to select the appropriate mapping to minimize the unavoidable error ω. When we compare it to Figure 4.2a, we can see that a smaller choice for $\alpha(\lambda)$ is not always the best option for a smaller ω.

Next, we look at Figures 4.2c and 4.2d to understand how Γ_4 and Γ_5 are affected by $\alpha(\lambda)$. The constant Γ_4 determines the number of clients to be recruited, M (see (4.13)). When the client recruitment is a bottleneck for the application, the practitioner should consider choosing $\alpha(\lambda)$ accordingly to obtain good performance with fewer clients. Moreover, the number of clients directly affects the cumulative regret as the arms will be played more before termination. Finally, we have $\Gamma_5 = \Gamma_3\Gamma_4$ which scales the cumulative regret in (4.36). Observe that a bigger Γ_3 implies bigger confidence bounds, E_p, making the elimination of suboptimal arms harder and increasing the cumulative regret.

The choice of $\alpha(\lambda)$ determines the unavoidable error margin and the number of clients, and it also affects the cumulative regret. Therefore, choosing a suitable mapping that best fits a specific problem's needs is crucial. In Section 4.5.3, we make experiments to investigate the effect of $\alpha(\lambda)$.

4.5 Experimental Results

In this section, we run simulations for a synthetic MAB problem with $K = 10$ arms and various attack schemes to evaluate the Fed2-UCB and Fed-MoM-UCB algorithms. We present our findings for three distinct scenarios. The main difference between the scenarios is the upper bound on the fraction of the Byzantine clients and the attack structure. In the respective sections, we explain our objectives and findings for each scenario in detail, and a comprehensive overview of the scenarios can be found in Table 4.3. The unknown global mean arm outcomes

μ_k for $k \in [10]$ are the same for all the scenarios and are given in Table 4.2. The client and arm sampling uncertainties are modeled as zero-mean additive i.i.d. Gaussian noise with variances $\sigma_c^2 = 0.01^2$ and $\sigma^2 = 0.5^2$, respectively. We use $\delta = 0.001$ and perform 100 independent experiments for each algorithm and scenario, and report the mean and standard deviations for the cumulative regret and the communication cost. We calculate the cumulative regret using (4.3) for both Fed2-UCB and Fed-MoM-UCB, where the β-optimal arms do not incur any regret.

Table 4.2: Unknown global mean arm outcomes

k	$\rightarrow$	1	2	3	4	5	6	7	8	9	10
μ_k	$\rightarrow$	0.4	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.95

4.5.1 Scenario 1

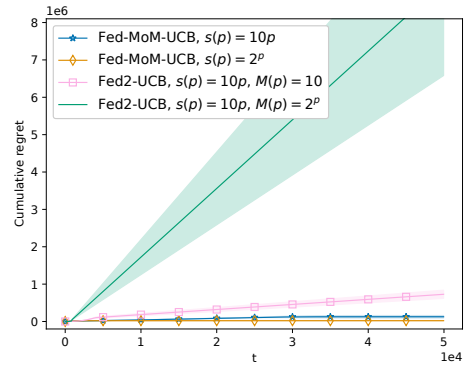
We set the fraction for Byzantine clients to $\lambda = 0.2$ and use the arithmetic mapping for Fed-MoM-UCB, which is $\alpha(\lambda) = (1+2\lambda)/2$ (see Table 4.1). We chose the arithmetic mapping because it produces the smallest value for the unavoidable error margin at $\lambda = 0.2$, which is $\omega = 0.139$, as well as the smallest constants Γ_4 and Γ_5, which scale the number of clients to be recruited and the cumulative regret. Then we set $\epsilon = 0.05$ and get $\beta = 0.189$. The β-suboptimal arm set in that case is $\neg\mathcal{K}_\beta = \{1, \ldots, 7\}$. The cost of communication is set to $C = 1$. Finally, for a Byzantine client m, consider the following attack scheme:

$$X_{m,k}(t) = \begin{cases} \texttt{Unif}[0,\ 0.1] & \text{if } k \in \mathcal{K}_\beta \\ \texttt{Unif}[0.9,\ 1] & \text{if } k \in \neg\mathcal{K}_\beta \end{cases} \tag{4.44}$$

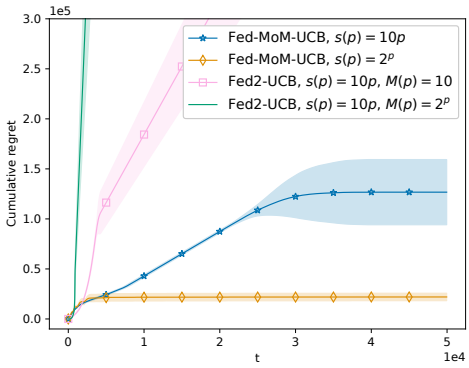
where $X_{m,k}(t)$ denotes the local observation of the Byzantine client m for the arm k at round t. That is, a Byzantine client m records a random value in $[0,\ 0.1]$ for the β-optimal arms and a random value in $[0.9,\ 1]$ for the β-suboptimal arms, to make it more challenging for the parameter server to identify the β-optimal arms.

Table 4.3: Parameters for different scenarios. The "M", "B", and "G" columns show the number of total clients recruited, the number of clients in a group, and the number of groups for the Fed-MoM-UCB algorithm, respectively. The "M(p)" and "M" columns show the number of clients recruited in phase p and the number of total clients recruited by the Fed2-UCB algorithm, respectively. ($\pm$ standard deviation).

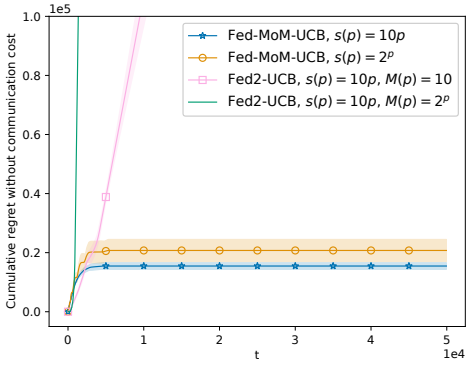
									Fed-MoM-UCB				Fed2-UCB	
	T	λ	ω	ϵ	β	$\neg\mathcal{K}_\beta$	C	Attack	$\alpha(\lambda)$	M	B	G	M(p)	M
Sce. 1	5×10^4	0.2	0.139	0.05	0.189	$\{1, \ldots, 7\}$	1	(4.44)	Arithmetic	89	1	89	10 2^p	1237 ± 88 16710 ± 2294
Sce. 2	10^5	0.2 0.25	0.139 0.149	0.05	0.189 0.199	$\{1, \ldots, 7\}$	0.1	(4.45)	Arithmetic	92 96	1 1	92 96	20	1559 ± 120 1277 ± 84
Sce. 3	5×10^4	0.1	≈ 0.1	0.01	0.11	$\{1, \ldots, 8\}$	0.1	(4.44)	Geometric Harmonic Polynomial	123 193 362	2 2 4	61 96 90	N/A	N/A



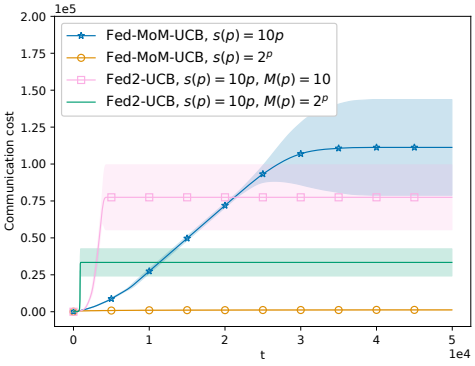
(a) Cumulative Regrets



(b) Cumulative Regrets (zoomed)



(c) Regrets without communication costs



(d) Communication costs only

Figure 4.3: Regrets and communication costs for "Scenario 1" ($\pm$ 2 std. dev.).

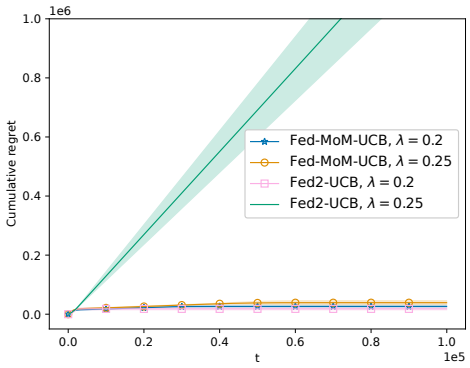
Fed2-UCB, unlike Fed-MoM-UCB, continues to recruit more clients over phases to reduce the uncertainty resulting from client sampling. That does not help in the presence of Byzantine clients, as we discussed in Section 4.4. We set $s(p) = 10p$ and run Fed2-UCB with two different configurations, $M(p) = 10$ and $M(p) = 2^p$, where $M(p)$ is the number of new clients recruited at the beginning p. Fed2-UCB terminates, as one can observe from Figure 4.3d. That is, only one active arm remains, and the communication procedure is halted. The remaining arm, however, is not a β-optimal arm because Fed2-UCB incurs linear regret in both cases, as seen in Figures 4.3a and 4.3b. Furthermore, while recruiting more clients does not assist Fed2-UCB in eliminating the β-suboptimal arms, the cumulative regret increases because the more new clients recruited, the more the suboptimal arms are played. The severe difference between the total number of

clients recruited by Fed2-UCB and Fed-MoM-UCB when they terminate is shown in Table 4.3. Fed2-UCB's unlimited client recruitment scheme not only causes the cumulative regret to explode when there are Byzantine clients, but it may also be impractical or even impossible for some applications.

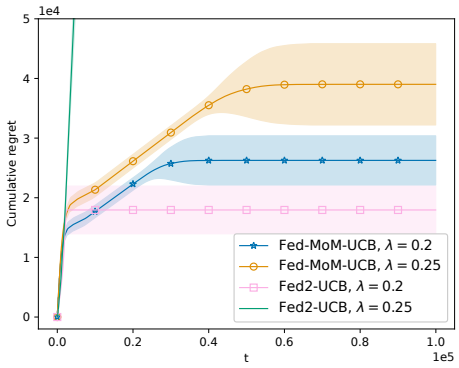
Fed-MoM-UCB recruits $M = 89$ clients only once at the beginning and creates $G = 89$ groups. In that case, it calculates the median of individual means for each arm at the parameter server. Although we can set $B = 2$ while maintaining the robustness, Figure 4.2 suggests that the arithmetic mapping is the best option when $\lambda = 0.2$. We run Fed-MoM-UCB with two different configurations, $s(p) = 10p$ and $s(p) = 2^p$. Figure 4.3d shows that Fed-MoM-UCB terminates and the communication is halted in both cases. Figure 4.3b suggests that Fed-MoM-UCB eliminates all β-suboptimal arms as the cumulative regret plateaus. Moreover, as we discussed in Remark 4.2, the communication cost is extremely smaller for $s(p) = 2^p$ compared to $s(p) = 10p$. However, Figure 4.3c shows that the cumulative regret is smaller for $s(p) = 10p$ when the communication cost is not considered (i.e., $C = 0$). That is the case as the suboptimal arms are eliminated sooner when $s(p) = 10p$ than $s(p) = 2^p$ due to more frequent communication with the parameter server. The tradeoff between the communication cost and the regret incurred from playing the suboptimal arms is crucial, and it should be carefully treated depending on the application and the communication cost, C. While the overall cumulative regret is smaller with $s(p) = 2^p$ as we can see in Fig. 4.3b, it can be the other way around when the communication cost is smaller.

4.5.2 Scenario 2

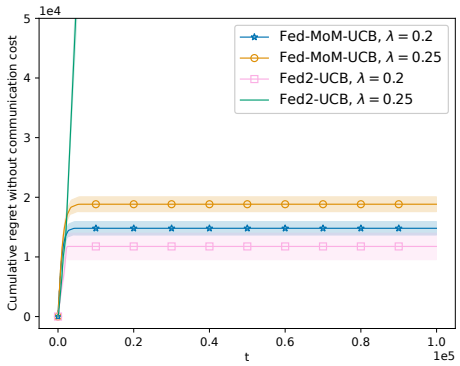
In this section, we consider a very similar experimental setup to scenario 1, except that the attack scheme is changed. We also change the communication cost to $C = 0.1$ in order to provide a more nuanced comparison in the plots in Figure 4.4. The difference in M between scenarios 1 and 2 when $\lambda = 0.2$ is caused by the longer time horizon T. We consider the following attack scheme for a



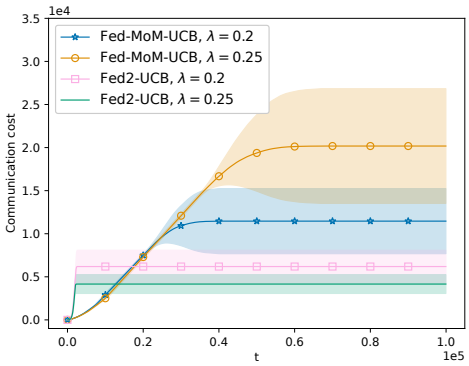
(a) Cumulative Regrets



(b) Cumulative Regrets (zoomed)



(c) Regrets without communication costs



(d) Communication costs only

Figure 4.4: Regrets and communication costs for "Scenario 2" ($\pm$ 2 std. dev.).

Byzantine client m,

$$X_{m,k}(t) = \begin{cases} \texttt{Unif}[0,\ 0.1] & \text{if } k \in \mathcal{K}_\beta \\ \mu_{m,k} + \zeta_t & \text{if } k \in \neg\mathcal{K}_\beta \end{cases} \tag{4.45}$$

That is, a Byzantine client m sets its local observations to a random value in $[0,\ 0.1]$ for the β-optimal arms as in scenario 1, while recording honest samples for the β-suboptimal arms (see (4.1)). The attack scheme in (4.45) is milder than (4.44).

We make experiments with two separate values for the fraction of Byzantine clients, $\lambda = 0.2$ and $\lambda = 0.25$. Since the set of β-suboptimal arms, $\neg\mathcal{K}_\beta$, are the same for both, we set $\beta = 0.189$ in both scenarios while calculating the cumulative regret for a meaningful comparison. In both setups, we set $s(p) = 10p$

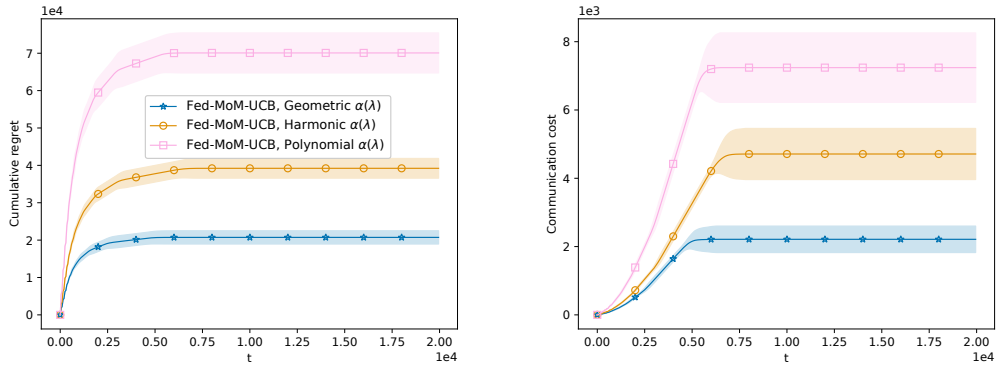
and $M(p) = 20$ for Fed2-UCB, and $s(p) = 10p$ for Fed-MoM-UCB. Figure 4.4d shows that all the algorithms terminates and the communication stops. For $\lambda = 0.2$, Fed2-UCB yields lower cumulative regret than Fed-MoM-UCB as we see in Figures 4.4a and 4.4b. The difference in the cumulative regrets partially depends on the communication cost as pointed out by Figure 4.4d. Figure 4.4c shows that the algorithms' cumulative regrets are closer when the communication cost is ignored. That is the case since even though Fed-MoM-UCB eliminates the suboptimal arms and have the active arm set $\mathcal{A}_p = \{8, 9, 10\}(\mathcal{K}_\beta)$, which does not incur any additional regret, the communication continues unless either $\mathcal{A}_p = \{10\}$ or $E_p \leq \beta/4$. Since the confidence bounds (see (4.11)) employed by Fed-MoM-UCB have a non-decreasing term due to the fixed group size B, Fed-MoM-UCB takes longer to terminate either by eliminating all the arms except the globally optimal arm (i.e., $k = 10$), or by reducing the confidence bound to the point where $E_p \leq \beta/4$.

Last but not least, although Fed2-UCB works well when $\lambda = 0.2$, it immediately fails when $\lambda = 0.25$. It still terminates when $\lambda = 0.25$ as we see in Figure 4.4d. However, it eliminates all the β-optimal arms and decides on a β-suboptimal arm as the globally optimal arm, thus yielding linear regret as we see in Figure 4.4a. While Fed2-UCB may perform well with Byzantine clients, slight changes to the attacking scheme or the fraction of the malicious clients can render it ineffective. On the other hand, Fed-MoM-UCB eliminates all the β-suboptimal arms when $\lambda = 0.25$.

4.5.3 Scenario 3

In this part, we investigate how different choices for $\alpha(\lambda)$ affects the performance. We set the communication cost to $C = 0.1$ and the Byzantine client fraction to $\lambda = 0.1$. The attack scheme is the same as scenario 1 and it follows (4.44).

We first refer to Figure 4.2 to decide on a mapping, $\alpha(\lambda)$. Figure 4.2b reveals that the unavoidable error margin is very close for the geometric, harmonic, and polynomial mappings with $\omega \approx 0.1$. Figures 4.2c and 4.2d then suggest that the



(a) Cumulative Regrets (b) Communication costs

Figure 4.5: Regrets and communication costs for "Scenario 3" ($\pm$ 2 std. dev.).

geometric mapping is the best option since it leads to the smallest constants Γ_4 and Γ_5, which scale the number of clients and the cumulative regret. Similarly, the polynomial mapping appears to be the worst option. To test our presumption, we experiment with Fed-MoM-UCB using all three mappings. As we can see from Figure 4.5a, the order between the cumulative regrets is as hypothesized. The geometric mapping yields the best performance with the minimum number of clients $M = 123$, while the polynomial mapping offers the worst with the maximum number of clients $M = 362$. As discussed in Section 4.4.3, the mapping choice is vital since it affects a variety of parameters that are not sufficient alone to decide the best mapping. The practitioner must carefully analyze the tradeoff between the unavoidable error margin ω, the number of required clients M (determined by Γ_4), and other constants which scale the cumulative regret before choosing a mapping.

Chapter 5

Combinatorial Gaussian Process Bandits with Probabilistically Triggered Arms

5.1 Introduction

In this final chapter, we study the "combinatorial multi-armed bandits with probabilistically triggered arms" (CMAB-PTA) discussed in Section 2.3. We propose the ComGP-UCB algorithm in Section 5.2 which uses Gaussian processes to model and learn the unknown expected base arm outcomes. We provide high probability upper bounds on its cumulative regret in Section 5.4.1 and demonstrate its effectiveness against the state-of-the-art baselines when the expected arm outcomes are correlated in Section 5.5.

5.2 Problem Formulation

The learner interacts with its environment through m base arms indexed by the set $[m] \coloneqq \{1, \ldots, m\}$ over rounds indexed by $t \in [T]$. Every base arm $i \in [m]$ has a context vector $\boldsymbol{x}_i \in \mathcal{X}$, where $\mathcal{X} \coloneqq [0,1]^d$ is the d-dimensional context space. The following actions take place in order at each round:

1. The learner selects a super arm $S(t) \in \mathcal{I}$, where $\mathcal{I} \subseteq 2^{[m]}$ is the feasible super arm set.
2. The base arms in $S(t)$ may trigger others depending on their outcomes through a stochastic triggering process, D^{trig}. We denote by $S'(t) \supseteq S(t)$ the set of arms after the triggering process is completed. Also, we denote the set of arms that can be possibly triggered by the arms in S by $\tilde{S}$.
3. The learner observes the individual rewards for each base arm in $S'(t)$, and a reward that depends on $S'(t)$ (*semi-bandit feedback*).

5.2.1 Base Arm Outcomes

We model the expected base arm outcomes as a sample from a Gaussian process (GP) on $\mathcal{X}$ [22]. A GP can be thought of as a probability distribution over functions or an infinite-dimensional multivariate Gaussian distribution. It is characterized by its mean, $m(\boldsymbol{x})$, and kernel, $k(\boldsymbol{x}, \boldsymbol{x}')$, functions. GPs enjoy non-parametric flexibility and analytical tractability properties. Most importantly, one can model the correlations between expected base arm outcomes through the kernel function using GPs.

At each round t, the environment realizes an outcome vector $\boldsymbol{Y}(t) \coloneqq (Y_1(t), \ldots, Y_m(t))$ drawn from the true outcome distribution $\mathcal{D}$ unknown to the learner, where $Y_i(t)$ is the outcome of base arm i at round t. We have $Y_i(t) \coloneqq \mu_i + \epsilon_{i,t}$, where $\epsilon_{i,t}$ are independent and identically distributed zero-mean

Gaussian random variables with known variance σ^2. Let $\mu_i \coloneqq \mathbb{E}_{\boldsymbol{Y}\sim\mathcal{D}}[Y_i(t)]$ denote the expected outcome of base arm i, $\boldsymbol{\mu} \coloneqq (\mu_1, \ldots, \mu_m)$ denote the expected base arm outcome vector, and $\boldsymbol{\mu}_S$ denote the projection of $\boldsymbol{\mu}$ onto S. We also denote by $f : \mathcal{X} \to \mathbb{R}$ the mapping such that $f(\boldsymbol{x}_i) = \mu_i$. We attempt to learn f efficiently using a GP surrogate.

5.2.2 Triggering Process

The learner chooses a super arm $S(t) \in \mathcal{I}$ based on its previous observations in each round t. All of the arms in the selected super arm $S(t)$ are instantly activated. These arms can trigger other arms, and triggered arms can trigger other arms again and again until no more arms are triggered. By the end of the triggering process, we have a set of triggered arms $S'(t) \sim D^{trig}(S(t), \boldsymbol{Y}(t))$ such that $S(t) \subseteq S'(t) \subseteq \tilde{S}(t) \subseteq [m]$. We describe the triggering process through a set of triggering probabilities. For every base arm $i \in [m]$, and feasible super arm $S \in \mathcal{I}$, we denote by $p_i^{\mathcal{D}',S}$ the triggering probability of the base arm i when super arm S is played and that underlying the base arm outcome distribution is $\mathcal{D}'$. We let $p_i^S \coloneqq p_i^{\mathcal{D},S}$, where $\mathcal{D} \in \mathbb{D}$ is the true arm outcome distribution, and $\mathbb{D}$ is the class of distributions that $\mathcal{D}$ belongs to. $\mathcal{D}$ is unknown by the learner, but $\mathbb{D}$ is known. In our case, $\mathbb{D}$ is the class of m-dimensional multivariate Gaussian distributions as we assume a GP prior over the expected base arm outcomes. We define $p_i \coloneqq \min\{p_i^S \mid S \in \mathcal{I}, p_i^S > 0\}$ as the minimum non-zero triggering probability of the base arm i, where $i \in \{1, \ldots, m\}$ and $p^* \coloneqq \min_{i\in[m]} p_i$ as the minimum non-zero triggering probability.

5.2.3 Reward

Upon selecting a super arm $S(t)$, the learner observes a reward, $R(S'(t), \boldsymbol{Y}(t))$ that depends on the set of triggered arms $S'(t)$ and the true expected base arm outcome vector. To keep the notation uncluttered, we use $R(t) = R(S'(t), \boldsymbol{Y}(t))$.

Assumption 5.1. The expected reward of a super arm is only a function of the

super arm and the true outcome vector, and denote $\mathbb{E}[R(t)|\boldsymbol{\mu}] = r(S(t), \boldsymbol{\mu})$.

We denote by $OPT(\boldsymbol{\theta}) \coloneqq \text{argmax}_{S \in \mathcal{I}} r(S, \boldsymbol{\theta})$ the set of optimal super arms for the expected base arm outcome θ. OPT denotes the exact oracle, which is a problem-specific computation black box. The oracle knows the problem structure (e.g., cascading bandit, influence maximization), and it can solve for the optimal super-arm for a given expected base arm outcome vector (i.e., $\boldsymbol{\theta}$). For instance, in the influence maximization problem, the oracle knows the graph structure, and it can provide the optimal solution given the triggering probabilities of the edges. Let $S^* \in OPT(\boldsymbol{\mu})$ represent an optimal super arm. The learner aims to minimize the cumulative regret over the time horizon T, which can then be written as,

$$Reg_{\boldsymbol{\mu}}(T) \coloneqq \sum_{t=1}^{T} r(S^*, \boldsymbol{\mu}) - r(S(t), \boldsymbol{\mu}) \ .$$

Next, we make some standard assumptions on the reward function R [106, 107, 137].

Assumption 5.2. (Monotonicity) The expected reward, $r(S, \boldsymbol{\mu})$, is non-decreasing w.r.t. $\boldsymbol{\mu}$, that is, given $\boldsymbol{\mu}$, $\boldsymbol{\nu} \in \mathbb{R}^M$ such that $\mu_i \geq \nu_i$, $\forall i \in [M]$, we have $r(S, \boldsymbol{\mu}) \geq r(S, \boldsymbol{\nu}), \forall S \in \mathcal{I}$.

Assumption 5.3. (Lipschitz continuity) There exists a constant $B > 0$, such that for every super arm S and every pair of mean base arm outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we have $|r(S, \boldsymbol{\mu}) - r(S, \boldsymbol{\nu})| \leq B||\boldsymbol{\mu}_{\tilde{S}} - \boldsymbol{\nu}_{\tilde{S}}||_1$, where $||\cdot||_1$ denotes the ℓ_1 norm.

Assumption 5.4. (Triggering probability modulated Lipschitz continuity) There exists a constant $\tilde{B} > 0$, such that for every super arm S and every pair of base arm outcome distributions $\mathcal{D}$ and $\mathcal{D}'$ with mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ respectively, we have,

$$|r(S, \boldsymbol{\mu}) - r(S, \boldsymbol{\nu})| \leq \tilde{B} \sum_{i \in \tilde{S}} p_i^{\mathcal{D}, S} |\mu_i - \nu_i| \ .$$

5.2.4 Observation Model

Individual outcomes of every base arm in $S'(t)$ are revealed after each round. That is, $Q(S'(t), \boldsymbol{Y}(t)) \coloneqq \{(i, Y_i(t) : i \in S'(t))\}$ is observed. We let $Q(t) = Q(S'(t), \boldsymbol{Y}(t))$. By these definitions, the information available to the learner to guide its actions in round $t+1$ is $\mathcal{F}_t \coloneqq \{(S(\tau), Q(\tau)) : \tau \in [t]\}$. After making a set of observations, we can update the posterior mean and variance at $\boldsymbol{x} \in \mathcal{X}$ as follows,

$$\begin{aligned} k_{N(t)}(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x},\ \boldsymbol{x}') \\ &\quad - \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T (\boldsymbol{K}_{N(t)} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}_{N(t)}(\boldsymbol{x}') \end{aligned} \tag{5.1}$$

$$\hat{\sigma}^2_{N(t)}(\boldsymbol{x}) = k_{N(t)}(\boldsymbol{x},\ \boldsymbol{x}) \tag{5.2}$$

$$\hat{\mu}_{N(t)}(\boldsymbol{x}) = \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T (\boldsymbol{K}_{N(t)} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y}_t, \tag{5.3}$$

where $\boldsymbol{Y}_t \coloneqq [\boldsymbol{Y}^T(1), \ldots, \boldsymbol{Y}^T(t-1)]^T$ denotes the vector of observations made until round t, and $N(t) \coloneqq |\boldsymbol{Y}_t|$ denotes the total number of observations made until round t. Moreover, $k_{N(t)}(\boldsymbol{x},\ \boldsymbol{x}')$ denotes the posterior covariance between $\boldsymbol{x}$ and $\boldsymbol{x}'$, and $\hat{\mu}_{N(t)}(\boldsymbol{x})$ and $\hat{\sigma}^2_{N(t)}(\boldsymbol{x})$ denote the posterior mean and variance at $\boldsymbol{x} \in \mathcal{X}$ at round t, respectively. $\boldsymbol{k}_{N(t)}(\boldsymbol{x}) \coloneqq [k(\boldsymbol{x}^1,\ \boldsymbol{x}), \ldots, k(\boldsymbol{x}^{N(t)},\ \boldsymbol{x})]^T$ denotes the vector of covariances between $\boldsymbol{x} \in \mathcal{X}$, and past observations $[\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{N(t)}]$, where $\boldsymbol{x}^i$ is the ith base arm picked from the beginning. $\boldsymbol{K}_{N(t)}$ is the gram matrix, $\boldsymbol{I}$ is the $N(t) \times N(t)$ identity matrix, and σ^2 is the noise variance that we include in our calculations.

To summarize, $([m], \mathcal{I}, \mathcal{D}, D^{trig}, R)$ tuple describes a CMAB-PTA problem instance, among which only $\mathcal{D}$ is unknown to the learner.

5.3 ComGP-UCB Algorithm

Pseudo-code for ComGP-UCB is given in Algorithm 3, and Figure 5.1 demonstrates the algorithm flow. ComGP-UCB uses Gaussian process based upper confidence bounds to form optimistic estimates of the expected base arm outcomes [10]. First, it initiates a GP, g, in the base arm domain. At the beginning

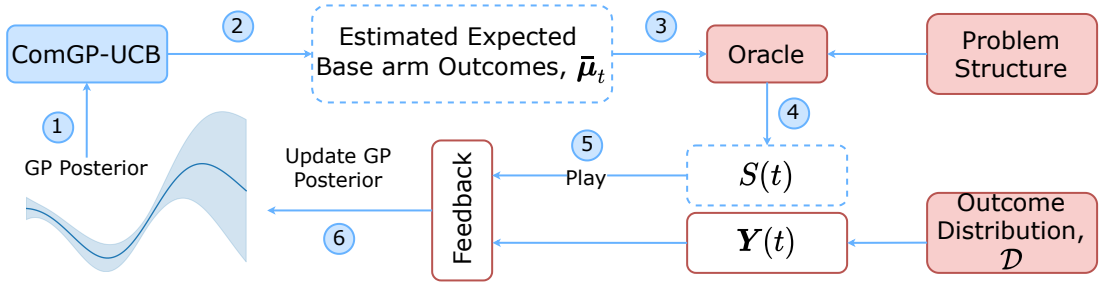


Figure 5.1: *Algorithm Flow.* At each round t, the environment draws a random outcome vector $\boldsymbol{Y}(t)$ from the true outcome distribution $\mathcal{D}$. ComGP-UCB first forms an optimistic estimate for the expected base arm outcome vector, $\bar{\boldsymbol{\mu}}_t$, using the posterior GP distribution and feeds it to the oracle. The oracle knows the problem structure and it returns an optimal super-arm $S(t)$ for the given parameter set, $\bar{\boldsymbol{\mu}}_t$. Then, the learner plays the super-arm $S(t)$ and observes a feedback, $Q(t)$, which depends on $\boldsymbol{Y}(t)$. Finally, the learner updates the GP distribution based on the feedback it observes. Blue boxes are the variables that are created and updated by the learner, whereas the red boxes are external and environmental entities.

Algorithm 3 ComGP-UCB

Initialize $g \sim GP(\hat{\mu}_0(\boldsymbol{x}),\ k_0(\boldsymbol{x}, \boldsymbol{x}'))$
for $t = 1, \dots, T$ **do**
  $\bar{\boldsymbol{\mu}}_t = \{\hat{\mu}_{N(t)}(\boldsymbol{x}_i) + \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)\}_{i=1}^m$
  $S(t) = OPT(\bar{\boldsymbol{\mu}}_t)$
  Play $S(t)$ and observe $Q(t)$
  Update $\mathcal{F}_t$ based on $Q(t)$
  Update $\hat{\mu}_{N(t+1)}(\boldsymbol{x}_i)$ and $\hat{\sigma}_{N(t+1)}(\boldsymbol{x}_i)$, $i \in [m]$
end for

of each round, the algorithm calculates an upper confidence bound (UCB) for base arms as follows,

$$\bar{\mu}_t(\boldsymbol{x}_i) = \hat{\mu}_{N(t)}(\boldsymbol{x}_i) + \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i),$$

where $i \in [m]$, and $\hat{\mu}_{N(t)}$ and $\hat{\sigma}^2_{N(t)}$ denote the posterior mean and variance vectors of g at round t, respectively. Without loss of generality, we assume bounded variance, that is, $\hat{\sigma}^2_{N(t)}(x) = k_{N(t)}(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ [10]. $\beta_n := \frac{8}{3}\log(\frac{mT\pi_n}{\delta})$ is the exploration coefficient where $\sum_{n=1}^{\infty} 1/\pi_n = 1$ (e.g., $\pi_n = \pi^2 n^2/6$). Note that β_n is strictly increasing in n. We define $\bar{\boldsymbol{\mu}}_t := [\bar{\mu}_t(\boldsymbol{x}_i), \dots, \bar{\mu}_t(\boldsymbol{x}_m)]$ to denote the vector of estimated optimistic expected base arm outcomes at round t. Then, $\bar{\boldsymbol{\mu}}_t$

is passed to the exact oracle to select the super-arm in round t, $S(t)$.[1] After $S(t)$ is selected, $Q(t)$ is observed and the mean vector and covariance matrix of g are updated according to (5.1) and (5.3).

5.4 Theoretical Analysis

5.4.1 Main results

In this section, we bound the regret of ComGP-UCB when the reward function satisfies (TPM) Lipschitz continuity. First, we derive an upper bound on the regret when the reward function satisfies Lipschitz continuity.

Theorem 5.1. *Given $\delta,\ \rho \in (0,1)$, $\alpha \in [mT]$, and under Assumptions 5.1, 5.2, and 5.3, the cumulative regret of ComGP-UCB after round T is upper bounded with at least $1-\delta-\frac{m}{\alpha}-\frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ probability as follows,*

$$\mathbb{P}\big\{Reg_{\boldsymbol{\mu}}(T) \leq 4mB\sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2m\alpha B\sqrt{\beta_{mT}}\big\} \geq 1-\delta-\frac{m}{\alpha}-\frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]\ ,$$

where B is the Lipschitz constant.

We observe that the right hand-side in Theorem 5.1 is greater than zero when $\alpha > \frac{m+\frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{1-\delta}$. This condition is translated into a condition on T in the following corollary. Also, note that $\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ should be finite in order for this bound to make sense. One simple example is the case when $p^* \geq p^*_{\min}$ for all $\boldsymbol{\mu}$ for a constant $p^*_{\min} > 0$.

Corollary 5.1. *Setting $\alpha = \sqrt{T}$ in Theorem 5.1, we have,*

$$\mathbb{P}\big\{Reg_{\boldsymbol{\mu}}(T) \leq 4mB\sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2mB\sqrt{T\beta_{mT}}\big\} \geq 1-2\delta\ ,$$

when $\sqrt{T} > \frac{m+\frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{\delta}$.

[1] We note that our regret analysis can easily be extended to the cases when approximation oracles are used, by redefining the regret to be α-approximation regret [104].

Next, we bound the regret when the expected reward function satisfies the TPM Lipschitz continuity. The regret bound for this case depends on a new term called *probabilistic triggering modulated (pseudo-) information gain.*

Definition 5.1. (Probabilistic Triggering Modulated Pseudo-Information Gain)

$$\gamma_{T,\boldsymbol{\mu}}^{PTA} \coloneqq \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in\tilde{S}(t)}(p_i^{S(t)})^2\log(1+\sigma^{-2}\hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i))\ . \tag{5.4}$$

Probabilistic triggering modulated pseudo-information gain is very similar to the standard information gain term [10]. In the following lemma, we consider the CMAB framework with no probabilistically triggered arms, in which the super-arm S is built using a greedy approach. That is, within round t, the learner chooses a single base arm, observes its outcome, and updates the algorithm's parameters before moving on to the next base arm. We consider such a case to provide an analogy between the standard information gain term and the *probabilistic triggering modulated (pseudo-) information gain* term. More information can be found in the supplementary material.

Lemma 5.1. *The information gain by the end of round T in a CMAB problem (no probabilistic triggering) where the super-arm S is constructed with a greedy approach can be expressed as,*

$$I(\boldsymbol{y}_{[T]};\boldsymbol{f}_{[T]}) = \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)} log(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)), \tag{5.5}$$

where $\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)$ is the conditional variance at $\boldsymbol{x}_i \in \mathcal{X}$ conditioned on all the selected base arms before ith arm at round t .

However, there are two critical distinctions between the two terms. The first is that when calculating the pseudo-information gain, we sum over the triggering set $\tilde{S}(t)$. Even if a base arm is not triggered, it still contributes to the sum, a situation that arises from the probabilistic triggering segment of the problem under consideration. Another distinction is that ComGP-UCB performs batch updates on the mean vector and covariance matrix; they are updated after a

round is completed rather than after each base arm is selected and its outcome is observed. Despite the differences, pseudo-information gain characterizes a piece of information similar to the standard information gain term. Although we do not provide explicit theoretical bounds on the pseudo-information gain term, we show how it grows over time in Section 5.5. We are now ready to present the following theorem.

Theorem 5.2. *Given $\delta \in (0,1)$, $C \coloneqq \frac{8\tilde{B}^2}{\log(1+\sigma^{-2})}$, where $\tilde{B}$ is the TPM Lipschitz constant, and under Assumptions 5.1, 5.2, and 5.4, the cumulative regret of ComGP-UCB after round T is upper bounded with at least $1-\delta$ probability as follows,*

$$\mathbb{P}\big\{Reg_{\boldsymbol{\mu}}(T) \leq \sqrt{CmT\beta_{mT}\gamma^{PTA}_{T,\boldsymbol{\mu}}}\big\} \geq 1-\delta.$$

In terms of time dependence, the bound in Theorem 5.2 matches the standard results in GP bandits [10]. However, due to the combinatorial nature of the problem, we have an extra $\sqrt{m}$ term, as well as the pseudo-information gain term instead of the standard information gain term. It should be noted that the vanilla GP-UCB algorithm can also be used to solve a CMAB problem by treating each super arm as a single base arm, with the context vector of each super-arm defined as the concatenation of the context vectors of its base arms. In this case, the $\sqrt{m}$ term in the regret bound will vanish. This approach, however, has one immediate disadvantage: the information related to underlying base arms shared by different super arms will be discarded. Another disadvantage is that the number of possible super arms grows combinatorially with the number of base arms, and GP-based inference on a high-dimensional space may degrade performance.

Theorem 5.3. *Let f be bounded such that $c_l \leq f(\boldsymbol{x}) \leq c_u$ for $\boldsymbol{x} \in \mathcal{X}$, $c_l, c_u \in \mathbb{R}$, and $K \coloneqq c_u - c_l$. Given $\delta \in (0,1)$, $\alpha \in [mT]$, and under Assumptions 5.1, 5.2, and 5.4, the cumulative regret of ComGP-UCB after round T is upper bounded with at least $1-\delta-\frac{1}{m\alpha}$ probability as follows,*

$$\mathbb{P}\big\{Reg_{\boldsymbol{\mu}}(T) \leq 2m\tilde{B}K + 4(2+\sqrt{2})m\tilde{B}K\sqrt{T\beta_{mT}} + 2m\alpha\tilde{B}K\big\} \geq 1-\delta-\frac{1}{m\alpha}\ ,$$

where $\tilde{B}$ is the TPM Lipschitz constant.

Corollary 5.2. *Setting $\alpha = \sqrt{T}$ in Theorem 5.3, we have,*

$$\mathbb{P}\big\{Reg_{\boldsymbol{\mu}}(T) \leq 2m\tilde{B}K + 4(2+\sqrt{2})m\tilde{B}K\sqrt{T\beta_{mT}} + 2m\tilde{B}K\sqrt{T}\big\} \geq 1 - 2\delta \ ,$$

when $\sqrt{T} > \frac{1}{m\delta}$.

5.4.2 Preliminaries for Theorems 1 and 2

Let us define the following events,

$$\begin{aligned} H_i(t) &\coloneqq \{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\} \qquad (5.6)\\ H(t) &\coloneqq \{\exists i \in [m] : H_i(t)\} \\ \mathcal{G} &\coloneqq \{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| \leq \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \\ &\qquad \forall i \in [m], \forall t \in [T]\} \\ \mathcal{J} &\coloneqq \Big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} < \alpha\Big\}, \qquad (5.7) \end{aligned}$$

where $\alpha \in [mT]$ is fixed, $M_i(t) \coloneqq \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in \tilde{S}(\tau)\}$ denotes the number of times base arm i was in the triggering set, and $N_i(t) \coloneqq \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in S'(\tau)\}$ denotes the number of observations made at base arm i, up to round t. Also, let τ_w^i denote the round number where ith base arm is in the triggering set of the selected super arm for the wth time, and $\tau_0^i = 0$.

5.4.3 Supplemental Lemmas

We have the following three lemmas for which the detailed proofs can be found in the supplementary material.

Lemma 5.2. *Given $\delta \in (0,1)$, the event $\mathcal{G}$ holds with at least $1-\delta$ probability.*

Lemma 5.3. *Given $\alpha \in [mT]$, the event $\mathcal{J}$ holds with at least $1 - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ probability.*

Lemma 5.4. *Instantaneous variance of the Gaussian process at $\boldsymbol{x}_i \in \mathcal{X}$ can be upper bounded by the noise variance and the number of times base arm i was triggered as, $\sigma^2/N_i(t) \geq \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)$.*

5.4.4 Proof of Theorem 1

For the proof of Theorem 1, we assume that the events $\mathcal{G}$ and $\mathcal{J}$ hold, which means that $\bar{\mu}_t(\boldsymbol{x}_i) \geq \mu_i$ for all $i \in [m]$ and $t \in [T]$, and $\sum_{i=1}^{m}\sum_{t=1}^{T} \mathbb{I}\{H_i(t)\} < \alpha$ for a fixed $\alpha \in [mT]$.

$$\begin{aligned}
\mathrm{Reg}_{\boldsymbol{\mu}}(T) &= \sum_{t=1}^{T} r(S^*, \boldsymbol{\mu}) - r(S(t), \boldsymbol{\mu}) \\
&\leq \sum_{t=1}^{T} r(S^*, \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) && (5.8)\\
&\leq \sum_{t=1}^{T} r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) && (5.9)\\
&\leq \sum_{t=1}^{T} \mathbb{I}\{\neg H(t)\}(r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu})) && (5.10)\\
&\quad + \sum_{t=1}^{T} \mathbb{I}\{H(t)\} \times 2mB\sqrt{\beta_{mT}}, && (5.11)
\end{aligned}$$

where (5.8) follows from Assumption 5.2, (5.9) is due to the definition of the OPT operator, and (5.11) is obtained by observing that for all $S \in \mathcal{I}$,

$$\begin{aligned}
&|r(S, \bar{\boldsymbol{\mu}}_t) - r(S, \boldsymbol{\mu})| \\
&\quad \leq B \sum_{i \in \tilde{S}} |\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)| && (5.12)\\
&\quad \leq B \sum_{i \in \tilde{S}} \left(|\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i)| + |\sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)|\right) && (5.13)\\
&\quad \leq 2B \sum_{i \in \tilde{S}} \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) && (5.14)\\
&\quad \leq 2mB\sqrt{\beta_{mT}}, && (5.15)
\end{aligned}$$

where (5.12) follows from Assumption 5.3, (5.14) holds since $\mathcal{G}$ holds, and (5.15) holds since $mT \geq N(t)$ for all $t \in [T]$, and $\hat{\sigma}^2_{N(t)}(x_i) \leq 1$ since we assume bounded variance.

5.4.4.1 Bounding (5.10)

When $\mathcal{G}$ and $\neg H(t)$ hold, we have,

$$
\begin{aligned}
&|r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu})| \\
&\leq 2B \sum_{i \in \tilde{S}(t)} \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) && (5.16) \\
&\leq 2B \sum_{i \in \tilde{S}(t)} \sqrt{\beta_{N(t)}} \frac{\sigma}{\sqrt{N_i(t)}} && (5.17) \\
&\leq 2B \sum_{i \in \tilde{S}(t)} \sqrt{\beta_{N(t)}} \frac{\sigma}{\sqrt{(1-\rho) p_i M_i(t)}}, && (5.18)
\end{aligned}
$$

where (5.16) follows from (5.14), (5.17) follows from Lemma 5.4, and (5.18) holds since $\neg H(t)$ holds. Then,

$$
\begin{aligned}
(5.10) &\leq \sum_{t=1}^{T} \mathbb{I}\{\neg H(t)\} 2B \sum_{i \in \tilde{S}(t)} \sqrt{\beta_{N(t)}} \frac{\sigma}{\sqrt{(1-\rho) p_i M_i(t)}} \\
&\leq \sqrt{\beta_{mT}} \sum_{t=1}^{T} \mathbb{I}\{\neg H(t)\} 2B \sum_{i \in \tilde{S}(t)} \frac{\sigma}{\sqrt{(1-\rho) p_i M_i(t)}} \\
&\leq \sqrt{\beta_{mT}} \sum_{i=1}^{m} \sum_{w=0}^{T} \sum_{t=\tau^i_w+1}^{\tau^i_{w+1}} \mathbb{I}\{i \in \tilde{S}(t), \neg H(t)\} \times 2B \frac{\sigma}{\sqrt{(1-\rho) p_i M_i(t)}} \\
&\leq \sqrt{\beta_{mT}} \sum_{i=1}^{m} \sum_{w=0}^{T} \mathbb{I}\{\neg H(\tau^i_{w+1})\} \times 2B \frac{\sigma}{\sqrt{(1-\rho) p_i M_i(\tau^i_{w+1})}} \\
&\leq \sqrt{\beta_{mT}} \sum_{i=1}^{m} \sum_{w=1}^{T} 2B \frac{\sigma}{\sqrt{(1-\rho) p^* w}} && (5.19) \\
&\leq 4mB \sqrt{\frac{T \beta_{mT} \sigma^2}{(1-\rho) p^*}}, && (5.20)
\end{aligned}
$$

where (5.19) holds since $M_i(\tau^i_1) = N_i(\tau^i_1) = 0$ for all i, meaning $\neg H(\tau^i_1)$ does not hold. Finally, (5.20) follows from the fact $\sum_{n=1}^{N} \sqrt{1/n} \leq 2\sqrt{N}$.

5.4.4.2 Bounding (5.11)

$$
\begin{aligned}
(5.11) &= 2mB\sqrt{\beta_{mT}}\sum_{t=1}^{T}\mathbb{I}\{H(t)\} \\
&\leq 2mB\sqrt{\beta_{mT}}\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} \\
&\leq 2m\alpha B\sqrt{\beta_{mT}}, \qquad (5.21)
\end{aligned}
$$

where (5.21) holds since $\mathcal{J}$ holds. Finally, when the events $\mathcal{G}$ and $\mathcal{J}$ hold, we can bound the cumulative regret after T rounds as follows,

$$
\begin{aligned}
Reg_{\boldsymbol{\mu}}(T) &\leq (5.10) + (5.11) \\
&\leq 4mB\sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2m\alpha B\sqrt{\beta_{mT}}.
\end{aligned}
$$

Theorem 5.1 then follows after combining Lemmas 5.2 and 5.3 with the fact that $\mathbb{P}\{\mathcal{G}\cap\mathcal{J}\} = 1-\mathbb{P}\{\neg\mathcal{G}\cup\neg\mathcal{J}\}$, where $1-\mathbb{P}\{\neg\mathcal{G}\cup\neg\mathcal{J}\} \geq 1-\mathbb{P}\{\neg\mathcal{G}\}-\mathbb{P}\{\neg\mathcal{J}\}$—by union bound—. Finally, $1-\mathbb{P}\{\neg\mathcal{G}\}-\mathbb{P}\{\neg\mathcal{J}\} = \mathbb{P}\{\mathcal{G}\}+\mathbb{P}\{\mathcal{J}\}-1$, and $\mathbb{P}\{\mathcal{G}\}+\mathbb{P}\{\mathcal{J}\}-1 \geq 1-\delta-\frac{m}{\alpha}-\frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ by Lemmas 5.2 and 5.3.

5.4.5 Proof of Theorem 2

For the proof of Theorem 5.2, we assume that the event $\mathcal{G}$ holds.

$$
\begin{aligned}
Reg_{\boldsymbol{\mu}}(T) &\leq \sum_{t=1}^{T} r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) \qquad (5.22) \\
&\leq 2\tilde{B}\sum_{t=1}^{T}\sum_{i\in\tilde{S}(t)} p_i^{S(t)}\sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \qquad (5.23) \\
&\leq 2\tilde{B}\sqrt{\beta_{mT}}\sum_{t=1}^{T}\sum_{i\in\tilde{S}(t)} p_i^{S(t)}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \qquad (5.24)
\end{aligned}
$$

where (5.22) follows from (5.9), (5.24) holds since $mT \geq N(t)$ for all $t \in [T]$, and (5.23) is obtained by observing that for all $S \in \mathcal{I}$,

$$
\begin{aligned}
&|r(S, \bar{\boldsymbol{\mu}}_t) - r(S, \boldsymbol{\mu})| \\
&\leq \tilde{B} \sum_{i \in \tilde{S}} p_i^S |\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i)| && (5.25) \\
&\leq \tilde{B} \sum_{i \in \tilde{S}} p_i^S \big(|\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i)| + \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \big) && (5.26) \\
&\leq 2\tilde{B} \sum_{i \in \tilde{S}} p_i^S \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i), && (5.27)
\end{aligned}
$$

where (5.25) follows from Assumption 5.4, and (5.27) holds since $\mathcal{G}$ holds. Let us define $p_{i,S(t)} = p_i^{S(t)}$ to keep the notation uncluttered. By Cauchy-Schwartz inequality, we can write,

$$
\begin{aligned}
Reg_{\boldsymbol{\mu}}^2(T) &\leq 4\tilde{B}^2 T \beta_{mT} \sum_{t=1}^{T} \Big(\sum_{i \in \tilde{S}(t)} p_{i,S(t)} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \Big)^2 \\
&\leq 4\tilde{B}^2 mT \beta_{mT} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) && (5.28) \\
&= 4\tilde{B}^2 mT \beta_{mT} \sigma^2 \times \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \sigma^{-2} \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \\
&\leq 8\tilde{B}^2 mT \beta_{mT} \times \frac{1}{2} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \frac{\log(1 + \sigma^{-2} \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i))}{\log(1 + \sigma^{-2})} && (5.29) \\
&= CmT \beta_{mT} \gamma_{T,\boldsymbol{\mu}}^{PTA}, && (5.30)
\end{aligned}
$$

where $C \coloneqq \frac{8\tilde{B}^2}{log(1+\sigma^{-2})}$, is a constant, and $\gamma_{T,\boldsymbol{\mu}}^{PTA}$ is *pseudo-information gain* in Definition 5.1. (5.28) follows applying Cauchy-Schwarz inequality again, (5.29) follows from the fact that $s^2 \leq \frac{\sigma^{-2} log(1+s^2)}{log(1+\sigma^{-2})}$ for $s \in [0, \sigma^{-2}]$, and $\sigma^{-2} \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \leq \sigma^{-2}$, since $\hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \leq 1$.

5.5 Experimental Results

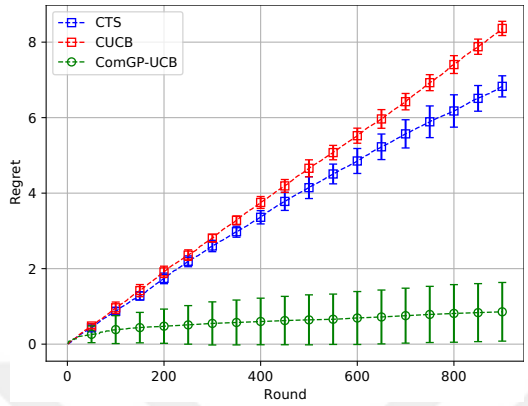
5.5.1 Cascading Bandit

In a disjunctive cascading bandit problem (DCB) [12], which can be treated as a CMAB-PTA instance [108], we compare the performance of ComGP-UCB against other state-of-the-art algorithms. We consider a scenario where a search engine tries to recommend the most appealing list of web pages to its users. The users arrive one by one, and the engine chooses K pages from a set of R pages for each one. Each K-sized super-arm (i.e., S) is represented by a virtual arm that is immediately triggered. The base arms that can be triggered in virtual arm S make up $\tilde{S}$. In this experiment, the oracle (OPT) returns an ordered list of K best webpages from R webpages based on their estimated click probabilities at each t round (i.e., $\bar{\boldsymbol{\mu}}_t$). The user looks through the pages in the recommended order and clicks on the first one she finds interesting. If a user finds none of the pages appealing, she does not click on them. The search engine's goal is to increase the total number of clicks.

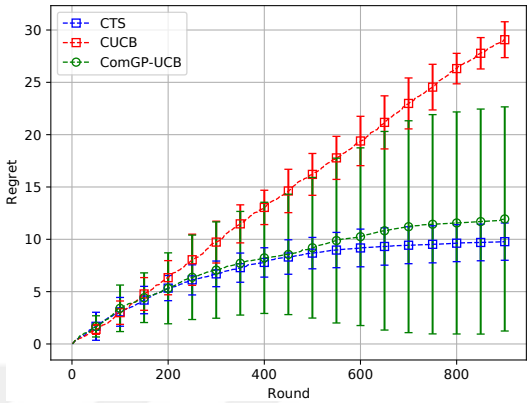
Each webpage $j \in [R]$ is treated as a base arm. The user is interested in the page j with a probability of p_j. The super arms are lists of size K that contain the user's recommended webpages. Given a super arm S, let $S(k)$ denote the k-th page recommended to the user. The triggering probabilities for each base arm can then be expressed in S as follows,

$$p^S_{(j)} = \begin{cases} 1 & \text{if } j = S(1) \\ \prod_{k'=1}^{k-1}(1 - p_{S(k')}) & \text{if } \exists k \neq 1 : j = S(k) \\ 0 & \text{otherwise} \end{cases}$$

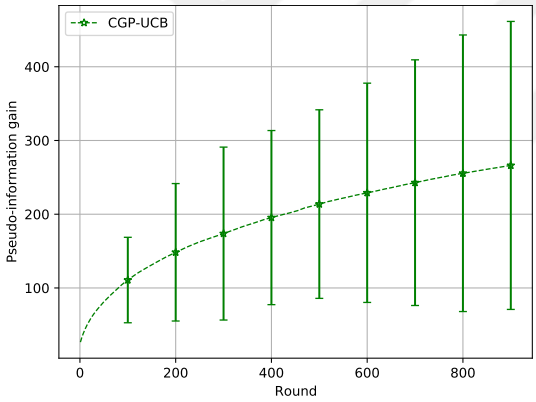
We monitor feedback for the first recommendations (pages) immediately and monitor feedback for the subsequent recommendations only if the user does not click on the previous recommendations. The expected reward for playing super



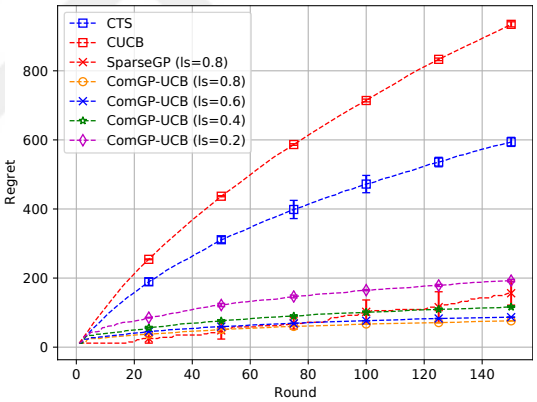
(a) Regrets for DCB, when there is high correlation between expected base arm outcomes.



(b) Regrets for DCB, when expected base arm outcomes are sampled randomly.



(c) Pseudo-information gain term over time.



(d) Regrets for SLB with sparse GPs True length-scale (ls) is set to 0.8.

Figure 5.2: Regrets and communication costs for "Scenario 1" ($\pm$ 2 std. dev.).

arm S is then written as,

$$r(S, \boldsymbol{p}) = \left(1 - \prod_{k=1}^{K} (1 - p_{S(k)})\right)$$

for which Assumptions 5.3 and 5.4 hold when $B = 1$ and $\tilde{B} = 1$.

We consider a scenario where the recommendation engine is tasked to pick 10 webpages among 900 web pages to recommend to a user, that is, $R = 900$ and $K = 10$. We assume that webpages have 2-dimensional context vectors ($\boldsymbol{x}$) and generate p_j's by first sampling from a Gaussian process with squared-exponential

kernel[2] on a 2D grid, and then passing the output through a sigmoid function [138]. We run ComGP-UCB, CUCB in [107], and CTS in [139] for 900 rounds and average their regrets over 10 different runs. Note that CTS and CUCB do not use the side information (context) related to webpages to exploit the dependencies between their click probabilities. In Figure 5.2a we see that ComGP-UCB outperforms its competitors. That is because, unlike ComGP-UCB, other algorithms cannot utilize the correlations between base arm outcomes. On the other hand, when p_j are sampled randomly, we see in Figure 5.2b that ComGP-UCB is similar to CTS in performance as expected. Figure 5.2c shows that the pseudo-information gain term, $\gamma^{PTA}_{T,\boldsymbol{\mu}}$, increases sublinearly in time.

5.5.2 Sparse GPs and Hyper-parameter Mismatch

A particular setback of GPs is that they are computationally expensive. ComGP-UCB has a computational complexity of $O(n^3)$, where n denotes the number of base arm observations. One can consider using sparse GPs for inference to reduce the computational complexity to $O(m^2 n)$, where m is a chosen constant (number of inducing points) [140]. We test the sparse GPs' performance in a synthetic linear bandit (SLB) problem where the reward of a super-arm S equals the sum of its base arms' rewards. Formally, we have $R(S(t), \boldsymbol{Y}(t)) = \sum_{i \in S(t)} \bar{r}_i^t$, where $\bar{r}_i^t$ denotes the outcome of base arm $i \in [m]$ at round t. Figure 5.2d shows that sparse GPs still outperform CTS and CUCB and have a competitive performance against ComGP-UCB. Furthermore, we assume that the kernel hyper-parameters are known in advance. These parameters can be estimated from a training dataset or tuned with domain knowledge. Figure 5.2d shows that ComGP-UCB outperforms its competitors even when there is a hyper-parameter mismatch; that is, the GP does not know the true kernel hyper-parameters a priori.

[2]We set variance parameter to 2, lengthscale parameter to 0.8, and likelihood variance to 0.01.

Chapter 6

Concluding Remarks and Future Directions

In Chapter 3, we studied a prevalent problem in medicine, *safe leveling*, and proposed TACO, a novel acquisition strategy tailored for this problem structure. As safety is crucial in healthcare, we proposed a safe exploration strategy to render TACO safe. Combining these two blocks, we proposed ESCADA, a *safe* and *efficient* learning algorithm, and provided safety guarantees and upper bounds on its cumulative regret. We made extensive *in silico* experiments on the bolus-insulin dose allocation problem for type-1 diabetes disease to show our algorithms' effectiveness over the rule-based dose calculators and straightforward adaptations of the GP-UCB algorithm. We also compared ESCADA's performance against a clinician's to provide external validation and discussed its potential as a complementary instrument in clinical settings. ESCADA can also be used in other safety-critical decision-making problems where the goal is to control a target variable safely.

In Chapter 4, we studied the federated multi-armed problem considering Byzantine clients and proposed a robust median-of-means-based algorithm, Fed-MoM-UCB. We showed that when at most half the client pool consists of Byzantine clients, it is possible to maintain robustness and provide an upper bound on

the cumulative regret while also considering the communication cost, a primary bottleneck of the federated learning systems. We showed using a median-based estimator leads to an unavoidable error margin due to the non-increasing size of the client groups, and that that recruiting more clients does not alleviate that problem. We derived an expression for the required number of clients to provide high probability guarantees on the cumulative regret. We analyzed the interplay between the different algorithm parameters and performance metrics. Finally, we compared Fed-MoM-UCB's performance against Fed2-UCB's and confirmed its effectiveness under Byzantine attacks. Interesting future work includes a detailed investigation on reducing the error margin emanating from client sampling while retaining robustness against adversaries.

In Chapter 5, we proposed the ComGP-UCB algorithm to operate in a CMAB-PTA setting, and we provided guarantees on its cumulative regret under (triggering probability modulated) Lipschitz continuity and monotonicity assumptions on the expected reward function. One of our bounds includes the $1/p^*$ term that is unavoidable. The other includes a pseudo-information gain term for which we have provided simulations regarding its behavior in time. Interesting future work includes a more detailed investigation of the pseudo-information gain, for instance, providing explicit upper bounds. Another exciting research direction may include getting rid of the $1/p^*$ term. Overall, we showed that when the expected base-arm outcomes are not completely independent but rather a sample from a GP, for example, ComGP-UCB algorithm significantly outperforms the state-of-the-art benchmarks CTS and CUCB.

Bibliography

[1] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to Reinforcement Learning*. MIT Press Cambridge, 1998.

[2] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[4] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[5] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory (COLT)*, pp. 39–1, 2012.

[6] R. Agrawal, "Sample mean based index policies by o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[7] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 24, pp. 2312–2320, 2011.

[8] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[9] A. Garivier and O. Cappé, "The kl-ucb algorithm for bounded stochastic bandits and beyond," in *Conference on Learning Theory (COLT)*, pp. 359–376, 2011.

[10] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: no regret and experimental design," in *International Conference on Machine Learning (ICML)*, pp. 1015–1022, 2010.

[11] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

[12] B. Kveton, C. Szepesvári, Z. Wen, and A. Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *International Conference on Machine Learning (ICML)*, pp. 767–776, 2015.

[13] Z. Wen, B. Kveton, M. Valko, and S. Vaswani, "Online influence maximization under independent cascade model with semi-bandit feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3026–3036, 2017.

[14] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[15] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Royal Society Open Science*, vol. 4, no. 11, p. 171377, 2017.

[16] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *ACM International Conference on Web Search and Data Mining*, pp. 357–365, 2019.

[17] V. Srivastava, P. Reverdy, and N. E. Leonard, "Surveillance in an abruptly changing world via multiarmed bandits," in *IEEE Conference on Decision and Control*, pp. 692–697, 2014.

[18] M. Y. Cheung, J. Leighton, and F. S. Hover, "Autonomous mobile acoustic relay positioning as a multi-armed bandit with switching costs," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3368–3373, 2013.

[19] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[20] S. S. Villar, J. Bowden, and J. Wason, "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*, vol. 30, no. 2, p. 199, 2015.

[21] C. Shen, Z. Wang, S. Villar, and M. Van Der Schaar, "Learning for dose allocation in adaptive clinical trials with safety constraints," in *International Conference on Machine Learning (ICML)*, pp. 8730–8740, 2020.

[22] C. E. Rasmussen, "Gaussian processes in machine learning," *Advanced Lectures on Machine Learning*, 2004.

[23] A. Bastaki *et al.*, "Diabetes mellitus and its treatment," *International Journal of Diabetes and Metabolism*, vol. 13, no. 3, p. 111, 2005.

[24] A. Cassel, S. Mannor, and A. Zeevi, "A general approach to multi-armed bandits under risk criteria," in *Conference On Learning Theory (COLT)*, pp. 1295–1306, 2018.

[25] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.

[26] S. Vakili and Q. Zhao, "Mean-variance and value at risk in multi-armed bandit problems," in *Annual Allerton Conference on Communication, Control, and Computing*, pp. 1330–1335, IEEE, 2015.

[27] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[28] M. C. Steinbach, "Markowitz revisited: Mean-variance models in financial portfolio analysis," *SIAM Review*, vol. 43, no. 1, pp. 31–85, 2001.

[29] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *Asian Conference on Machine Learning*, pp. 245–260, 2013.

[30] A. R. Cardoso and H. Xu, "Risk-averse stochastic convex bandit," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 39–47, 2019.

[31] N. Khajonchotpanya, Y. Xue, and N. Rujeerapaiboon, "A revised approach for risk-averse multi-armed bandits under CVaR criterion," *Operations Research Letters*, vol. 49, no. 4, pp. 465–472, 2021.

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

[33] Y. Wu, R. Shariff, T. Lattimore, and C. Szepesvári, "Conservative bandits," in *International Conference on Machine Learning (ICML)*, pp. 1254–1262, 2016.

[34] T. M. Moldovan and P. Abbeel, "Safe exploration in Markov decision processes," in *International Conference on Machine Learning (ICML)*, 2012.

[35] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe exploration and optimization of constrained MDPs using Gaussian processes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[36] M. A. Gelbart, J. Snoek, and R. P. Adams, "Bayesian optimization with unknown constraints," in *Uncertainty in Artificial Intelligence (UAI)*, pp. 250–259, 2014.

[37] J. M. Hernandez-Lobato, M. A. Gelbart, R. P. Adams, M. W. Hoffman, Z. Ghahramani, *et al.*, "A general framework for constrained Bayesian optimization using information-based search," *Journal of Machine Learning Research*, vol. 17, no. 160, pp. 1–53, 2016.

[38] G. R. Petroni, N. A. Wages, G. Paux, and F. Dubois, "Implementation of adaptive methods in early-phase clinical trials," *Statistics in Medicine*, vol. 36, no. 2, pp. 215–224, 2017.

[39] S. Amani, "Linear stochastic bandits under safety constraints," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] K. Khezeli and E. Bitar, "Safe linear stochastic bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10202–10209, 2020.

[41] A. Moradipari, S. Amani, M. Alizadeh, and C. Thrampoulidis, "Safe linear thompson sampling with side information," *IEEE Transactions on Signal Processing*, 2021.

[42] M. Abeille and A. Lazaric, "Linear Thompson Sampling Revisited," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 176–184, 2017.

[43] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe convex learning under uncertain constraints," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2106–2114, 2019.

[44] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Regret bound for safe Gaussian process bandit optimization," in *Conference on Learning for Dynamics and Control*, pp. 158–159, 2020.

[45] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with Gaussian processes," in *International Conference on Machine Learning (ICML)*, pp. 997–1005, 2015.

[46] Y. Sui, vincent Zhuang, J. Burdick, and Y. Yue, "Stagewise safe Bayesian optimization with Gaussian processes," in *International Conference on Machine Learning (ICML)*, pp. 4781–4789, 2018.

[47] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration for interactive machine learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[48] R. Mirnezami, J. Nicholson, and A. Darzi, "Preparing for precision medicine," *New England Journal of Medicine*, vol. 366, no. 6, pp. 489–491, 2012.

[49] E. A. Ashley, "Towards precision medicine," *Nature Reviews Genetics*, vol. 17, no. 9, p. 507, 2016.

[50] M. D. Kraft, I. F. Btaiche, G. S. Sacks, and K. A. Kudsk, "Treatment of electrolyte disorders in adult patients in the intensive care unit," *American Journal of Health-System Pharmacy*, vol. 62, no. 16, pp. 1663–1682, 2005.

[51] K. A. Nerenberg, K. B. Zarnke, A. A. Leung, K. Dasgupta, S. Butalia, K. McBrien, K. C. Harris, M. Nakhla, L. Cloutier, M. Gelfer, *et al.*, "Hypertension Canada?s 2018 guidelines for diagnosis, risk assessment, prevention, and treatment of hypertension in adults and children," *Canadian Journal of Cardiology*, vol. 34, no. 5, pp. 506–525, 2018.

[52] C. Shi and C. Shen, "Federated multi-armed bandits," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, pp. 9603–9611, 2021.

[53] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *International Conference on Machine Learning (ICML)*, pp. 1467–1474, 2012.

[54] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Annual Network and Distributed System Security Symposium*, 2017.

[55] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed parameter estimation with heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4918–4933, 2019.

[56] L. Lamport, R. E. Shostak, and M. C. Pease, "The byzantine generals problem," in *Concurrency: The Works of Leslie Lamport* (D. Malkhi, ed.), pp. 203–226, 2019.

[57] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, 2016.

[58] K. Bonawitz, H. Eichner, *et al.*, "Towards federated learning at scale: System design," in *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.

[59] P. Kairouz, McMahan, *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[60] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 65–75, 2013.

[61] Y. Chen, S. Kar, and J. M. Moura, "The internet of things: Secure distributed inference," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 64–75, 2018.

[62] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[63] C. Szegedy, W. Zaremba, *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[64] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2938–2948, 2020.

[65] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[66] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[67] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *International Conference on Machine Learning (ICML)*, pp. 634–643, 2019.

[68] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, and P. S. Yu, "Privacy and robustness in federated learning: Attacks and defenses," *arXiv preprint arXiv:2012.06337*, 2020.

[69] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis, "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4583–4596, 2020.

[70] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning (ICML)*, pp. 5650–5659, 2018.

[71] R. Guerraoui, S. Rouault, *et al.*, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning (ICML)*, pp. 3521–3530, 2018.

[72] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to {Byzantine-Robust} federated learning," in *USENIX Security Symposium*, pp. 1605–1622, 2020.

[73] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1893–1905, 2014.

[74] S. Shen, S. Tople, and P. Saxena, "Auror: Defending against poisoning attacks in collaborative deep learning systems," in *Annual Conference on Computer Security Applications*, p. 508?519, 2016.

[75] X. Wei and C. Shen, "Federated learning over noisy channels," in *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2021.

[76] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3796–3811, 2021.

[77] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[78] K. Liu and Q. Zhao, "Decentralized multi-armed bandit with multiple distributed players," in *IEEE Information Theory and Applications Workshop (ITA)*, pp. 1–10, 2010.

[79] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[80] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proceedings of IEEE INFOCOM*, pp. 1–9, 2010.

[81] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *IEEE European Control Conference (ECC)*, pp. 243–248, 2016.

[82] P. Landgren, V. Srivastava, and N. E. Leonard, "Social imitation in cooperative multiarmed bandits: Partition-based algorithms with strictly local information," in *IEEE Conference on Decision and Control (CDC)*, pp. 5239–5244, 2018.

[83] D. Martínez-Rubio, V. Kanade, and P. Rebeschini, "Decentralized cooperative stochastic bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[84] Z. Zhu, J. Zhu, J. Liu, and Y. Liu, "Federated bandit: A gossiping approach," in *ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pp. 3–4, 2021.

[85] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[86] T. Li and L. Song, "Privacy-preserving communication-efficient federated multi-armed bandits," *IEEE Journal on Selected Areas in Communications*, 2022.

[87] A. Dubey and A. Pentland, "Differentially-private federated linear bandits," *Advances in Neural Information Processing Systems, (NeurIPS)*, vol. 33, pp. 6003–6014, 2020.

[88] R. Huang, W. Wu, J. Yang, and C. Shen, "Federated linear contextual bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[89] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331, 1995.

[90] K.-S. Jun, L. Li, Y. Ma, and J. Zhu, "Adversarial attacks on stochastic bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[91] F. Liu and N. Shroff, "Data poisoning attacks on stochastic bandits," in *International Conference on Machine Learning (ICML)*, pp. 4042–4050, 2019.

[92] T. Lykouris, V. Mirrokni, and R. Paes Leme, "Stochastic bandits robust to adversarial corruptions," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

[93] A. Gupta, T. Koren, and K. Talwar, "Better algorithms for stochastic bandits with adversarial corruptions," in *Conference on Learning Theory (COLT)*, pp. 1562–1578, 2019.

[94] I. Bogunovic, A. Krause, and J. Scarlett, "Corruption-tolerant gaussian process bandit optimization," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1071–1081, 2020.

[95] E. Garcelon, B. Roziere, L. Meunier, J. Tarbouriech, O. Teytaud, A. Lazaric, and M. Pirotta, "Adversarial attacks on linear contextual bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 14362–14373, 2020.

[96] I. Bogunovic, A. Losalka, A. Krause, and J. Scarlett, "Stochastic linear bandits robust to adversarial attacks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 991–999, 2021.

[97] J. Altschuler, V.-E. Brunel, and A. Malek, "Best arm identification for contaminated bandits," *Journal of Machine Learning Research*, vol. 20, no. 91, pp. 1–39, 2019.

[98] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7711–7717, 2013.

[99] A. Mitra, H. Hassani, and G. Pappas, "Exploiting heterogeneity in robust federated best-arm identification," *arXiv preprint arXiv:2109.05700*, 2021.

[100] D. Vial, S. Shakkottai, and R. Srikant, "Robust multi-agent multi-armed bandits," in *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 161–170, 2021.

[101] S. Maghsudi and E. Hossain, "Multi-armed bandits with application to 5g small cells," *IEEE Wireless Communications*, vol. 23, no. 3, pp. 64–73, 2016.

[102] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

[103] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 535–543, 2015.

[104] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning (ICML)*, pp. 151–159, 2013.

[105] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning (ICML)*, pp. 100–108, 2014.

[106] A. Hüyük and C. Tekin, "Thompson sampling for combinatorial network optimization in unknown environments," *IEEE/ACM Transactions on Networking*, vol. 28, no. 6, pp. 2836–2849, 2020.

[107] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1746–1778, 2016.

[108] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1161–1171, 2017.

[109] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[110] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[111] Y. Zhang, E. B. Laber, M. Davidian, and A. A. Tsiatis, "Interpretable dynamic treatment regimes," *Journal of the American Statistical Association*, vol. 113, no. 524, pp. 1541–1549, 2018.

[112] S. Vakili, K. Khezeli, and V. Picheny, "On information gain and regret bounds in gaussian process bandits," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 82–90, 2021.

[113] A. Krause and C. S. Ong, "Contextual Gaussian process bandit optimization.," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2447–2455, 2011.

[114] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," *Annual Review of Biomedical Data Science*, vol. 4, 2020.

[115] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: addressing ethical challenges," *PLoS Medicine*, vol. 15, no. 11, p. e1002689, 2018.

[116] W. Price and I. Nicholson, "Regulating black-box medicine," *Mich. L. Rev.*, vol. 116, p. 421, 2017.

[117] W. N. Price, "Big data and black-box medical algorithms," *Science Translational Medicine*, vol. 10, no. 471, 2018.

[118] M. Aziz, E. Kaufmann, and M.-K. Riviere, "On multi-armed bandit designs for dose-finding clinical trials," *Journal of Machine Learning Research*, vol. 22, pp. 1–38, 2021.

[119] H.-S. Lee, C. Shen, J. Jordon, and M. Schaar, "Contextual constrained learning for dose-finding clinical trials," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2645–2654, 2020.

[120] J. Xie, "Simglucose v0.2.1." url=https://github.com/jxx123/simglucose, 2018.

[121] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, "In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes," 2009.

[122] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[123] T. Zhu, K. Li, L. Kuang, P. Herrero, and P. Georgiou, "An insulin bolus advisor for type 1 diabetes using deep reinforcement learning," *Sensors*, vol. 20, no. 18, p. 5058, 2020.

[124] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial Intelligence in Medicine*, vol. 104, p. 101836, 2020.

[125] A. Huang, L. Leqi, Z. Lipton, and K. Azizzadenesheli, "Off-policy risk assessment in contextual bandits," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[126] S. Lee, J. Kim, S. W. Park, S.-M. Jin, and S.-M. Park, "Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 536–546, 2020.

[127] I. Fox, J. Lee, R. Pop-Busui, and J. Wiens, "Deep reinforcement learning for closed-loop blood glucose control," in *Machine Learning for Healthcare Conference (ML4HC)*, pp. 508–536, 2020.

[128] Y. Chandak, S. Niekum, B. da Silva, E. Learned-Miller, E. Brunskill, and P. S. Thomas, "Universal off-policy evaluation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[129] Y. Chandak, G. Theocharous, S. Shankar, M. White, S. Mahadevan, and P. Thomas, "Optimizing for the future in non-stationary mdps," in *International Conference on Machine Learning (ICML)*, pp. 1414–1425, 2020.

[130] B. P. Kovatchev, D. J. Cox, A. Kumar, L. Gonder-Frederick, and W. L. Clarke, "Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data," *Diabetes Technology & Therapeutics*, vol. 5, no. 5, pp. 817–828, 2003.

[131] C. R. Tomson and S. N. Van Der Veer, "Learning from practice variation to improve the quality of care," *Clinical Medicine*, vol. 13, no. 1, p. 19, 2013.

[132] J. Walsh, R. Roberts, and T. Bailey, "Guidelines for optimal bolus calculator settings in adults," *Journal of Diabetes Science and Technology*, vol. 5, no. 1, pp. 129–135, 2011.

[133] S. Schmidt and K. Nørgaard, "Bolus calculators," *Journal of Diabetes Science and Technology*, vol. 8, no. 5, pp. 1035–1041, 2014.

[134] E. Even-Dar, S. Mannor, Y. Mansour, and S. Mahadevan, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems.," *Journal of Machine Learning Research*, vol. 7, no. 6, 2006.

[135] P. Laforgue, G. Staerman, and S. Clémençon, "Generalization bounds in the presence of outliers: a median-of-means study," in *International Conference on Machine Learning (ICML)*, pp. 5937–5947, 2021.

[136] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[137] S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," in *International Conference on Machine Learning (ICML)*, pp. 5114–5122, 2018.

[138] M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman, "A framework for interdomain and multioutput Gaussian processes," *arXiv:2003.01115*, 2020.

[139] A. Huyuk and C. Tekin, "Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," in *International*

Conference on Artificial Intelligence and Statistics (ICML), pp. 1322–1330, 2019.

[140] M. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 567–574, 2009.

Appendix A

ESCADA: Efficient Safety and Context-Aware Dose Allocation for Precision Medicine

A.1 Proof of Theorem 3.1

We prove by induction when f is L-Lipschitz continuous and the event $\mathcal{E}$ in Lemma 3.1 holds. When $\mathcal{E}$ holds, we have $f(\boldsymbol{z}_n, d) \in [\mu_{n-1}(\boldsymbol{z}_n, d) - \beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}_n, d), \mu_{n-1}(\boldsymbol{z}_n, d) + \beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}_n, d)]$, that is,

$$l_n(\boldsymbol{z}_n, d) \leq f(\boldsymbol{z}_n, d) \leq u_n(\boldsymbol{z}_n, d) \ ,$$

for every $d \in \mathcal{D}$ and $\boldsymbol{z}_n \in \mathcal{Z}$. Also, by Lipschitz continuity, we have,

$$\begin{aligned} f(\boldsymbol{z}_n, d) &\geq f(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d, d') \\ &\geq l_n(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d, d') \ , \end{aligned}$$

and,

$$\begin{aligned} f(\boldsymbol{z}_n, d) &\leq f(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d, d') \\ &\leq u_n(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d, d') \ , \end{aligned}$$

that is,

$$l_n(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d, d') \leq f(\boldsymbol{z}_n, d) \leq u_n(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d, d') \ ,$$

for any $d, d' \in \mathcal{D}$.

Remember, we have,

$$\bar{l}_n(\boldsymbol{z}, d) = \max\left\{l_n(\boldsymbol{z}, d), l_n(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d')\right\} \tag{A.1}$$

$$\bar{u}_n(\boldsymbol{z}, d) = \min\left\{u_n(\boldsymbol{z}, d), u_n(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d')\right\} \ , \tag{A.2}$$

where $d' = \operatorname{argmin}_{d^* \in \bar{\mathcal{D}}_z} q_{\boldsymbol{z}}(d, d^*)$. Then, we have, $f(\boldsymbol{z}_n, d) \geq \bar{l}_n(\boldsymbol{z}_n, d)$ and $f(\boldsymbol{z}_n, d) \leq \bar{u}_n(\boldsymbol{z}_n, d)$. For the base case, $n = 0$, we have $f(\boldsymbol{z}, d) \in [T_{\min}, T_{\max}]$ for any $d \in S_0(\boldsymbol{z})$ by definition. For the inductive step, first consider the safety condition from below, that is, $f(\boldsymbol{z}_n, d_n) \geq T_{\min}$. For $n \geq 1$, assume that $f(\boldsymbol{z}_n, d) \geq T_{\min}$ for any $d \in S_{n-1}(\boldsymbol{z}_n)$. Then by definition of the safe set expansion operator, for all $d^* \in S_n(\boldsymbol{z}_n) \setminus S_{n-1}(\boldsymbol{z}_n)$, there exists $d' \in S_{n-1}(\boldsymbol{z}_n)$ such that,

$$\begin{aligned} T_{\min} &\leq \bar{l}_n(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d', d^*) \\ &\leq f(\boldsymbol{z}_n, d') - Lq_{\boldsymbol{z}_n}(d', d^*) \\ &\leq f(\boldsymbol{z}_n, d^*) \ . \end{aligned} \tag{A.3}$$

Similarly, to check the condition from above, we can observe,

$$\begin{aligned} T_{\max} &\geq \bar{u}_n(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d', d^*) \\ &\geq f(\boldsymbol{z}_n, d') + Lq_{\boldsymbol{z}_n}(d', d^*) \\ &\geq f(\boldsymbol{z}_n, d^*) \ , \end{aligned} \tag{A.4}$$

where (A.3) and (A.4) follow from Lipschitz continuity, and we are done.

A.2 Proof of Theorem 3.2

Assume $\mathcal{E}$ in Lemma 3.1 holds. Then,

$$\begin{aligned} R_N &= \sum_{n=1}^{N} \mid f(\boldsymbol{z}_n, d_n) - T \mid \\ &\leq \sum_{n=1}^{N} \big(\mid f(\boldsymbol{z}_n, d_n) - \mu_{n-1}(\boldsymbol{z}_n, d_n) \mid + \mid (\mu_{n-1}(\boldsymbol{z}_n, d_n) - T) \mid \big) \qquad \text{(A.5)} \\ &\leq \sum_{n=1}^{N} 2\beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d_n) \ , \qquad \text{(A.6)} \end{aligned}$$

where (A.5) follows from triangle inequality on $\mathbb{R}$. For the first term in (A.5), we have,

$$|f(\boldsymbol{z}_n, d_n) - \mu_{n-1}(\boldsymbol{z}_n, d_n)| \leq \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d_n) \ ,$$

by Lemma 3.1 under the good event $\mathcal{E}$. For the second term, notice that in the first step of TACO, we choose the doses whose confidence intervals contain the target value T, i.e., $\bar{l}_n(\boldsymbol{z}_n, d_n) \leq T \leq \bar{u}_n(\boldsymbol{z}_n, d_n)$. By (A.1) and (A.2), we have,

$$l_n(\boldsymbol{z}_n, d_n) \leq T \leq u_n(\boldsymbol{z}_n, d_n) \ ,$$

which is equivalent to,

$$\mu_{n-1}(\boldsymbol{z}_n, d_n) - \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d_n) \leq T \leq \mu_{n-1}(\boldsymbol{z}_n, d_n) + \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d_n) \ , \quad \text{(A.7)}$$

by definitions of $l_n(\boldsymbol{z}_n, d_n)$ and $u_n(\boldsymbol{z}_n, d_n)$ in (3.3) and (3.4). Finally, we have, $|\mu_{n-1}(\boldsymbol{z}_n, d_n) - T| \leq \beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}_n, d_n)$ by (A.7). Note that since there is not any safety constraint, we have at least one dose in $\mathbb{D}_n = \mathcal{D}$ (e.g., $d^*_{\boldsymbol{z}_n}$) whose confidence interval contains the target value, T, which makes the passage from (A.5) to (A.6)

always possible. Then, by Cauchy-Schwartz inequality,

$$
\begin{aligned}
R_N^2 &\leq N \sum_{n=1}^{N} 4\beta_n \sigma_{n-1}^2(\boldsymbol{z}_n, d_n) \\
&\leq 4N\beta_N \sigma^2 \sum_{n=1}^{N} \sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}_n, d_n) && \text{(A.8)} \\
&\leq 8N\beta_N \sigma^2 \frac{1}{2} \sum_{n=1}^{N} \frac{\sigma^{-2}\log(1+\sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}_n, d_n))}{\log(1+\sigma^{-2})} && \text{(A.9)} \\
&\leq CN\beta_N \max \frac{1}{2} \sum_{n=1}^{N} \log(1+\sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}_n, d_n)) \\
&\leq CN\beta_N \gamma_N^{vol1} \ , && \text{(A.10)}
\end{aligned}
$$

where (A.8) holds since β_n is increasing in n. (A.9) follows from the fact that $s^2 < \frac{\sigma^{-2}log(1+s^2)}{log(1+\sigma^{-2})}$ for $s \in [0, \sigma^{-2}]$, and $\sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}_n, d_n) \leq \sigma^{-2}k(\boldsymbol{x}_n, \boldsymbol{x}_n) \leq \sigma^{-2}$, and (A.10) follows from combining Lemma 3.2 and the definition of γ_N^{vol1} in Section 4. Theorem 3.2 follows from taking square roots of both sides.

A.3 Auxiliary Results

Before we prove Theorem 3.3, it is useful to highlight the critical difference between the setup considered in Theorem 3.2 and Theorem 3.3, and motivate the proof of Theorem 3.3. In Theorem 3.2, note that the passage between (A.5) and (A.6) is always possible. This is because it is always guaranteed that there exists at least one dose $d \in \mathbb{D}_n$, whose confidence interval contains the target, T, where $\mathbb{D}_n = \mathcal{D}$ (e.g. $d_{\boldsymbol{z}_n}^*$). That allows us to upper-bound the term $\mid (\mu_{n-1}(\boldsymbol{z}_n, d_n) - T) \mid$ by $\beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}_n, d_n)$. However, for the setup considered in Theorem 3.3, we are not allowed to choose any $d \in \mathcal{D}$, but choose $d \in S_n(\boldsymbol{z}_n)$ in round n, to guarantee the safety of the recommendation. Until we are certain that the optimal dose $d_{\boldsymbol{z}}^* \in S_n(\boldsymbol{z})$, we will incur a linear regret at the worst case. Once we have $d_{\boldsymbol{z}}^* \in S_n(\boldsymbol{z})$, the regret will converge to the same order as in Theorem 3.1, since the relation between (A.5) and (A.6) will be valid.

In the next three lemmas, we show that if $d_{\boldsymbol{z}}^* \in R_\epsilon^m(S_0(\boldsymbol{z}))$ for some $m \in \mathbb{N}$,

we have $d^*_{\boldsymbol{z}} \in S_n(\boldsymbol{z})$ after a finite number of rounds, and then the the passage between (A.5) and (A.6) is also possible for the safety constrained setting in a single context scenario.

Lemma A.1. *Posterior variance of the Gaussian process after round n at $\boldsymbol{x} \in \mathcal{X}$ can be upper bounded by the noise variance and the number of times the function evaluated at $\boldsymbol{x}$ up to round n ($n_{\boldsymbol{x}}$) as, $\sigma^2/n_{\boldsymbol{x}} \geq \sigma^2_{n-1}(\boldsymbol{x})$.*

Proof. Given $A \subseteq B$, $H(\mu|A) \geq H(\mu|B)$, since conditioning on more observations will reduce entropy. Let $\boldsymbol{Y}^{\boldsymbol{x}}_n$ denote the observations made at $\boldsymbol{x} \in \mathcal{X}$ up to round n, and $\boldsymbol{Y}_n$ denote all the observations up to round n. Then, we have $\boldsymbol{Y}^{\boldsymbol{x}}_n \subseteq \boldsymbol{Y}_n$, which implies $H(\mu|\boldsymbol{Y}^{\boldsymbol{x}}_n) \geq H(\mu|\boldsymbol{Y}_n)$. The entropy of a Gaussian random variable is $H(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2}\log(2\pi e\sigma^2)$. We have $H(\mu|\boldsymbol{Y}_n) = \frac{1}{2}\log(2\pi e\sigma^2_{n-1}(\boldsymbol{x}))$, and $H(\mu|\boldsymbol{Y}^{\boldsymbol{x}}_n) = \frac{1}{2}\log(2\pi e(n_{\boldsymbol{x}}/\sigma^2 + \sigma_0^{-2})^{-1}))$, where $\sigma_0^2 = k(\boldsymbol{x}, \boldsymbol{x})$ is the prior variance of GP at $\boldsymbol{x}$. Then we can write,

$$\frac{1}{2}\log(\frac{2\pi e}{n_{\boldsymbol{x}}/\sigma^2 + \sigma_0^{-2}}) \geq \frac{1}{2}\log(2\pi e\sigma^2_{n-1}(\boldsymbol{x})) \ . \tag{A.11}$$

Since $k(\boldsymbol{x}, \boldsymbol{x}) \geq 0$, the following is immediate from (A.11),

$$\frac{\sigma^2}{n_{\boldsymbol{x}}} \geq \sigma^2_{n-1}(\boldsymbol{x}) \ .$$

□

Let us first to recall the definition of a *safe path*.

Definition A.1 (Definition 3.1 restated)**.** For a fixed context $\boldsymbol{z} \in \mathcal{Z}$, we say that there exists a safe path between two doses $d_1, d_2 \in \mathcal{D}$ if the following is satisfied,

$$\eta(d_1, d_2) = \min\left(\min_{d\in[d_1,d_2]}\big(T_{\max} - \epsilon - f(\boldsymbol{z}, d)\big), \min_{d\in[d_1,d_2]}\big(f(\boldsymbol{z}, d) - T_{\min} - \epsilon\big)\right) > 0 \ , \tag{A.12}$$

Lemma A.2. *For a fixed context $\boldsymbol{z} \in \mathcal{Z}$, if there exists at least one dose $d \in S_0(\boldsymbol{z})$ such that there exists a safe path between d and $d^*_{\boldsymbol{z}}$, and we have $q_{\boldsymbol{z}}(d^1, d^2) = K(|d^1 - d^2|)$ for some monotonically increasing mapping $K : \mathbb{R}^+ \to \mathbb{R}^+$ and for all $d^1, d^2 \in \mathcal{D}$, then we have $d^*_{\boldsymbol{z}} \in R^m_\epsilon(S_0(\boldsymbol{z}))$ for some $m \in \mathbb{N}$.*

Proof. We first note that $R_\epsilon^l(A) \subseteq R_\epsilon^l(B)$ for any $l \in \mathbb{N}$ if $A \subseteq B$ by definition of the reachability operator in (3.6). That is, $R_\epsilon^k(\{d\}) \subseteq R_\epsilon^k(S_0(\boldsymbol{z}))$ since $d \in S_0(\boldsymbol{z})$. At the heart of the proof lies the idea that if there is a safe path between d and $d_{\boldsymbol{z}}^*$, the reachability operator will keep expanding towards $d_{\boldsymbol{z}}^*$. Now, if $d = d_{\boldsymbol{z}}^*$, we are already done. Consider the case where $d_{\boldsymbol{z}}^* > d$. Let $d_1 > d$ be such that,

$$Lq_{\boldsymbol{z}}(d, d_1) = \eta \ , \tag{A.13}$$

where $\eta \coloneqq \eta(d, d_{\boldsymbol{z}}^*)$. Then, by (A.12), we have,

$$\begin{aligned} &Lq_{\boldsymbol{z}}(d, d_1) \leq T_{\max} - \epsilon - f(\boldsymbol{z}, d) \\ \Rightarrow &f(\boldsymbol{z}, d) + Lq_{\boldsymbol{z}}(d, d_1) + \epsilon \leq T_{\max} \ , \end{aligned} \tag{A.14}$$

and,

$$\begin{aligned} &Lq_{\boldsymbol{z}}(d, d_1) \leq f(\boldsymbol{z}, d) - T_{\min} - \epsilon \\ \Rightarrow &f(\boldsymbol{z}, d) - Lq_{\boldsymbol{z}}(d, d_1) - \epsilon \geq T_{\min} \ . \end{aligned} \tag{A.15}$$

By (A.14), (A.15), and (3.6), we have $d_1 \in R_\epsilon^1(\{d\})$, where $d_1 \geq d$ and $q_{\boldsymbol{z}}(d, d_1) = \eta/L$. Now, if $q_{\boldsymbol{z}}(d, d_1) = K(|d - d_1|)$ for some monotonically increasing mapping $K : \mathbb{R}^+ \to \mathbb{R}^+$, where $K(x) = 0 \iff x = 0$ (since $q_{\boldsymbol{z}}(\cdot, \cdot)$ is a metric), then we have $|d - d_1| = K^{-1}(\eta/L)$, which implies that,

$$d_1 = d + K^{-1}(\eta/L) \ , \tag{A.16}$$

and $d_1 > d$ since $\eta > 0$. Now, if $d_1 \geq d_{\boldsymbol{z}}^* > d$, one can simply check that $Lq_{\boldsymbol{z}}(d, d_{\boldsymbol{z}}^*) \leq \eta$ by (A.13) since K is monotonically increasing. This then implies that $d_{\boldsymbol{z}}^* \in R_\epsilon^1(\{d\})$, and we are done. If $d_{\boldsymbol{z}}^* > d_1$, we will keep expanding the reachable set. Note that we have $R_\epsilon^1(\{d_1\}) \subseteq R_\epsilon^2(\{d\})$, since $\{d_1\} \subseteq R_\epsilon^1(\{d\})$, where $R_\epsilon^2(\{d\}) = R_\epsilon^1(R_\epsilon^1(\{d\}))$. Similar to before, we can show that there exists $d_2 \in \mathcal{D}$ such that $d_2 \in R_\epsilon^1(\{d_1\})$ and,

$$\begin{aligned} d_2 &= d_1 + K^{-1}(\eta/L) \\ &= d + 2K^{-1}(\eta/L) \ , \end{aligned} \tag{A.17}$$

which follows from (A.16). By this iteration, we have $d_n \in R_\epsilon^n(\{d\})$, where

$$d_n = d + nK^{-1}(\eta/L) \ , \tag{A.18}$$

for $n \in \mathbb{N}$ such that $d_n \leq d^*_{\boldsymbol{z}}$. Let $M \coloneqq |d^*_{\boldsymbol{z}} - d|$, and $m \in \mathbb{N}$ be the smallest integer such that $mK^{-1}(\eta/L) \geq M$. This implies that $d_{m-1} < d^*_{\boldsymbol{z}}$. Then, by (A.18), after one more iteration, we have $d_m \geq d^*_{\boldsymbol{z}}$, and $d^*_{\boldsymbol{z}} \in R^m_\epsilon(\{d\})$, which completes the proof for the case where $d^*_{\boldsymbol{z}} > d$. For $d^*_{\boldsymbol{z}} < d$, similar arguments will follow. □

Consider the fixed context scenario where the contexts are $\boldsymbol{z} \in \mathcal{Z}$ for all rounds $n \in \{1, \ldots, N\}$. We define the bad event in round n as

$$\mathcal{G}_n = \{T \notin [\bar{l}_n(\boldsymbol{z}, d), \bar{u}_n(\boldsymbol{z}, d)],\ \forall d \in S_n(\boldsymbol{z})\}\ . \tag{A.19}$$

When the event $\mathcal{G}_n$ occurs in round n, we incur the following instantaneous worst-case regret,

$$\mid f(\boldsymbol{z}, d_n) - T \mid \leq \overline{T}\ , \tag{A.20}$$

since $f \in [0, \overline{T}]$. When the event $\neg\mathcal{G}_n$ occurs, the instantenous regret can be bounded similar to before when the event $\mathcal{E}$ holds, that is,

$$\begin{aligned} \mid f(\boldsymbol{z}, d_n) - T \mid &\leq \mid f(\boldsymbol{z}, d_n) - \mu_{n-1}(\boldsymbol{z}, d_n) \mid + \mid \mu_{n-1}(\boldsymbol{z}, d_n) - T \mid \\ &\leq 2\beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}, d_n)\ , \end{aligned} \tag{A.21}$$

The next lemma shows that the number of rounds in which $\mathcal{G}_n$ can occur in a single context scenario is bounded (independent of time horizon N) given that the optimal dose $d^*_{\boldsymbol{z}}$ is ϵ-reachable from $S_0(\boldsymbol{z})$ after a finite number of rounds.

Lemma A.3. *Assume that the event $\mathcal{E}$ in Lemma 3.1 holds. Then, if $d^*_{\boldsymbol{z}} \in R^k_\epsilon(S_0(\boldsymbol{z}))$ for some $k \in \mathbb{N}$, the event $\mathcal{G}_n$, $n \geq 1$, can only occur finitely many times.*

Proof. If we can show that after finite number of rounds that the bad event $\mathcal{G}_n$ occurs, we have $d^*_z \in S_n(\boldsymbol{z})$, we are done since it is guaranteed that there exists at least one dose in $S_n(\boldsymbol{z})$ (d^*_z) whose confidence interval contains the target value. Assume to the contrary that the bad event $\mathcal{G}_n$ occurs infinitely many times. We first show that there exists $n_1 \in \mathbb{N}$ such that $R_\epsilon(S_0(\boldsymbol{z})) =: R_1 \subseteq S_{n_1}(\boldsymbol{z})$. Assume to the contrary. When $\mathcal{G}_n$ occurs, TACO chooses a dose from the finite set

$\bar{S}_n(\boldsymbol{z}) \coloneqq S_n(\boldsymbol{z}) \cap \bar{\mathcal{D}}_{\boldsymbol{z}} \subseteq R_1 \cap \bar{\mathcal{D}}_{\boldsymbol{z}}$ in round n. In that scenario, let $\bar{S}(\boldsymbol{z})$ denote the biggest finite set such that $\bar{S}_n(\boldsymbol{z}) \subseteq \bar{S}(\boldsymbol{z}) \subseteq R_1 \cap \bar{\mathcal{D}}_{\boldsymbol{z}}$ for every $n \in \mathbb{N}$, and let $\bar{S}_{m_1}(\boldsymbol{z}) = \bar{S}(\boldsymbol{z})$. TACO chooses the dose with the maximum confidence interval in $\bar{S}_n(\boldsymbol{z})$ when $\mathcal{G}_n$ occurs. Then, this implies that the posterior variances of all doses in $\bar{S}_{m_1}(\boldsymbol{z})$ will go to zero by Lemma A.1. For any $d \in S_n(\boldsymbol{z})$, the lower and upper confidence bounds are calculated as follows,

$$\begin{aligned}\bar{l}_n(\boldsymbol{z}, d) &= \max\{l_n(\boldsymbol{z}, d), l_n(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d')\} \\ \bar{u}_n(\boldsymbol{z}, d) &= \min\{u_n(\boldsymbol{z}, d), u_n(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d')\} \ ,\end{aligned}$$

where $d' = \operatorname{argmin}_{d^* \in \bar{\mathcal{D}}_{\boldsymbol{z}}} q_{\boldsymbol{z}}(d, d^*)$. Then, we have,

$$\begin{aligned}w_n(\boldsymbol{z}, d) &\leq u_n(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') - \big(l_n(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d')\big) \\ &= 2\beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}, d') + 2Lq_{\boldsymbol{z}}(d, d') \ , \end{aligned} \tag{A.22}$$

where $\sigma_{n-1}(\boldsymbol{z}, d')$ goes to zero for every $d' \in \bar{S}_{m_1}(\boldsymbol{z})$. Also, by definition of $\bar{\mathcal{D}}_{\boldsymbol{z}}$, we have $2Lq_{\boldsymbol{z}}(d, d') < 2L\frac{\lambda}{2L} = \lambda$. Finally, by choosing $\lambda < \epsilon$ and by (A.22), we can conclude that there exists $n_1 - 1 > m_1$ such that $w_n(\boldsymbol{z}, d) < \lambda < \epsilon$ for every $d \in S_n(\boldsymbol{z})$ for $n \geq n_1 - 1$. That is, we have, $f(\boldsymbol{z}, d) - \epsilon < \bar{l}_{n_1-1}(\boldsymbol{z}, d)$ and $f(\boldsymbol{z}, d) + \epsilon > \bar{u}_{n_1-1}(\boldsymbol{z}, d)$, for every $d \in S_{n_1-1}(\boldsymbol{z})$, where $S_0(\boldsymbol{z}) \subseteq S_{n_1-1}(\boldsymbol{z})$.

Now, let us restate the safe set expansion rule in (3.5) and the reachability operator in (3.6),

$$S_n(\boldsymbol{z}_n) = S_{n-1}(\boldsymbol{z}_n) \cup \left(\bigcup_{d \in S_{n-1}(\boldsymbol{z}_n)} \{d' \in \mathcal{D} \mid \hat{l}_n(\boldsymbol{z}_n, d, d') \geq T_{\min} \wedge \hat{u}_n(\boldsymbol{z}_n, d, d') \leq T_{\max}\} \right) ,$$

$$\begin{aligned}R_\epsilon(S_0(\boldsymbol{z})) = S_0(\boldsymbol{z}) \cup \{d \in \mathcal{D} \mid \exists d' \in S_0(\boldsymbol{z}), f(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d') - \epsilon \geq T_{\min} \\ \wedge f(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') + \epsilon \leq T_{\max}\} \ .\end{aligned}$$

For every dose $d \in R_1 \cap S_0(\boldsymbol{z})$, we have $d \in S_{n_1}(\boldsymbol{z})$ since $S_0(\boldsymbol{z}) \subseteq S_{n_1}(\boldsymbol{z})$. Now, for every dose $d \in R_1 \setminus S_0(\boldsymbol{z})$, there exists $d' \in S_0(\boldsymbol{z})$ such that $f(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d') - \epsilon \geq T_{\min}$, and $f(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') + \epsilon \leq T_{\max}$. Then, since $\bar{l}_{n_1-1}(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d') \geq f(\boldsymbol{z}, d') - Lq_{\boldsymbol{z}}(d, d') - \epsilon \geq T_{\min}$, and $\bar{u}_{n_1-1}(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') \leq$

$f(\boldsymbol{z}, d') + Lq_{\boldsymbol{z}}(d, d') + \epsilon \leq T_{\max}$, we can conclude that for every $d \in R_1$, we have $d \in S_{n_1}(\boldsymbol{z})$, that is, $R_1 \subseteq S_{n_1}(\boldsymbol{z})$ for some $n_1 \in \mathbb{N}$, which is a contradiction. After that point, we can define $R_2 = R_\epsilon(R_1)$, where $R_1 \subseteq S_{n_1}(\boldsymbol{z})$, and show that that there exists $n_2 \geq n_1$ such that $R_2 \subseteq S_{n_2}(\boldsymbol{z})$ and so on. Finally, we have $N_z = n_k$ such that $R_k \subseteq S_{N_z}$, where $N_z \in \mathbb{N}$, meaning that $d_z^* \in S_n(\boldsymbol{z})$ for every $n \geq N_z$, which completes the proof (Note that $R_k \coloneqq R_\epsilon^k(S_0(\boldsymbol{z}))$). □

A.4 Proof of Theorem 3.3

For an event $\mathcal{G}$, let $\mathbb{I}\{\mathcal{G}\} = 1$ if $\mathcal{G}$ holds and 0 otherwise. Assume that the event $\mathcal{E}$ in Lemma 3.1 holds,

$$
\begin{aligned}
R_N &= \sum_{n=1}^{N} \mid f(\boldsymbol{z}, d_n) - T \mid \\
&= \sum_{n=1}^{N} \mathbb{I}\{\mathcal{G}\} \times \mid f(\boldsymbol{z}, d_n) - T \mid + \sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times \mid f(\boldsymbol{z}, d_n) - T \mid \\
&\leq \sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times \mid f(\boldsymbol{z}, d_n) - T \mid + \overline{T} N_z \qquad \text{(A.23)} \\
&\leq \sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times \big(\mid f(\boldsymbol{z}, d_n) - \mu_{n-1}(\boldsymbol{z}, d_n) \mid + \mid (\mu_{n-1}(\boldsymbol{z}, d_n) - T) \mid \big) + \overline{T} N_z \\
& \qquad \text{(A.24)} \\
&\leq \sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times 2\beta_n^{1/2} \sigma_{n-1}(\boldsymbol{z}, d_n) + \overline{T} N_z \ , \qquad \text{(A.25)}
\end{aligned}
$$

where (A.23) follows from (A.20) and Lemma A.3, (A.24) follows from triangle inequality on $\mathbb{R}$, and (A.25) follows from (A.21). The rest of the proof is similar to Theorem 3.2.

Let $\bar{R}_N \coloneqq \sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times 2\beta_n^{1/2}\sigma_{n-1}(\boldsymbol{z}, d_n)$. Then, by Cauchy-Schwartz ineqaulity,

$$\begin{aligned}
\bar{R}_N^2 &\leq N\sum_{n=1}^{N} \mathbb{I}\{\neg\mathcal{G}\} \times 4\beta_n\sigma_{n-1}^2(\boldsymbol{z}, d_n) \\
&\leq 4N\beta_N\sigma^2 \sum_{n=1}^{N} \sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}, d_n) \\
&\leq 8N\beta_N\sigma^2 \frac{1}{2}\sum_{n=1}^{N} \frac{\sigma^{-2}\log(1+\sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}, d_n))}{\log(1+\sigma^{-2})} \\
&\leq CN\beta_N \max\frac{1}{2}\sum_{n=1}^{N} \log(1+\sigma^{-2}\sigma_{n-1}^2(\boldsymbol{z}, d_n)) \\
&= CN\beta_N\gamma_N^{vol2} \ , \qquad \text{(A.26)}
\end{aligned}$$

where (A.26) follows from combining Lemma 3.2 and the definition of γ_N^{vol2} in Section 4. Taking square roots of both sides, we have $\bar{R}_N \leq \sqrt{CN\beta_N\gamma_N^{vol2}}$. Finally, combining (A.25) with the definition of $\bar{R}_N$, we have $R_N \leq \sqrt{CN\beta_N\gamma_N^{vol2}} + \overline{T}N_z$.

Appendix B

Combinatorial Gaussian Process Bandits with Probabilistically Triggered Arms

B.1 Preliminaries

Let us remember the definitions of the events,

$$
\begin{aligned}
H_i(t) &\coloneqq \{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\} \\
H(t) &\coloneqq \{\exists i \in [m] : H_i(t)\} \\
\mathcal{G} &\coloneqq \{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| \leq \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \\
&\qquad \forall i \in [m], \forall t \in [T]\} \\
\mathcal{J} &\coloneqq \Big\{ \sum_{i=1}^{m} \sum_{t=1}^{T} \mathbb{I}\{H_i(t)\} < \alpha \Big\},
\end{aligned}
$$

where $\alpha \in [mT]$ is fixed, $M_i(t) \coloneqq \sum_{\tau=1}^{t-1} \mathbb{I}\{i \in \tilde{S}(\tau)\}$ denotes the number of times base arm i was in the triggering set, and $N_i(t) \coloneqq \sum_{\tau=1}^{t-1} \mathbb{I}\{i \in S'(\tau)\}$ denotes the number of observations made at base arm i, up to round t. Also, let τ_w^i denote the round number where ith base arm is in the triggering set of the selected super

arm for the wth time, and $\tau_0^i = 0$.

Also, let us remind the definition of the exploration coefficient, $\beta_n = 2\log(\frac{mT\pi_n}{\delta})$, where $\sum_{n=0}^{\infty} \frac{1}{\pi_n} = 1$ (e.g., $\pi_n = \pi^2 n^2/6$). Note that β_n is strictly increasing in n.

Fact B.1. *(Multiplicative Chernoff bound [107]) Let $X_1, \ldots, X_n$ be Bernoulli random variables taking values in $\{0,1\}$ such that $\mathbb{E}[X_t|X_1, \ldots, X_{t-1}] \geq \mu$ for all $t \leq n$, and $Y = X_1 + \cdots + X_n$. Then, for all $\delta \in (0,1)$,*

$$\Pr\left\{Y \leq (1-\delta)\mu n\right\} \leq e^{-\frac{\delta^2 \mu n}{2}} .$$

B.2 Proof of Lemma 5.1

The main goal of this lemma is to provide an analogy between the standard information gain, and the pseudo-information gain term that appears in our regret analysis. For the proof of this lemma, we consider a scenario the where base arms are selected with a greedy approach, that is, after a base arm is chosen, the statistics of the learner's policy are updated before choosing the next base arm in the same round. As aforementioned, this lemma serves to provide an analogy for our pseudo-information gain term. Even though this assumption may not hold in general, there are several cases it holds. For instance, in the disjunctive form of cascading bandit problem, where a search engine recommends webpages to its users, we can recommend webpages to a user one at a time and update the GP statistics before our next recommendation. The proof this lemma is inspired by the proof of Lemma 5.3 in [10]. Let $\boldsymbol{x}_t^k$ denote the kth arm in round t for which the learner observes a feedback, and y_t^k denote the observation at $\boldsymbol{x}_t^k$. Furthermore, let $\boldsymbol{y}_t^k$ be the vector of observations up to (not including) kth observation in round t, and n_t denote the total number of observations in round t. Also, let N denote the total number of observations by the end of round T. Conditioned on $\boldsymbol{y}_t^k$, $\boldsymbol{x}_1^1, \boldsymbol{x}_1^2, \ldots, \boldsymbol{x}_t^k$ are deterministic, and the conditional variance $\hat{\sigma}^2_{N(t,k)}(\boldsymbol{x}_t^k)$ does not

depend on $\boldsymbol{y}_t^k$. Then, we have,

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = H(\boldsymbol{y}_T) - \frac{1}{2}\log|2\pi e\sigma^2 \boldsymbol{I}| \tag{B.1}$$
$$= H(\boldsymbol{y}_T) - \frac{N}{2}\log(2\pi e\sigma^2)$$

where $\boldsymbol{y}_T$ denotes the vector of observations by the end of round T. We can rewrite $H(\boldsymbol{y}_T)$ as,

$$\begin{aligned}
H(\boldsymbol{y}_T) &= H(\boldsymbol{y}_T^{n_T}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= H(\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T-1}|\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= H(\boldsymbol{y}_1^2) + H(y_1^2|\boldsymbol{y}_1^2) + \ldots + H(y_T^{n_T-1}|\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)} \log(2\pi e\sigma^2(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)) \\
&= \frac{N}{2}\log(2\pi e\sigma^2) + \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)} \log(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i))
\end{aligned} \tag{B.2}$$

Finally, by the end of round T, given a set of base-arm observations $\boldsymbol{y}_T$, the information gain expression in (5.5) follows from combining (B.1) and (B.2).

B.3 Proof of Lemma 5.2

Fact B.2. *Given $r \sim \mathcal{N}(0,1)$, $\mathbb{P}\{|r| > c\} \leq e^{-c^2/2}$, by tail inequality.*

At round t, we have $\mu_i \mid \mathcal{F}_t \sim \mathcal{N}(\hat{\mu}_{N(t)}(\boldsymbol{x}_i), \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i))$. By using Fact B.2, we can write,

$$\begin{aligned}
\mathbb{P}\big\{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| > \sqrt{\beta_n}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \mid \mathcal{F}_t\big\} &= \mathbb{P}\big\{\frac{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i|}{\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)} > \sqrt{\beta_n} \mid \mathcal{F}_t\big\} \\
&\leq e^{-\beta_n/2}
\end{aligned}$$

for a fixed $i \in [m]$ and n. Calculating the union bound over all m base arms, time horizon T, and number of arms that can be chosen up to round t, $N(t)$, we

observe

$$\begin{aligned}
&\bigcup_{i,t,n} \mathbb{P}\big\{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| > \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \cap N(t) = n \mid \mathcal{F}_t\big\} \\
&\qquad \leq \sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{m(t-1)} e^{-\beta_n/2} P\{N(t) = n \mid \mathcal{F}_t\} \\
&\qquad \leq \sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{\infty} \frac{\delta}{mT\pi_n} \qquad \text{(B.3)} \\
&\qquad \leq \frac{\delta}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{\infty} \frac{1}{\pi_n} \\
&\qquad \leq \delta \sum_{n=0}^{\infty} \frac{1}{\pi_n} \\
&\qquad \leq \delta \qquad \text{(B.4)}
\end{aligned}$$

where (B.3) follows from the definition of β_n, and (B.4) follows from fact that $\sum_{n=0}^{\infty} 1/\pi_n = 1$. Finally, by definition of the event $\mathcal{G}$, and (B.4), we can observe that $\mathbb{P}\big\{\mathcal{G}\big\} \geq 1 - \delta$.

B.4 Proof of Lemma 5.3

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\} \middle| \boldsymbol{\mu}\right] \\
&\quad \leq \mathbb{E}\left[\sum_{w=0}^{T}\sum_{t=\tau_w^i+1}^{\tau_{w+1}^i} \mathbb{I}\{i \in \tilde{S}(t),\ N_i(t) \leq (1-\rho)p_i M_i(t)\} \middle| \boldsymbol{\mu}\right] \\
&\quad \leq \mathbb{E}\left[\sum_{w=0}^{T} \mathbb{I}\{N_i(\tau_{w+1}^i) \leq (1-\rho)p_i M_i(\tau_{w+1}^i)\} \middle| \boldsymbol{\mu}\right] \\
&\quad \leq 1 + \sum_{w=1}^{T} \Pr\{N_i(\tau_{w+1}^i) \leq (1-\rho)p_i M_i(\tau_{w+1}^i) | \boldsymbol{\mu}\} \\
&\quad \leq 1 + \sum_{w=1}^{T} e^{-\frac{\rho^2 p^* w}{2}} \qquad \text{(B.5)} \\
&\quad \leq 1 + \frac{2}{\rho^2 p^*}
\end{aligned}$$

where (B.5) is due to Fact B.1. We can then write,

$$\mathbb{E}\Big[\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\mid \boldsymbol{\mu}\Big] \leq 1+\frac{2}{\rho^2 p^*}\ .$$

Then, we have

$$\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\mid \boldsymbol{\mu}\Big] \leq m+\frac{2m}{\rho^2 p^*}\ . \tag{B.6}$$

We can now use (B.6) and tower rule to observe

$$\mathbb{E}_{\boldsymbol{\mu}}\big[\mathbb{E}[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\mid \boldsymbol{\mu}]\big] \leq m+\frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$$

which yields,

$$\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big] \leq m+\frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]\ . \tag{B.7}$$

Since $\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}$ is non-negative, we can use Markov's inequality for a fixed $\alpha\in[mT]$ and write

$$\mathbb{P}\big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\geq\alpha\big\} \leq \frac{\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big]}{\alpha}\ . \tag{B.8}$$

Finally, we combine the definition of event $\mathcal{J}$, (B.7), and (B.8) to observe

$$\begin{aligned}
\mathbb{P}\big\{\mathcal{J}\big\} &= \Pr\big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}<\alpha\big\}\\
&\geq 1-\frac{\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big]}{\alpha}\\
&\geq 1-\frac{m+\frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{\alpha}\\
&\geq 1-\frac{m}{\alpha}-\frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]\ .
\end{aligned}$$

B.5 Proof of Lemma 5.4

Given $A\subseteq B$, $H(\mu|A)\geq H(\mu|B)$, since conditioning on more observations will reduce entropy. Let $\boldsymbol{Y}_t^i$ denote the observations made at base arm i up to round

t. Remember that $\boldsymbol{Y}_t$ denotes all the observations up to round t. Then, we have $\boldsymbol{Y}_t^i \subseteq \boldsymbol{Y}_t$, which implies $H(\mu|\boldsymbol{Y}_t^i) \geq H(\mu|\boldsymbol{Y}_t)$. The entropy of a Gaussian random variable is $H(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2}\log(2\pi e\sigma^2)$. We have $H(\mu|\boldsymbol{Y}_t) = \frac{1}{2}\log(2\pi e\hat{\sigma}^2_{N(t)})$, and $H(\mu|\boldsymbol{Y}_t^i) = \frac{1}{2}\log(2\pi e(N_i(t)/\sigma^2 + \sigma_i^{-2})^{-1}))$. Then we can write,

$$\frac{1}{2}\log(\frac{2\pi e}{N_i(t)/\sigma^2 + \sigma_i^{-2}}) \geq \frac{1}{2}\log(2\pi e\hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)) \tag{B.9}$$

where $\sigma_i^2 = k(\boldsymbol{x}_i, \boldsymbol{x}_i)$ is the prior variance of GP at $\boldsymbol{x}_i$. Since $k(\boldsymbol{x}_i, \boldsymbol{x}_i) \geq 0$, the following is immediate from (B.9),

$$\frac{\sigma^2}{N_i(t)} \geq \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i) \ .$$