Adaptive Online Decision Making Based on Interconnected Data

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

MSc. Behzad Nourani Koliji

aus Bookan / Iran

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.11.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Philipp Hennig
2. Berichterstatter/-in:	Prof. Dr. Georg Martius

To my parents.

Abstract

Online decision-making problems, in which an agent must select the optimal action from multiple alternatives at each step, are frequently encountered in various real-world scenarios. Some of the main applications of online decision making frameworks are in healthcare, finance, dynamic pricing, recommender systems, anomaly detection, and telecommunication [27]. Multi-armed bandits (MAB) [126] provides rich mathematical formulations for modelling online decision making problems. In MAB settings, the only feedback provided to the learning algorithm (agent) is a possibly noisy reward signal of the chosen decision. This property of MAB severely restricts the agent and slows down the learning process as the size of the action space becomes (exponentially) large. However, in reality, data is generally naturally structured (interconnected). *Hence, it is critical to be able to learn such structures on the fly, and also to learn from the properties these structures create in the data with the goal to accelerate the learning and improve the performance of online decision making algorithms.* This is the key idea that forms the foundation of the research in this thesis.

In order to model structures and interrelations of the data, graphs have been used extensively within MAB problems. Consequently, researchers have managed to introduce effective frameworks for exploiting the structures to accelerate the learning of MAB agents and cope with the dimensions of MAB problems in big environments. However, first of all, there are still some real-world problems that require novel structured MAB frameworks to be solved. Second, state-of-the-art structured MAB frameworks mostly ignore the underlying structure in choosing their strategies in piecewise-stationary environments. Moreover, the current literature of structured MAB frameworks ignore the natural behaviour of structured environments in spreading the negative effects of adversarial corruptions within social networks and consequently fail to perform in a robust manner. In this regard, with the ultimate goal of addressing these issues, we engage in the study of structured MAB settings.

In the first project, we develop a novel combinatorial semi-bandit framework with causally related rewards, where we model the causal relations by a directed graph in a Structural Equation Model (SEM). We deploy our framework to demonstrate a novel application of the MAB framework: analyzing the spread of Covid-19 within a country and identifying the optimal regions for intervention to stop the epidemic. In the second project, we study a novel structured MAB in a piecewise-stationary environment such

that the distribution of arms' instantaneous rewards as well as the relationships between the arms' rewards are subject to changes across time. Within the same project, we study the benefits of adapting the piecewise-stationary MAB strategies according to the underlying structure of the data. For the third project, along the direction of multi-task bandit settings where there is a graph structure linking the bandit tasks, we introduce a novel framework that is more data efficient, in some large-scale real-world scenarios, in comparison to the state-of-the-art. In the final project, we study the online influence maximization problem in social networks in the presence of some corrupted nodes whose damaging effects diffuse throughout the network structure and we introduce an algorithm that is robust against the diffusion of malicious effects of corruptions within the network. In this document, we substantiate the research with in-depth literature reviews and analyses, the development of various novel algorithms, rigorous theoretical justifications, and supporting experimental results.

Kurzfassung

Online-Entscheidungsprobleme, bei denen ein Agent in jedem Schritt die optimale Aktion aus mehreren Alternativen auswählen muss, sind in verschiedenen realen Szenarien häufig anzutreffen. Einige der wichtigsten Anwendungen von Online-Entscheidungsframeworks sind im Gesundheitswesen, im Finanzwesen, bei der dynamischen Preisgestaltung, bei Empfehlungssystemen, bei der Erkennung von Anomalien und in der Telekommunikation zu finden. Mehrarmige Bandits (MABs) bieten umfangreiche mathematische Formulierungen / Formeln für die Modellierung von Online-Entscheidungsproblemen. In MAB-Settings ist die einzige Rückmeldung, die der Lernalgorithmus (Agent) erhält, ein möglicherweise verrauschtes Belohnungssignal der gewählten Entscheidung. Diese Eigenschaft von MABs schränkt den Agenten stark ein und verlangsamt den Lernprozess, da der Aktionsraum (exponentiell) groß wird. In der Realität sind die Daten jedoch im Allgemeinen natürlich strukturiert. Daher ist es von entscheidender Bedeutung in der Lage zu sein, solche Strukturen spontan zu erlernen – und auch von den Eigenschaften zu lernen, die diese Strukturen in den Daten erzeugen. Beides zielt darauf ab, das Bedauern von Online-Entscheidungsalgorithmen zu verbessern. Dies ist der Kerngedanke, der die Grundlage für die Forschung in dieser Doktorarbeit bildet.

Um Strukturen und Zusammenhänge der Daten zu modellieren, wurden bei MAB-Problemen in großem Umfang Graphen verwendet. Folglich ist es Forschern gelungen, wirksame Frameworks zur Nutzung der Strukturen einzuführen. Diese haben das Ziel, das Lernen von MAB-Agenten zu beschleunigen und die Dimensionen von MAB-Problemen in großen Umgebungen zu bewältigen. Allerdings gibt es erstens immer noch einige reale Probleme, deren Lösung neuartige strukturierte MAB-Frameworks erfordert. Zweitens ignorieren strukturierte MAB-Frameworks auf dem Stand der Technik meist die zugrunde liegende Struktur bei der Wahl ihrer Strategien in stückweise-stationären Umgebungen. Darüber hinaus ignoriert die derzeitige Literatur zu strukturierten MAB-Frameworks das natürliche Verhalten strukturierter Umgebungen bei der Verbreitung negativer Auswirkungen gegnerischer Korruptionen / Beschädigungen innerhalb sozialer Netzwerke und ist daher nicht robust genug. Mit dem ultimativen Ziel, diese Probleme zu lösen, befassen wir uns mit der Untersuchung strukturierter MAB-Settings.

Zunächst entwickeln wir ein neuartiges kombinatorisches Semi-Bandit-Framework mit kausal verbundenen Belohnungen. Bei diesem modellieren wir die kausalen Beziehungen durch einen gerichteten Graphen / Digraphen in einem Strukturgleichungsmodell (SGM / Structural Equation Model, SEM). Wir setzen unser Framework ein, um eine neuartige Anwendung des MAB-Frameworks zu demonstrieren: die Analyse der Ausbreitung von Covid-19 innerhalb eines Landes und die Identifizierung der optimalen Regionen für eine Intervention zur Eindämmung der Epidemie. Zweitens untersuchen wir einen neuartig strukturierten MAB in einer stückweise-stationären Umgebung. In dieser unterliegen die Verteilung der unmittelbaren Belohnungen der Arme sowie die Beziehungen zwischen den Belohnungen der Arme zeitlichen Veränderungen. An gleicher Stelle untersuchen wir die Vorteile einer Anpassung der stückweise-stationären MAB-Strategien an die zugrunde liegende Struktur der Daten. Drittens stellen wir im Hinblick auf Multi-Task-Bandit-Settings, bei denen die Bandit-Aufgaben durch eine Graphenstruktur miteinander verbunden sind, ein neuartiges Framework vor. Dieses ist im Vergleich zum Stand der Technik in einigen groß angelegten realen Szenarien dateneffizienter. Schließlich untersuchen wir das Problem der Online-Einflussmaximierung in sozialen Netzwerken bei Vorhandensein einiger korrumpierter / beschädigter Knoten, deren schädliche Auswirkungen sich in der gesamten Netzwerkstruktur verbreiten. Und wir stellen einen Algorithmus vor, der gegen die Verbreitung bösartiger Auswirkungen von Korruptionen innerhalb des Netzwerks robust ist. In diesem Dokument untermauern wir die Forschung mit ausführlichem Literaturüberblick und eingehenden Literaturanalysen, der Entwicklung verschiedener neuartiger Algorithmen, gründlichen theoretischen Begründungen und unterstützenden experimentellen Ergebnissen.

Acknowledgments

I would like to express my heartfelt gratitude to my family, especially my brother Azad, for motivating and inspiring me to pursue a career in Artificial Intelligence. I am profoundly thankful to my beloved partner Julia Lehmann for her unwavering support throughout these years.

I am grateful to professor Setareh Maghsudi, professor Claire Vernade, and professor Philipp Hennig for their management, guidance, and care, which were instrumental in supporting me to complete my PhD.

I am very thankful to Dr. Andrea Schaub, professor Thilo Stehle, and Kirsten Sonnenschein for their support and for providing the funding that enabled me to successfully complete my PhD. Furthermore, I would also like to thank my colleagues and friends, especially Dr. Sofien Dhouib, Steven Bilaj, Mariam Yahya, Amir Rezaei Balef, Ioannis Tsetis, and Xiaotong Cheng whose brilliant minds and encouraging presence have helped me grow both as a researcher and as a person. In addition, I want to thank my teammates, Michela, Nicolas, Onno, Ziyad, and Cagatay for the great time that we had together in Lifelong Reinforcement Learning Lab.

Behzad Nourani Koliji, Tübingen

Contents

Chapter 1

Introduction

This chapter serves as both an introduction and a summary outline of this dissertation.The thesis aims to explore characteristics of sequential decision-making problems in interconnected environments in order to develop innovative frameworks that address real-world challenges. The research falls under the broader umbrella of online decision-making and online learning. In the following, we briefly introduce online decision making and elaborate more on the term "structured bandits" and review state-of-the-art research in handling structural properties within online decision making settings. We highlight some of the gaps and issues within the current state-of-the-art literature and discuss our solutions in a brief and high level fashion as we leave the more detailed discussions for the following chapters of the thesis.

1.1 Online decision making and multi-armed bandits

Online learning frameworks are a subset of machine learning (ML) techniques that are aiming at learning from a data that arrives sequentially. These methods are capable of updating the model incrementally upon the arrival of the new data. This makes them powerful candidates for real-time learning and adaptations. Provided with the feedback to the previous learning tasks, the online learning framework tries to minimize a loss function. In the last couple of decades, online decision making has become one of the most active research fields within online learning. Online decision-making concentrates on applying a range of machine learning, optimization, and statistical methods to ensure decisions are made both promptly and efficiently. The goal is to create systems that are providing optimal outcomes in real-world, real-time scenarios. One could identify *adaptive sampling* and *partial feedback* as two of the main characteristics of most of online decision making settings. Adaptive sampling implies that the learning algorithm is actively involved in the collection of data from the environment in an interactive manner. Moreover, partial feedback indicates that the learning algorithm only receives feedback related to the taken action. Receiving only the partial feedback, the learning agent re-

quires to gather information about the environment (exploration) while minimizing its loss based on the previously collected information (exploitation). Consequently, online decision making with partial feedback naturally inherits the trade-off between exploration and exploitation at any point in time.

Multi-armed bandits (MAB) [126] frameworks provide a rich mathematical formulation for modelling this exploration-exploitation trade-off. Hence, MABs have been used extensively in order to address this trade-off in a variety of applications of online decision making. In the classical stochastic K-armed bandit problem, a decision maker chooses one out of the K possible arms (actions) and experiences an instantaneous reward, which is chosen from an unknown distribution associated with that arm. The goal is to learn from the experience the arm with the best expected reward, *i.e.* the arm with the highest reward on average. A variety of strategies exist as solutions to the MAB problem, such as ε-greedy [140], UCB [8], and Thompson sampling [145]. The performance of these learning strategies is measured in terms of regret, which is the difference between the reward incurred by the algorithm and the optimal reward. The minimax regret scales as $\mathcal{O}(\sqrt{KT})$ where T is the time horizon. Achieving strong performance requires balancing the exploration of less understood actions with the exploitation of existing knowledge to maximize rewards. MAB frameworks have found applications in many fields, including finance [133, 72], dynamic pricing [111], recommender systems [174], and adaptive routing [10].

1.2 Structured bandits

Why are structures important in MAB? In many MAB problems, a common assumption is that the collected data are realizations of independent and identically distributed (i.i.d.) random variables, and the reward obtained from selecting an action does not provide any information about the other available actions. However, this assumption is often violated in numerous realistic scenarios. When this assumption does not hold, it becomes theoretically and practically challenging for a learning agent to effectively solve the bandit problem, especially in large-scale, interconnected environments. This challenge is particularly evident in settings with a vast number of potential actions, such as movie recommendation systems, where the set of all possible movies forms the action space. Another challenging setting is when a decision maker must select a subset of items in each round of decision-making, while adhering to combinatorial constraints, a scenario known as combinatorial bandits [86].

Aside from the above-mentioned challenges that are posed by the mathematical formulation of these types of bandit settings, from an application perspective, in some real-world applications, particularly in areas like healthcare, finance, and robotics, obtaining interaction data can be both expensive and sometimes unattainable. The objective, therefore, is to leverage any available knowledge about the relationships of data entities to mitigate the impact of the large number of actions on the regret of the decision mak-

ing algorithm. Mathematically speaking, we aim to replace the dependence on the total number of actions, K, in the minimax regret of the MAB problem with a smaller quantity that reflects the structure of the data. This would not only improve the algorithm's performance but also provide more clear intuition and interpretation regarding how the information structure of the data influences the bandit problem.

Graphs as mathematical objects to model structures in machine learning. Graphs have been widely used in various applications to model and analyze interconnected data entities in real-world scenarios. Examples include drug discovery [165], gene regulatory network analysis [130], social network analysis [13], fraud detection [5], traffic prediction [99], and computer vision [112]. In these domains, large-scale, complex graphs or networks are common, and learning algorithms are often designed to exploit any knowledge of these structures.

Structured bandits. As the result of the aforementioned, there has been significant interest in addressing bandit problems where complex information structures exist, allowing the learning agent to utilize this knowledge to become more sample-efficient. Hence, one can roughly define structured bandits as the following; Structured bandits is the problem of online decision-making under uncertainty, where the presence of structural information plays a crucial role in guiding and optimizing the decision-making process [155].

Despite notable successes in this area, challenges remain that hinder the widespread adoption of MAB algorithms in real-life applications. This thesis aims to develop frameworks that leverage the understanding of the underlying structure of the data to inspire the creation of novel algorithms. These algorithms are intended to enhance the performance of Multi-Armed Bandit (MAB) algorithms beyond the current state-of-the-art. The following sub-sections will introduce key work in bandit problems that involve information structures closely related to our studies in this dissertation. For more detailed discussions on other types of structured bandits, readers may refer to [152].

1.3 Thesis contribution to structured bandits literature

In the following, we briefly review some of the main research directions and the related references within the more general setting of structured bandits and highlight the fundamental missing points within the state-of-the-art that we want to address in this thesis. Consequently, we mention our contributions in addressing the issues in structured bandits literature.

1. Causally structured bandits. Motivated by applications in epidemiology, marketing strategies, computational biology, and analysis of gene regulatory networks, researchers have tried to marry multi-armed bandits settings and causality [122]. The goal would be to identify the best intervention on a causal system in order to optimize the expected value of a certain signal (reward signal in this setting) in the causal system [17, 90, 131]. Causal graphs are used for representing causal relationships among interacting variables in the bandit setting. They explicitly account for the causal relationships between actions and outcomes, which makes the decision-making more informative in comparison with traditional bandits. In these papers, directed graphs that encode the pattern of interaction among components in the causal system are considered in the problem modelling. However, these works rely on the prior knowledge of the directed graphs between the actions. This property makes their algorithms impractical in realistic scenarios where the underlying graph structure is not available a priori.

Lack of knowledge w.r.t. the underlying graph structure can be managed by employing online learning strategies that both infer system dynamics and guide decision-making in real-time. The central concept is to discern the underlying graph processes and leverage this understanding to enhance action optimization. To achieve this objective, we propose an algorithm that determines the causal relations by learning the network's topology and simultaneously exploits this knowledge to optimize the decision-making process. Compared to previous works, our proposed framework does not require any prior knowledge of the structural dependencies. In addition, the framework presents a completely novel form of combinatorial semi-bandit setting. We establish a sublinear regret bound for the proposed algorithm. We present a novel application of MAB algorithms by applying our framework to analyze the development of Covid-19 in a country. We show that our proposed policy is able to detect the regions that contribute the most to the spread of Covid-19 in the country. We elaborate more on this project and present it in complete details in Chapter 3.

2. Piecewise-stationary structured bandits. Unlike stationary stochastic settings, real-world scenarios often feature evolving reward distributions. For example, in recommender systems, user feedback changes over time. In some cases the changes are such that the environment can be modelled as piecewise stationary [23]. Generally speaking, there are two main approaches in modeling this type of nonstationarity in bandit problems; the *switching case* (abruptly changing) [173] and the *dynamic case* (smoothly changing) [44] [150]. In the switching case, reward distributions of arms stay constant for certain intervals but change instantly for a subset of arms when the environment varies. The point where distributions change is a *change-point* (or breakpoint) and an interval between any two consecutive change-point is a *stationary segment* [173]. Conversely, in the dynamic case, base arms' mean rewards evolve gradually, constrained by a variation budget, rather than changing abruptly [22]. Let us focus on the switching case, also referred to as *piecewise* stationary bandit model [23].

In structured bandit problems within piecewise stationary settings, reward generation process can change due to changes in reward distributions of arms or the structural relationships between variables. Although not addressed in the MAB literature, this model applies to real-world scenarios like financial markets, where both investors' stock purchasing behavior and causal effects among stock prices change over time. Optimal investors monitor and adapt to both types of variations.

Generally, there are two primary approaches to tracking the piecewise stationary behavior of arm distributions; *passively adaptive* approach [58] and *actively adaptive* approach [23]. Methods in the first category do not recognize change-points and base their decisions on the most recent observations. Conversely, methods in the second category employ a change detection algorithm to track distribution changes and decide accordingly. Clearly, the effectiveness of actively adaptive approaches significantly depends on the agent's ability to handle breakpoints efficiently. Actively adaptive algorithms use either *global* or *local restart* to relearn the expected reward of arms. Global restart resets learnings for all arms when any change is detected, while local restart only resets learning of the detected. The choice of the restarting strategy introduces a trade-off on the control of the *false alarm* and the *delay* of the change point detectors and the overall influence of these two parameters on the regret of piecewise stationary bandit algorithms. While global restart is vulnerable to the negative effects of false alarms, local restart suffers from a bigger effects of delayed detections. Basically, the above restarting approaches have a major drawback: they overlook potential relationships among arms' distributions when deciding on restarting them, i.e. the relationships between the arms are ignored in the design of the algorithms that rely on change point detectors.

Therefore, introducing a new restarting strategy for bandit algorithms in structured piecewise stationary environments is essential for two motivational reasons. First, social networks, a famous application area of bandit algorithms, show high modularity [114] [26] in their network structures. Second, changes within networks are often interconnected, spreading locally through contagion [53], social influence [53], or diffusion [144], as seen in marketing campaigns and rumor propagation [132]. In this regard, we theoretically study the advantages of having restarting strategies that consider the underlying structure of the data. Hence, we introduce the notion of *group restart* as a new alternative restarting strategy in the decision making process in structured environments and we theoretically prove its advantage over other restarting strategies in controlling the delay and the false alarm of change point detectors. Furthermore, we study the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In our non-stationary environment, variations in the base arms' distributions, causal relationships between rewards, or both, change the reward generation process. Our UCB-based algorithm integrates a mechanism to trace the variations of the underlying graph structure, which captures the causal relationships between the rewards in the bandit setting. Theoretically, we establish a regret upper bound that reflects the effects of the number of structural- and distribution changes on the performance. We will expand on this project and provide a comprehensive explanation and presentation of results in Chapter 4.

3. Multi-task contextual bandits [38, 169]. In this setting, the learner is given a graph encoding relations between the bandit tasks. Each bandit task represents a user who needs to be provided with services by the online recommender system. Depending on the problem setup, sampling one action can, explicitly or implicitly, provide information to the decision-maker about other actions.

One solution to this problem relies on the notion of *homophily* in social networks [109, 53]. Accordingly, given the large number of users in social networks, user preferences can potentially be identified more rapidly by utilizing the similarities between them, as indicated by the social network's structure. In the context of modeling social networks, users' preference relationships are depicted using an undirected graph, where neighboring nodes correspond to users with similar preferences. Utilizing this information and integrating it into bandit algorithms can greatly enhance performance [169]. For instance, the paper of [169] achieves a single-user cumulative regret of $\mathcal{O}(\psi d\sqrt{T})$ in stochastic linear bandit problem [94], where T is the time horizon, d is the dimensionality of the feature vectors and $\psi \in (0,1)$ is a scalar parameter that depends on the graph structure between users in the social network. This is a clear improvement over the famous LinUCB algorithm [95] that considers the users independent of one another and achieves the cumulative regret of $\mathcal{O}(d\sqrt{T})$. Indeed, the knowledge of user relations allows the algorithm to tackle the data sparsity issue that is inherent to bandit settings. Both papers of Cesa-Bianchi *et al.* [38] and Yang *et al.* [169] propose UCB-style algorithms and exhibit the importance of using the social network graph to achieve lower regrets using Laplacian regularization. Consequently, both methods promote smoothness among the preference vectors of users in order to transfer the collected information between them.

A second solution to the multi-task problem is to consider the high *modularity* measures exhibited by social networks [114, 26]. Accordingly, users can be grouped based on the graph topology, allowing for a preference vector to be learned for each cluster. This approach significantly reduces the dimensionality of the problem. Sequentially clustering bandit tasks was first introduced in [59] and later improved in [97]. In other papers, the algorithm presented in [115] utilizes K-means clustering to group users, whereas the algorithm in [46] relies on hedonic games for the online clustering of bandits, and [168] make use of community detection techniques on graphs to find user clusters.

For the first solution, the Laplacian regularization in [38, 169] does not account for the smoothness heterogeneity introduced by a piecewise constant behavior over the social network graph [162]. For the second approach, the frameworks in [59, 97] aim at gradually forming clusters as connected components. However, their approach can cause overconfidence in the constructed clusters, potentially leading to error accumulation.

In our attempt to address these issues, we assume access to a graph that encodes the relationships between bandit tasks, where the task parameter vectors are piecewise constant across the graph, indicating that tasks naturally form clusters. We propose an algorithm that utilizes this prior knowledge of the piecewise constant structure to update tasks without the need to explicitly identify the clusters. This approach effectively ad-

dresses the limitations discussed earlier: it integrates the piecewise constant smoothness directly into our regularizer, and by avoiding explicit cluster estimation, our algorithm sidesteps the risk of overconfidence. We provide a regret upper bound for our setting. Our bound highlights the advantage of our algorithm in high dimensional settings, and for large graphs. In Chapter 5, we will delve into this project in greater details, offering a thorough and complete presentation of results involved.

4. Online influence maximization and diffusion of corruption in social network. Online social networks facilitate the frequent collection and updating of information regarding users' connections and communications, making Influence Maximization (IM) [79] easier to implement. Typically, in an IM setting, a social network is represented as a graph, with users depicted as nodes. The edges symbolize relationships between users, while the edge weights indicate the influence probabilities between them. Influence then spreads through the network according to a chosen diffusion model [79]. In practical scenarios, even when the network's topology is known, the influence probabilities remain unknown in advance. This challenge emphasizes the importance of the online influence maximization (OIM) problem [157, 164, 166, 92]. In OIM, the task of estimating activation probabilities falls to the learner, who must determine these probabilities through direct interactions with the network (exploration) while trying to maximize influence based on the gained knowledge (exploitation), hence the application of MAB mathematical frameworks in solving OIM problems.

Researchers have explored the OIM problem from various perspectives [164, 157, 166, 98, 176]; however, most approaches assume that all users in the social network are fully cooperative, influencing others willingly and automatically. This assumption overlooks the critical issue of potentially having corrupted users or nodes. In real-world applications, malicious users can deceive the system with disruptive behaviors, spreading corruption by disrupting the information flow within the network even if they are not selected as seeds. Therefore, it is crucial to develop an algorithm that is robust against corruption in online influence maximization settings.

We introduce a novel algorithm for corruption-robust OIM, which is based on an OIM algorithm combined with a corruption-robust linear bandit approach [67]. By incorporating weighted regression into the OIM algorithm, our method effectively mitigates the issues caused by inaccurate estimations due to corrupted users. We provide a theoretical analysis showing that the proposed algorithm not only achieves sublinear regret but also maintains robustness against malicious behaviors. In Chapter 6, we will thoroughly elaborate on this project, presenting it in full detail and covering the results comprehensively.

1.4 List of publications

This thesis is based on the following papers. The detailed contributions of the author of the thesis to each paper are provided in the next subsection.

I **Behzad Nourani-Koliji**, Saeed Ghoorchian, and Setareh Maghsudi, "Linear Combinatorial Semi-Bandit with Causally Related Rewards", published in proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (2022), (IJCAI-22).

II **Behzad Nourani-Koliji**, Steven Bilaj, Amir Rezaei Balef, and Setareh Maghsudi. "Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards", published in proceedings of the Twenty-Sixth European Conference on Artificial Intelligence, ECAI 2023, pages 1787–1794. IOS Press, 2023.407

III Sofien Dhouib, Steven Bilaj, **Behzad Nourani-Koliji**, and Setareh Maghsudi. "Clusters Agnostic Network Lasso Bandits". in [Openreview.net/submitted to ICML2025]

IV Xiaotong Cheng, **Behzad Nourani-Koliji**, Setareh Maghsudi, "Online Influence Maximization with Semi-Bandit Feedback under Corruptions", published in IEEE Transactions on Network Science and Engineering.

1.5 Author's contributions

In this section, I clarify my contributions to each paper in this dissertation.

Paper I:

Linear Combinatorial Semi-Bandit with Causally Related Rewards

I proposed and developed the entire mathematical problem formulation and algorithm. I also designed and performed all the synthetic data and the real-data experiments. The theoretical analysis was done by me and Saeed Ghoorchian. I wrote the article entirely and Saeed Ghoorchian helped with rephrasing and editing some parts.

Paper II:

Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards

The idea of extending the framework in Paper I to the nonstationary environment was introduced by me. I noticed the scientific contributions of this idea after my literature reviews. I proposed the problem formulation and the algorithm entirely. I designed and performed all the synthetic and real-data experiments. Based on my studies, I highlighted the place where the theoretical contribution should take place. The method for performing the theoretical analysis was proposed by Amir Rezaei Balef and discussed with me and Steven Bilaj. The theoretical analysis was finalized and written by Steven Bilaj and Amir Rezaei Balef. I wrote the entire paper and edited the theoretical analysis.

Paper III:

Clusters Agnostic Network Lasso Bandits

After discussions between me and Sofien Dhouib, Sofien Dhouib proposed the initial idea. Sofien Dhouib proposed the problem formulation and the algorithm entirely. I solved the optimization part of the proposed algorithm and implemented the solution. I performed some literature review and helped Sofien Dhouib with the design, implementation, and execution of the experiments. Sofien Dhouib and Steven Bilaj did the theoretical analysis. The paper is mostly written by Sofien Dhouib. Steven Bilaj helped with writing the theoretical part and I helped with writing the introduction and related works.

Paper IV:

Online Influence Maximization with Semi-Bandit Feedback under Corruptions

Xiaotong Cheng proposed an initial idea. The final idea is the result of our discussions with Xiaotong Cheng. Xiaotong Cheng proposed the problem formulation and the algorithm entirely. I performed the real-data experiment and some of the synthetic data experiments. I also wrote the introduction and the related work after my literature review. Xiaotong Cheng performed the analysis and wrote the rest of the paper and I helped with editing the text and proof-readings.

Chapter 2

Technical Background

In this chapter, we briefly explain the fundamental mathematical background of methods and models in this thesis. For more detailed explanations on variations of MAB, we refer the reader to [91]. The references for other topics are mentioned in the related sections.

2.1 Multi-armed bandits (MAB)

2.1.1 Stochastic MAB

In the general form of multi-armed bandit problem, the MAB framework [88, 126] consists of a set of arms resembling different decisions and actions in an environment. An action's reward is the utility collected by the learning algorithm (also named player or agent) upon taking an action. The agent only gets to observe the feedback related to the taken action. This partial feedback is also referred to as bandit feedback. The interaction between the agent and the environment happens over $T \in \mathbb{N}$ rounds of decision making.

Depending on the assumptions about the reward generating process, the MAB framework can be classified into the stochastic [8] or adversarial setting [9]. In this thesis, we exclusively focus on stochastic multi-armed bandits. In the stochastic setting, each arm is linked to a specific reward distribution, and every time an arm is pulled, a sample is drawn from its corresponding distribution. The mean of this distribution corresponds to the expected reward of the arm, but it is not known to the agent.

To assess the effectiveness of the MAB learning algorithm, we measure its performance against the optimal strategy, which selects the best possible action at every time step. This optimal approach reflects the decisions of an Oracle, who has full knowledge of the reward distributions for all available actions. This performance metric is typically referred to as cumulative (pseudo) regret. We use R_T to represent this measure and it is defined as

$$R_T = T \max_{i \in [K]} \mu_i - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{a_t}\right],$$

where a_t is the action taken by the agent at round t and μ_i is the mean of the reward distribution for arm i. The expectation is calculated based on the randomness in the actions chosen by the learning algorithm. Our primary focus is on how the regret scales with the time horizon T. The agent's objective is to minimize the gap between its own performance and that of the optimal strategy, as measured by cumulative regret.

From an algorithmic perspective, to solve the stochastic bandit problems, different strategies such as those based on Upper Confidence Bounds (UCBs) [91, 8] and Thompson Sampling [146, 3, 127] as well as greedy approaches [89] are proposed in the literature.

2.1.2 Stochastic linear bandits

In this section, we introduce a specific class of bandit problems within the stochastic framework, where the rewards are assumed to follow a linear structure. This assumption allows the learning process to benefit from insights gained from one action to be used for learning about other actions, enhancing overall learning efficiency. This modeling approach provides a rich and powerful framework, which will serve as a key focus for some of the research projects throughout this thesis.

In linear bandits [7, 94, 4, 158], every arm is associated with a d-dimensional vector $\mathbf{x}$ (a point in $\mathbb{R}^d$) and the reward function is an unknown linear function as $\mathbf{x}^T\boldsymbol{\theta}$ such that the aim is to learn the unknown d-dimensional vector $\boldsymbol{\theta}$. Two of the most influential algorithms in this setting are LINUCB and OFUL that were introduced in [94] and [1] respectively. Both algorithms achieve $\mathcal{O}(d\sqrt{T})$ regret.

2.1.3 Multi-task contextual linear bandits

In this section, we present the basic setting of Multi-task Contextual Linear Bandit (MLB) problem. One of our contributions is dedicated to solve this problem. In MLB [38, 169], the agent is required to solve multiple contextual linear bandit tasks. In practical applications, this setting mimics the task faced by a recommender system, which recommends items to multiple users. Each user represents a bandit instance and all users share the same set of candidate items (arm set). The bandit parameter vectors represent users' preference and items are arms characterised by arm features (contexts). The reward corresponds to user's preference to the recommended item.

The environment consists of a set of bandit problems $\mathcal{V}$. Over a time horizon T, the agent is required to solve bandit problems sequentially. At a round $t \in \mathbb{N}$, a user $m(t) \in \mathcal{V}$ is selected uniformly at random and served an arm with context vector $\mathbf{x}(t)$ from a finite action set $\mathcal{A}(t) \subset \mathbb{R}^d$ with size K, depending on their estimated preference vector $\hat{\boldsymbol{\theta}}_{m(t)}(t) \in \mathbb{R}^d$. We assume the expected reward to be linear, with an additive, σ-sub-Gaussian noise. Formally, the received reward $y(t)$ is given by $y(t) = \langle \boldsymbol{\theta}_{m(t)}, \mathbf{x}(t) \rangle + \eta(t)$ where $\eta(t)$ is the noise. The performance of our policy is assessed by the expected regret

over the T interaction rounds for all tasks:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_{m(t)}, \mathbf{x}^{\star}(t) - \mathbf{x}(t) \rangle\right],$$

where $\mathbf{x}^{\star}(t)$ is the optimal arm with respect to the bandit instance $m(t) \in \mathcal{V}$.

Consider a scenario where there are N contextual linear bandit instances to solve. As a result, each bandit is encountered approximately $T' = T/N$ times. If the agent treats each bandit separately, the cumulative regret for each instance can be bounded by $R(T') = \mathcal{O}(d\sqrt{T'})$. Consequently, the total regret across all bandits would be $R(T) = \mathcal{O}(Nd\sqrt{T'})$, which grows undesirably linearly with the number of bandit instances N. One of our key contributions, presented in Chapter 5, is the development of the algorithm *Network Lasso Policy*, which leverages prior knowledge of task similarities to reduce cumulative regret in multi-task contextual linear bandits.

2.1.4 Combinatorial semi-bandits

A constraint on the number of arms played by the learner in each round in the multi-armed bandit problem can present an issue in some real-world scenarios. Combinatorial multi-armed bandits [85, 37] is the framework that is designed to address this type of problems. Unlike the basic multi-armed bandit problem, in combinatorial multi-armed bandit, the actions can consist of a subset of the set of all available *base arms*. This subset is referred to as *super arm*. At the end of each round, the agent receives the reward corresponding to the selected subset of base arms. The primary challenge in combinatorial bandits lies in managing the exponentially large action space. To address this issue, current methods assume that the reward function has a structure over the set of arms and utilize an efficient oracle to solve an optimization problem over the combinatorial set of feasible actions. In this thesis, we consider semi-bandit feedback where the agent observes a reward for each selected arm. The combinatorial bandit problem is well-investigated in the literature by considering various settings [42, 43, 44]. We elaborate more on the mathematical formulation of combinatorial semi-bandit problems on Chapters 3, and 4.

2.1.5 Piecewise-stationary bandits

In many real-world applications, the environment is dynamic, with key statistical features of the random variables evolving as time goes on. This contrasts with stationary environments, where a single arm remains optimal for the duration of the process. In non-stationary settings, however, the optimal action can change as the environment changes. It is possible to address the nonstationary behaviors of rapidly-varying environments using the adversarial bandit framework [9]. However, in some cases, the environment changes slowly and less frequently. One of the main approaches in modeling this type

of nonstationarity in bandit problems is referred to as piecewise-stationary bandit model [23].

A piece-wise stationary bandit model is characterized by a set of K arms. We denote by $\mu_i(t)$ the mean reward of arm i at round t. At each round t, a decision maker has to select an arm $I_t \in \{1,\dots,K\}$. At time t, we denote by i_t^* an arm with maximal expected reward, i.e., $\mu_{i_t^*}(t) = \max_i \mu_i(t)$, called an optimal arm. We measure the performance of policy π using the notion of *piecewise stationary regret*, i.e., the regret w.r.t. an oracle that knows the best action in each stationary segment:

$$R_T^{\pi} = \mathbb{E}\left[\sum_{t=1}^{T}\left(\mu_{i_t^*}(t) - \mu_{I_t}(t)\right)\right].$$

We refer to the specific time instance where a change occurs in the environment as a *change point* or a *break point*.

A common approach to solve the problem is to use a sliding window or a discount factor to estimate the expected value of rewards with piecewise stationary generating processes [44]. Other approaches, such as those based on change-point detection [173], have also been proposed to solve the problem. In Chapter 5, we consider piecewise-stationary structured bandits and propose adaptive decision-making strategies to solve the formulated problem.

2.1.6 Online influence maximization

In social networks, influence propagates through the network structure under a specific diffusion model [79]. As a result, companies aim to select a fixed number of customers, called seeds or source nodes, that greatly influence others to receive reimbursement in return for advertising their products. That is referred to as Influence Maximization (IM) [79]. In influence maximization frameworks, companies aim at maximizing the influence spread given a limited budget. Even if the network topology is accessible, the influence probabilities are unknown a priori. That highlights the importance of online influence maximization (OIM) problem [156, 157]. Multi-armed bandit framework, especially the linear bandit model [95], is widely used to solve the online influence maximization problem [156, 164, 166].

In the influence maximization problem, a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is utilized to model the social network. $\mathcal{V} = \{1, 2, \dots, n\}$ is the set of users (nodes) and $\mathcal{E}$ is the set of edges with cardinality $m = |\mathcal{E}|$. Each edge $e \in \mathcal{E}$ is associated with an activation probability $p(e) \in [0,1]$. For example, an edge $e = (u,v)$ could represent user v follows user u on some social media and $p(u,v)$ represents the probability that user v (receiving node) will be activated/influenced by user u (giving node). Denote $\mathbf{P} = (p(e_1), \dots, p(e_m))$ to be the activation probability vector. For a given seed set $\mathcal{S} \subseteq \mathcal{V}$ with activation probability $\mathbf{P}$, the expected number of influenced users under the diffusion model D is $f_{D,\mathbf{P}}(\mathcal{S})$. By definition, the users/nodes in $\mathcal{S}$ are always influenced. Given $\mathcal{G}$ and a budget K on the

number of seeds to be selected, IM aims to find the seed set which will maximize the influence spread

$$\mathcal{S}^{\text{opt}} = \arg\max_{|\mathcal{S}|\leq K} f_{D,\mathbf{P}}(\mathcal{S}).$$

Relying on the mathematical modelling in linear bandits, we assume each edge $e \in \mathcal{E}$ is associated with a known feature vector $\mathbf{x}_e \in \mathbb{R}^d$ and an unknown coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the dimension of the feature vector. It is assumed that for all $e \in \mathcal{E}$, $p(e)$ is "well-approximated" by $\mathbf{x}_e^T \boldsymbol{\theta}$. The goal of the bandit agent is to estimate these activation probabilities in the social network as it tries to maximize the spread of influence in the network. We develop more on the mathematical formulation of online influence maximization in Chapter 6 as we present our project in this field.

2.2 Generalized likelihood ratio (GLR) change point detectors

In this section, we want to provide a quick overview and the basic formulations of the generalized likelihood ratio (GLR) change point detectors. Sequential change-point detection is a classical problem in statistical sequential analysis [106]. In our algorithm design in Chapter 4, we will use the GLR change-point detector presented in [23], which works for any sub-Bernoulli distribution.

In a piecewise stationary bandit game, let us assume that we have stored for an arm a time sequence $\{X_t\}_{t=1}^n$. Hypothesis $\mathcal{H}_0$ indicates that the entire sequence is drawn from a sub-Bernoulli distribution for any $t \leq n$, while Hypothesis $\mathcal{H}_1$ says that the sequence is generated from two different sub-Bernoulli distributions with an unknown change-point $s \in [1, n-1]$. Mathematically, we can formulate this as the following:

$$\mathcal{H}_0 : \exists f_0 : X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_0,$$

$$\mathcal{H}_1 : \exists f_0 \neq f_1, s \in [1, n-1] : X_1, \ldots, X_s \overset{i.i.d.}{\sim} f_0 \text{ and } X_{s+1}, \ldots, X_n \overset{i.i.d.}{\sim} f_1.$$

The GLR statistic for sub-Bernoulli distributions is:

$$\text{GLR}(n) = \sup_{s\in[1,n-1]} \left[s \times \text{kl}(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}) + (n-s) \times \text{kl}(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}) \right],$$

where $\hat{\mu}_{s:s'}$ is the mean of the observations collected between s and s', and $\text{kl}(x, y)$ is the binary relative entropy between Bernoulli distributions defined as the following,

$$\text{kl}(x, y) = x \log\left(\frac{x}{y}\right) + (1-x) \log\left(\frac{1-x}{1-y}\right).$$

If the GLR statistic $\text{GLR}(n)$ is large, it indicates that hypothesis $\mathcal{H}_1$ is more likely. Hence,

the sub-Bernoulli GLR change-point detector with threshold function $\beta(n,\delta)$, and confidence level $\delta \in (0,1)$ is

$$\tau := \inf\left\{ n \in \mathbb{N} : \sup_{s\in[1,n-1]} \left[s \times \mathrm{kl}(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}) + (n-s) \times \mathrm{kl}(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}) \right] \geq \beta(n,\delta) \right\},$$

where $\beta(n,\delta)$ can theoretically be chosen according to the lemma 2 in [23]. We can implement the GLR change point detector using the following function,

$$\begin{aligned}\mathbf{GLR}(X_1,\ldots,X_n;\delta) = \mathbb{1}\{\sup_{s\in[1,n-1]}[s \times \mathrm{kl}(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}) + \\ + (n-s) \times \mathrm{kl}(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n})] \geq \beta(n,\delta)\}.\end{aligned}$$

Thus, if $\mathbf{GLR}(X_1,\ldots,X_n;\delta) = 1$, the GLR change point detector sends a signal that indicates that a change in the statistics of the given sequence of numbers is detected.

In comparison to other change-point detection methods commonly used in piecewise-stationary MAB settings, the GLR change-point detector offers a significant advantage by requiring fewer parameters to be tuned and less prior information about the bandit instance. Specifically, the GLR detector only requires adjustment of the threshold parameter. In contrast, cumulative sum (CUSUM) change point detector [101] involves more parameter tuning and rely on prior knowledge of the smallest magnitude of all change-points. In Chapter 4 we delve deeper into the mathematical formulations of the GLR change point detector and discuss the effects of its parameters in details within the context of our project in piecewise stationary structured bandits.

2.3 Graph theory

In this section, we recall some definitions and basics of graph theory. These notions are used throughout the thesis.

Let us denote a graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1,\ldots,K\}$ is the set of all nodes, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set. For the graph $\mathcal{G}$, if $(u,v) \in \mathcal{E}$ implies $(v,u) \in \mathcal{E}$ for all $u,v \in \mathcal{V}$ then the graph is undirected. The graph is directed, if each edge has a direction, pointing from one vertex to another. One way of representing the graph is by using the adjacency matrix of the graph denoted as $\mathbf{A} \in \mathbb{R}^{K\times K}$, such that the (i,j)-th element is

$$[\mathbf{A}]_{ij} = \begin{cases} a_{ij} & \text{if there is an edge from } j \text{ to } i \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where a_{ij} is the weight of the edge that connects the node i and j. The degree matrix of the graph is defined as a $K \times K$ diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1,\ldots,d_k,\ldots,d_K)$ where d_k represents the number of times an edge terminates at vertex k. For an undirected graph, the matrix $\mathbf{A}$ is symmetric, i.e., $a_{ij} = a_{ji}$. This is not necessarily true for directed graphs.

Consider the case where we have a graph $\mathcal{G}$ with E edges, i.e. $|\mathcal{E}| = E$. Hence, an alternative representation of the graph is given by its incidence matrix $\mathbf{B} \in \mathbb{R}^{K \times E}$ such that,

$$[B]_{ij} = \begin{cases} -1 & \text{if the edge } j \text{ leaves node } i; \\ 1 & \text{if the edge } j \text{ enters node } i; \\ 0 & \text{otherwise.} \end{cases} \quad (2.2)$$

Most importantly, we define the Laplacian matrix as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Laplacian matrix is the most commonly used representation of the graph due to its interesting properties [136].

2.4 Graph signal processing

Graph signal processing is a multidisciplinary research area that is focused on the development of tools for analyzing data on irregular graph domains [120]. Roughly speaking, a graph signal can be defined as the set of values residing on the set of nodes of a graph [120]. Considering that the nodes in the graph are connected via the edges, the graph can help to extract lots of information from the data. As in classical signal processing, we can define the notion of smoothness for graph signals [49]. We can also define the notions of spectral representation of graph signals and graph Fourier transform [143]. As a result, it allows for the definitions of frequency, bandlimitedness, filtering, and sampling and reconstruction of graph signals [135, 103, 48, 151]. In some scenarios, if the underlying graph cannot be directly observed, we can also learn the structure from the observed graph signals [50, 129].

We consider a weighted and undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ where $\mathcal{V}$ is the vertex set and $\mathcal{E}$ is the edge set, and $\mathbf{A}$ is the adjacency matrix. The graph Laplacian operator can be defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal degree matrix. By construction, the graph Laplacian operator is a real symmetric and positive semi-definite matrix that admits a complete set of real orthonormal eigenvectors with corresponding non-negative eigenvalues. We denote the eigenvectors of the graph Laplacian by $\mathbf{U} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{K-1}]$, where K is the number of vertices, and the spectrum of the graph by

$$\Lambda(\mathbf{L}) = \{0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{K-1}\}.$$

A graph signal $\mathbf{y}$ in the vertex domain is defined as a real-valued function on the vertices of the graph $\mathcal{G}$. This implies that $\mathbf{y}(v)$ is in fact the value of the function at vertex $v \in \mathcal{V}$. We define the spectral domain representation of the graph signal to extract the critical information about the characteristics of graph signals. In this regard, the eigenvectors of the Laplacian operator can be used to perform harmonic analysis of the graph signal, and the corresponding eigenvalues carry a notion of frequency [135, 151]. The Laplacian eigenvectors form a Fourier basis, so that for any function $\mathbf{y}$ defined on the vertices of

the graph, the graph Fourier transform $\mathbf{s}$ is defined as

$$\mathbf{s} = \mathbf{U}^\top \mathbf{y},$$

while the inverse graph Fourier transform is

$$\mathbf{y} = \mathbf{U}\mathbf{s}.$$

A graph signal is smooth if signal samples at neighbouring nodes are similar [77, 49]. There are multiple methods in modelling smooth graph signals depending on the application [49, 77, 108, 144]. One of the most widely adopted modelling formulations is to use the Laplacian quadratic form [77] $\sum_{i,j} A_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2 = \text{tr}\left(\mathbf{Y}^\top \mathbf{L}\mathbf{Y}\right)$, where $\mathbf{y}_j$ is the nodal measurements at node j. This is to quantify the smoothness of all graph signals $\mathbf{y}^t$ across time that are stored on the columns of $\mathbf{Y}$. Smaller values of $\text{tr}\left(\mathbf{Y}^\top \mathbf{L}\mathbf{Y}\right)$ are indications of smoother graph signals. The above-mentioned techniques and notions and the corresponding references from graph signal processing have been used in the Chapters 3, 4, and 5 in this thesis.

2.5 Structural equation modelling

Structural equation models (SEMs) have been used popularly across various fields. Examples include sociology, psychometrics, and genetics [78, 62, 113, 32]. In particular, linear SEMs have proven effective for modeling causal relationships between variables in network structures [32]. In this thesis, we utilize SEM to model the causal interactions between reward outcomes of arms in multi-armed bandit (MAB) scenarios.

Let us assume to have a directed graph that shows the causal interactions between some variables. We use $\mathcal{G}(\mathcal{V},\mathcal{E})$ to represent the graph where $\mathcal{V}$ is the set of vertices/nodes, with cardinality $|\mathcal{V}| = K$, and $\mathcal{E}$ is the set of edges. We denote the adjacency matrix of the graph with $\mathbf{A} \in \mathbb{R}^{K\times K}$. Let y_{it} be the t-th measurement at node i. Assume to have a graph process over the network such that y_{it} depends on its one-hop neighbors in addition to an exogenous input x_{it}. Therefore, we can write the following,

$$y_{it} = \sum_{j\neq i} a_{ij} y_{jt} + b_{ii} x_{it}, \quad t = 1,\ldots,T$$

where $a_{ij} := [A]_{ij}$, as a non-zero element of the adjacency matrix $a_{ij} \neq 0$ shows that a directed edge from j to i is present. After concatenating nodal measurements into $\mathbf{y}_t := [y_{1t},\ldots,y_{Kt}]^\top$, and $\mathbf{x}_t := [x_{1t},\ldots,x_{Kt}]^\top$ per slot t, the matrix-vector version of the above equation can be written in a compact form as $\mathbf{y}_t = \mathbf{A}\mathbf{y}_t + \mathbf{B}\mathbf{x}_t$, $t = 1,\ldots,T$, where $[A]_{ii} = 0$ and $\mathbf{B} := \text{diag}(b_{11},\ldots,b_{KK})$. Consequently, if we collect the observations over T rounds in a matrix $\mathbf{Y} := [\mathbf{y}_1,\ldots,\mathbf{y}_T]$, we arrive at the noise-free matrix formulation of

SEM

$$\mathbf{Y} = \mathbf{AY} + \mathbf{BX}.$$

The above model indicates the causes-effect relationships in the system. It implies that the causes-effect do not necessarily happen instantaneously as causes $\{y_{jt}, x_{it}\}$ can happen at the beginning of the round and the effect y_{it} can be observed at the end of the round t (after the entire system is stabilized).

Graph Learning in SEMs. Structural equation models (SEMs) are highly regarded for their ability to infer network topology, offering a straightforward approach while effectively capturing the directed dependencies among variables in a system. Assuming to have the nodal measurements $\mathbf{Y}$ across the network as well as the exogenous inputs $\mathbf{X}$, the task of topology inference is to find the unknown adjacency matrix $\mathbf{A}$. Availability of the exogenous variables in linear SEMs plays a significant role in identifiability of the estimated adjacency matrix [20].

Depending on the application scenarios and the prior information w.r.t. the structure of the network in SEMs, we need to have different connectivity structures for the estimated networks, e.g. sparse, dense, clustered, static, dynamic, etc. Hence, a variety of different structure learning frameworks were introduced depending on the prior information regarding the behaviours of the system [14, 15, 20, 61].

The following figure provides an example for a linear SEM.

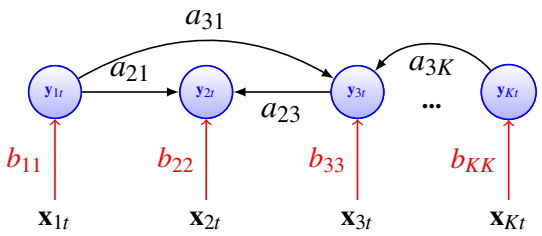


Figure 2.1: An illustration of a typical SEM network consisting of K vertices and their causal relations. The black directed edges represent the causal relationships amongst the vertices.

We are going to use our knowledge of linear SEM in Chapters 3, and 4.

Chapter 3

Linear Combinatorial Semi-Bandit with Causally Related Rewards

In a sequential decision-making problem, having a structural dependency amongst the reward distributions associated with the arms makes it challenging to identify a subset of alternatives that guarantees the optimal collective outcome. Thus, besides individual actions' reward, learning the causal relations is essential to improve the decision-making strategy. To solve the two-fold learning problem described above, we develop the 'combinatorial semi-bandit framework with causally related rewards', where we model the causal relations by a directed graph in a stationary structural equation model. The nodal observation in the graph signal comprises the corresponding base arm's instantaneous reward and an additional term resulting from the causal influences of other base arms' rewards. The objective is to maximize the long-term average payoff, which is a linear function of the base arms' rewards and depends strongly on the network topology. To achieve this objective, we propose a policy that determines the causal relations by learning the network's topology and simultaneously exploits this knowledge to optimize the decision-making process. We establish a sublinear regret bound for the proposed algorithm. Numerical experiments using synthetic and real-world datasets demonstrate the superior performance of our proposed method compared to several benchmarks.

3.1 Introduction

In the seminal form of the Multi-Armed Bandit (MAB) problem, an agent selects an arm from a given set of arms at sequential rounds of decision-making. Upon selecting an arm, the agent receives a reward, which is drawn from the unknown reward distribution of that arm. The agent aims at maximizing the average reward over the gambling horizon [126]. The MAB problem portrays the exploration-exploitation dilemma, where the agent decides between accumulating immediate reward and obtaining information that might result in larger reward only in the future [105]. To measure the performance of

a strategy, one uses the notion of *regret*. It is the difference between the accumulated reward of the applied decision-making policy and that of the optimal policy in hindsight.

In a combinatorial semi-bandit setting [42], at each round, the agent selects a subset of *base arms*. This subset is referred to as a *super arm*. She then observes the individual reward of each base arm that belongs to the selected super arm. Consequently, she accumulates the collective reward associated with the chosen super arm. The combinatorial MAB problem is challenging since the number of super arms is combinatorial in the number of base arms. Thus, conventional MAB algorithms such as [8] are not appropriate for combinatorial problems as they result in suboptimal regret bounds. The aforementioned problem becomes significantly more difficult when there are causal dependencies amongst the reward distributions.

In some cases, it is possible to model the causal structure that affects the rewards [90]. Therefore, exploiting the knowledge of this structure helps to deal with the aforementioned challenges. In our paper, we develop a novel combinatorial semi-bandit framework with causally related rewards, where we rely on Structural Equation Models (SEMs) [78] to model the causal relations. At each time of play, we see the instantaneous rewards of the chosen base arms as controlled stimulus to the causal system. Consequently, in our causal system, the solution to the decision-making problem is the choice over the exogenous input that maximizes the collected reward. We propose a decision-making policy to solve the aforementioned problem and prove that it achieves a sublinear regret bound in time. Our developed framework can be used to model various real-world problems, such as network data analysis of biological networks or financial markets. We apply our framework to analyze the development of Covid-19 in Italy. We show that our proposed policy is able to detect the regions that contribute the most to the spread of Covid-19 in the country.

Compared to previous works, our proposed framework does not require any prior knowledge of the structural dependencies. For example, in [141], the authors exploit the prior knowledge of statistical structures to learn the best combinatorial strategy. At each decision-making round, the agent receives the reward of the selected super arm and some side rewards from the selected base arms' neighbors. In [73] a Combinatorial Thompson Sampling (CTS) algorithm to solve a combinatorial semi-bandit problem with probabilistically triggered arms is proposed. The proposed algorithm has access to an oracle that determines the best decision at each round of play based on the already collected data. Similarly, the authors in [43] study a setting where triggering super arms can probabilistically trigger other unchosen arms. They propose an Upper Confidence Bound (UCB)-based algorithm that uses an oracle to improve the decision-making process. In [170], the authors formulate a combinatorial bandit problem where the agent has access to an influence diagram that represents the probabilistic dependencies in the system. The authors propose a Thompson sampling algorithm and its approximations to solve the formulated problem. Further, there are some works that study the underlying structure of the problem. For example, in [148], the authors attempt to learn the structure of a combinatorial bandit problem. However, they do not assume any causal relations

between rewards. Moreover, in [131], the MAB framework is employed to identify the best soft intervention on a causal system while it is assumed that the causal graph is only partially unknown.

The rest of the paper is organized as follows. In Section 3.2, we formulate the structured combinatorial semi-bandit problem with causally related rewards. In Section 3.3, we introduce our proposed algorithm, namely SEM-UCB. Section 3.4 includes the theoretical analysis of the regret performance of SEM-UCB. Section 3.5 is dedicated to numerical evaluation. Section 3.6 concludes the paper.

3.2 Problem setup

Let $[N] = \{1, 2, \ldots, N\}$ denote the set of *base arms*. $\mathbf{b}_t = [\mathbf{b}_t[1], \mathbf{b}_t[2], \ldots, \mathbf{b}_t[N]] \in [0,1]^N$ represents the vector of *instantaneous rewards* of the base arms at time t. The instantaneous rewards of each base arm $i \in [N]$ are independent and identically distributed (i.i.d.) random variables drawn from an unknown probability distribution with mean $\boldsymbol{\beta}[i]$. We collect the mean rewards of all the base arms in the mean reward vector of $\boldsymbol{\beta} = [\boldsymbol{\beta}[1], \boldsymbol{\beta}[2], \ldots, \boldsymbol{\beta}[N]]$.

We consider a causally structured combinatorial semi-bandit problem where an agent sequentially selects a subset of base arms over time. We refer to this subset as the *super arm*. More precisely, at each time t, the agent selects a *decision vector* $\mathbf{x}_t = [\mathbf{x}_t[1], \mathbf{x}_t[2], \ldots, \mathbf{x}_t[N]] \in \{0,1\}^N$. If the agent selects the base arm i at time t, we have $\mathbf{x}_t[i] = 1$, otherwise $\mathbf{x}_t[i] = 0$. The agent observes the value of $\mathbf{b}_t[i]$ at time t only if $\mathbf{x}_t[i] = 1$. The agent is allowed to select at most s base arms at each time of play. Hence, we define the set of all feasible super arms as

$$\mathcal{X} = \left\{ \mathbf{x} \mid \mathbf{x} \in \{0,1\}^N \wedge \|\mathbf{x}\|_0 \leq s \right\}, \tag{3.1}$$

where $\|\cdot\|_0$ determines the number of non-zero elements in a vector. In our problem, the parameter s is pre-determined and is given to the agent.

We take advantage of a directed graph structure to model the causal relationships in the system. We consider an unknown stationary sparse Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ denotes the set of N vertices, i.e., $|\mathcal{V}| = N$, $\mathcal{E}$ is the edge set, and $\mathbf{A}$ denotes the weighted adjacency matrix. By $p \leq N-1$, we denote the length of the largest path in the graph $\mathcal{G}$. We assume that the reward generating processes in the bandit setting follow an error-free Structural Equation Model (SEM) [61, 20]. The exogenous input vector and the endogenous output vector of the SEM at each time t are denoted by $\mathbf{z}_t = [\mathbf{z}_t[1], \mathbf{z}_t[2], \ldots, \mathbf{z}_t[N]]$ and $\mathbf{y}_t = [\mathbf{y}_t[1], \mathbf{y}_t[2], \ldots, \mathbf{y}_t[N]]$, respectively. At each time t, the exogenous input $\mathbf{z}_t$ represents the semi-bandit feedback in the decision-making problem. Formally,

$$\mathbf{z}_t = \mathrm{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{3.2}$$

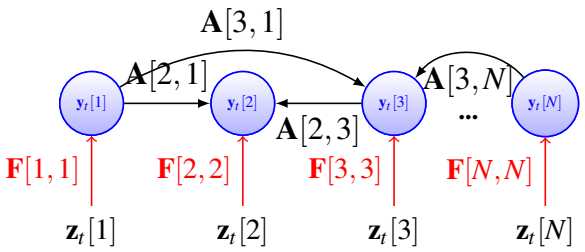


Figure 3.1: An exemplary illustration of a graph consisting of N vertices and their causal relations. The black directed edges represent the causal relationships amongst the vertices.

where $\mathrm{diag}(\cdot)$ represents the diagonalization of its given input vector. Consequently, we define the elements of the endogenous output vector $\mathbf{y}_t$ as

$$\mathbf{y}_t[i] = \sum_{i \neq j} \mathbf{A}[i,j]\mathbf{y}_t[j] + \mathbf{F}[i,i]\mathbf{z}_t[i], \quad \forall i = 1,\ldots,N, \tag{3.3}$$

where $\mathbf{F}$ is a diagonal matrix that captures the effects of the exogenous input vector $\mathbf{z}_t$. The SEM in Equation (3.3) implies that the output measurement $\mathbf{y}_t[i]$ depends on the single-hop neighbor measurements in addition to the exogenous input signal $\mathbf{z}_t[i]$. In our formulation, at each time t, the endogenous output $\mathbf{y}_t[i]$ represents the *overall reward* of the corresponding base arm $i \in [N]$. Therefore, at each time t, the overall reward of each base arm comprises two parts; one part directly results from its instantaneous reward, while the other part reflects the effect of causal influences of other base arms' overall rewards.

In Equation (3.3), the overall rewards are causally related. Thus, the adjacency matrix $\mathbf{A}$ represents the causal relationships between the overall rewards; accordingly, the element $\mathbf{A}[i,j]$ of the adjacency matrix $\mathbf{A}$ denotes the causal impact of the overall reward of base arm j on the overall reward of base arm i, and we have $\mathbf{A}[i,i] = 0$, $\forall i = 1,2,\ldots,N$. We assume that the agent is not aware of the causal relationships between the overall rewards. Hence, the adjacency matrix $\mathbf{A}$ is unknown a priori. In the following, we work with the matrix form of Equation (3.3), defined at time t as

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_t + \mathbf{F}\mathbf{z}_t. \tag{3.4}$$

In Figure 3.1, we illustrate an exemplary network consisting of N vertices and the underlying causal relations. Based on our problem formulation, the agent is able to observe both the exogenous input signal vector $\mathbf{z}_t$ and the endogenous output signal vector $\mathbf{y}_t$. As we see, there does not exist necessarily a causal relation between every pair of nodes.

By inserting Equation (3.2) in Equation (3.4) and solving for $\mathbf{y}_t$ we obtain

$$\mathbf{y}_t = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}\,\mathrm{diag}(\mathbf{b}_t)\mathbf{x}_t. \tag{3.5}$$

Finally, we define the *payoff* received by the agent upon choosing the decision vector $\mathbf{x}_t$ as

$$r(\mathbf{x}_t) = \mathbf{1}^\top \mathbf{y}_t = \mathbf{1}^\top (\mathbf{I} - \mathbf{A})^{-1} \mathbf{F} \text{diag}(\mathbf{b}_t) \mathbf{x}_t, \tag{3.6}$$

where $\mathbf{1}$ is the N-dimensional vector of ones. Since the graph $\mathcal{G}$ is a DAG, it implies that with a proper indexing of the vertices, the adjacency matrix $\mathbf{A}$ is a strictly upper triangular matrix. This guarantees that the matrix $(\mathbf{I} - \mathbf{A})$ is invertible. In our problem, since the agent directly observes the exogenous input, we assume that the effects of $\mathbf{F}$ on the exogenous inputs are already integrated in the instantaneous rewards. Therefore, to simplify the notation and without loss of generality, we assume that $\mathbf{F} = \mathbf{I}$ in the following.

Given a decision vector $\mathbf{x}_t \in \mathcal{X}$, the expected payoff at time t is calculated as

$$\mu(\mathbf{x}_t) = \mathbb{E}\left[r(\mathbf{X}) | \mathbf{X} = \mathbf{x}_t\right], \tag{3.7}$$

where the expectation is taken with respect to the randomness in the reward generating processes.

Ideally, the agent's goal is to maximize her total mean payoff over a time horizon T. Alternatively, the agent aims at minimizing the expected regret, defined as the difference between the expected accumulated payoff of an oracle that follows the optimal policy and that of the agent that follows the applied policy. Formally, the expected regret is defined as

$$\mathcal{R}_T(\mathcal{X}) = T\mu(\mathbf{x}^*) - \sum_{t=1}^{T} \mu(\mathbf{x}_t), \tag{3.8}$$

where $\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{X}}{\text{argmax}}\ \mu(\mathbf{x})$ is the optimal decision vector, and $\mathbf{x}_t$ denotes the selected decision vector at time t under the applied policy.

Remark 1. *The definition of payoff in Equation (3.6) implies that we are dealing with a linear combinatorial semi-bandit problem with causally related rewards. In general, due to the randomness in selection of the decision vector $\mathbf{x}_t$, the consecutive overall reward vectors $\mathbf{y}_t$ become non-identically distributed. In the following section, we propose our algorithm that is able to deal with such variables. This is an improvement over the previous methods, such as [43] and [73], that are not able to cope with our problem formulation, as they are specially designed to work with i.i.d. random variables.*

3.3 Decision-making strategy

In this section, we present our decision-making strategy to solve the problem described in Section 3.2. Our proposed policy consists of two learning components: (i) an online graph learning and (ii) an Upper Confidence Bound (UCB)-based reward learning. In the following, we describe each component separately and propose our algorithm, namely SEM-UCB.

3.3.1 Online graph learning

The payoff defined in Equation (3.6) implies that the knowledge of $\mathbf{A}$ is necessary to select decision vectors that result in higher accumulated payoffs. Therefore, the agent aims at learning the matrix $\mathbf{A}$ to improve her decision-making process. To this end, we propose an online graph learning framework that uses the collected feedback, i.e., the collected exogenous input and endogenous output vectors, to estimate the ground truth matrix $\mathbf{A}$. In the following, we formalize the online graph learning framework.

At each time t, we collect the feedback up to the current time in $\mathbf{Z}_t = [\mathbf{z}_1 \dots \mathbf{z}_t]$ and $\mathbf{Y}_t = [\mathbf{y}_1 \dots \mathbf{y}_t]$. Therefore,

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_t + \mathbf{Z}_t. \tag{3.9}$$

We assume that the right indexing of the vertices is known prior to estimating the ground truth adjacency matrix. We use the collected feedback $\mathbf{Y}_t$ and $\mathbf{Z}_t$ as the input to a parametric graph learning algorithm [61, 50]. More precisely, we use the following optimization problem to estimate the adjacency matrix at time t.

$$\begin{aligned} \hat{\mathbf{A}}_t = \underset{\mathbf{A}}{\operatorname{argmin}} \quad & \|\mathbf{Y}_t - \mathbf{A}\mathbf{Y}_t - \mathbf{Z}_t\|_2^2 + \mathrm{g}(\mathbf{A}) \\ \text{s.t.} \quad & \mathbf{A}[i,j] \geq 0, \quad \forall i,j \in [N], \\ & \mathbf{A}[i,j] = 0, \quad \forall i \geq j, \end{aligned} \tag{3.10}$$

where $\|\cdot\|_2$ represents the L^2-norm of matrices and $\mathrm{g}(\mathbf{A})$ is a regularization function that imposes sparsity over $\mathbf{A}$. In our numerical experiments, we work with different regularization functions to demonstrate the effectiveness of our proposed algorithm in different scenarios. As an example, we impose the sparsity property on the estimated matrix $\hat{\mathbf{A}}_t$ in Equation (3.10) by defining $\mathrm{g}(\mathbf{A}) = \lambda \|\mathbf{A}\|_1$, where $\|\cdot\|_1$ is the L^1-norm of the matrices and λ is the regularization parameter. Our choices of regularization function guarantee that the optimization problem in Equation (3.10) is convex.

3.3.2 SEM-UCB algorithm

We propose our decision-making policy in Algorithm 1. The key idea behind our algorithm is that it works with observations for each base arm, rather than the payoff observations for each super arm. As the same base arm can be observed while selecting different super arms, we can use the obtained information from selection of a super arm to improve our payoff estimation of other relevant super arms. This, combined with the fact that our algorithm simultaneously learns the causal relations, significantly improves the performance of our proposed algorithm and speed up the learning process.

For each base arm i, we define the empirical average of instantaneous rewards at time t as

$$\hat{\boldsymbol{\beta}}_t[i] = \frac{\sum_{\tau=1}^{t} \mathbf{b}_\tau[i] \mathbb{1}\left\{\mathbf{x}_\tau[i] = 1\right\}}{\mathbf{m}_t[i]}, \tag{3.11}$$

where $\mathbf{m}_t[i]$ denotes the number of times that the base arm i is observed up to time t. Formally,

$$\mathbf{m}_t[i] = \sum_{\tau=1}^{t} \mathbb{1}\left\{\mathbf{x}_\tau[i] = 1\right\}. \tag{3.12}$$

The initialization phase of SEM-UCB algorithm follows a specific strategy to create a rich data that helps to learn the ground truth adjacency matrix. At each time t during the first N times of play, SEM-UCB picks the column t of an upper-triangular initialization matrix $\mathbf{M} \in \{0,1\}^{N\times N}$, where $\mathbf{M}$ is created as follows. All diagonal elements of $\mathbf{M}$ are equal to 1. As for the column i, if $i \leq s$, we set all elements above diagonal to 1. If $s+1 \leq i \leq N$, we select $s-1$ elements above diagonal uniformly at random and set them to 1. The remaining elements are set to 0.

After the initialization period, our proposed algorithm takes two steps at each time t to learn the causal relationships and the expected instantaneous rewards of the base arms. First, it uses the collected feedback $\mathbf{Y}_t$ and $\mathbf{Z}_t$ and solves the optimization problem in Equation (3.10) to obtain the estimated adjacency matrix. It then uses the reward observations to calculate the UCB index $\mathbf{E}_t[i]$ for each base arm i, defined as

$$\mathbf{E}_t[i] = \hat{\boldsymbol{\beta}}_t[i] + \sqrt{\frac{(s+1)\ln t}{\mathbf{m}_t[i]}}. \tag{3.13}$$

Afterward, the algorithm selects a decision vector $\mathbf{x}_t$ using the current estimate of the adjacency matrix and the developed UCB indices of the base arms. Let $\mathbf{E}_t$ be the vector for all UCB indices at time t as $\mathbf{E}_t = [\mathbf{E}_t[1], \ldots, \mathbf{E}_t[N]]$. At time t, SEM-UCB selects $\mathbf{x}_t$ as

$$\begin{aligned} \mathbf{x}_t = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}} \quad & \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1} \operatorname{diag}(\mathbf{E}_{t-1})\mathbf{x} \\ \text{s.t.} \quad & \|\mathbf{x}\|_0 \leq s. \end{aligned} \tag{3.14}$$

Remark 2. *The initialization phase of our algorithm guarantees that all the base arms are pulled at least once and the matrix* $\mathbf{M}$ *is full rank. Consequently, the adjacency matrix* $\mathbf{A}$ *is uniquely identifiable from the collected feedback [20].*

Remark 3. *Let* $\mathbf{c}^\top = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1} \operatorname{diag}(\mathbf{E}_{t-1})$. *Since all the elements of both matrices* $\mathbf{E}_{t-1}$ *and* $\hat{\mathbf{A}}_{t-1}$ *are non-negative, we have* $\mathbf{c}[i] > 0$, $\forall i \in [N]$. *Thus, the optimization problem Equation (3.14) reduces to finding the s-biggest elements of* $\mathbf{c}$. *Therefore, Equation (3.14) can be solved efficiently based on the choice of sorting algorithm used to order the elements of* $\mathbf{c}$.

The computational complexity of the SEM-UCB algorithm varies depending on the solver that is used to learn the graph. For example, if we use OSQP solver [138], we achieve a computational complexity of order $\mathcal{O}(N^4)$.

Input: Parameter s, initialization matrix $\mathbf{M}$.

for $t = 1, \ldots, N$ **do**
  Select column t of the initialization matrix $\mathbf{M}$ as the decision vector $\mathbf{x}_t$.
  Observe $\mathbf{z}_t$ and $\mathbf{y}_t$.
end for
for $t = N+1, \ldots, T$ **do**
  Solve Equation (3.10) to obtain $\hat{\mathbf{A}}_{t-1}$.
  Calculate $\mathbf{E}_{t-1}[i]$ using Equation (3.13), $\forall i \in [N]$.
  Select decision vector $\mathbf{x}_t$ that solves Equation (3.14).
  Observe $\mathbf{z}_t$ and $\mathbf{y}_t$.
end for

1: SEM-UCB: Structural Equation Model-Upper Confidence Bound

3.4 Theoretical analysis

In this section, we prove an upper bound on the expected regret of SEM-UCB algorithm. We use the following definitions in our regret analysis. For any decision vector $\mathbf{x} \in \mathcal{X}$, let $\Delta(\mathbf{x}) = \mu(\mathbf{x}^*) - \mu(\mathbf{x})$. We define $\Delta_{\max} = \max_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})$ and $\Delta_{\min} = \min_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})$. Moreover, let $\mathbf{w}_t^\top = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{A}}_t)^{-1} \mathrm{diag}(\mathbf{x}_{t+1})$. We define $w_{\max} = \max_t \max_i \mathbf{w}_t[i]$.

The following theorem states an upper bound on the expected regret of SEM-UCB.

Theorem 1. *The expected regret of SEM-UCB algorithm is upper bounded as*

$$\mathcal{R}_T(\mathcal{X}) \leq \left[\frac{4 w_{\max}^2 s^2 (s+1) N \ln T}{\Delta_{\min}^2} + N + \frac{\pi^2}{3} s^p N \right] \Delta_{\max}. \tag{3.15}$$

Proof. See Section A of supplementary material. □

3.5 Experimental analysis

In this section, we present experimental results to provide more insight on the usefulness of learning the causal relations for improving the decision-making process. We evaluate the performance of our algorithm on synthetic and real-world datasets by comparing it to standard benchmark algorithms.

Benchmarks. We compare SEM-UCB with state-of-the-art combinatorial semi-bandit algorithms that do not learn the causal structure of the problem. Specifically, we compare our algorithm with the following policies: (i) CUCB [43] calculates a UCB index for each base arm at each time t and feeds them to an approximation oracle that outputs a

super arm. (ii) DFL-CSR [141] develops a UCB index for each base arm and selects a super arm at each time t based on a prior knowledge of a graph structure that shows the correlations among base arms. (iii) CTS [73] employs Thompson sampling and uses an oracle to select a super arm at each time t. (iv) FTRL [175] selects a super arm at each time t based on the method of Follow-the-Regularized-Leader. To be comparable, we apply these benchmarks on the vector of overall reward $\mathbf{y}_t$ at each time t. If a benchmark requires $\mathbf{y}_t$ to be in $[0,1]$, we feed the normalized version of $\mathbf{y}_t$ to the corresponding algorithm. Finally, in our experiments, we choose $s = 6$, meaning that the algorithms can choose 6 base arms at each time of play.

3.5.1 Synthetic dataset

Our simulation setting is as follows. We first create a graph consisting of $N = 20$ nodes. The elements of the adjacency matrix $\mathbf{A}$ are drawn from a uniform distribution over $[0.4, 0.7]$. The edge density of the ground truth adjacency matrix is 0.15. At each time t, the vector of instantaneous rewards $\mathbf{b}_t$ is drawn from a multivariate normal distribution with the support in $[0,1]^{20}$ and a spherical covariance matrix. As demonstrated in Section 3.2, we generate the vector of overall rewards according to the SEM in Equation (3.3). We use $\mathrm{g}(\mathbf{A}) = \lambda \|\mathbf{A}\|_1$ as the regularization function in Equation (3.10) when estimating the adjacency matrix $\mathbf{A}$. The regularization parameter λ is tuned by grid search over $[0.0001, 1000]$. We evaluate the estimated adjacency matrix at each time t by using the mean squared error defined as $\mathrm{MSE} = \frac{1}{N^2} \left\| \mathbf{A} - \hat{\mathbf{A}} \right\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

Comparison with the benchmarks. We run the algorithms using the aforementioned synthetic data with $T = 4000$. In Figure 3.2, we depict the trend of time-averaged expected regret for each policy. As we see, SEM-UCB surpasses all other policies. This is due to the fact that SEM-UCB learns the network's topology and hence, it has a better knowledge of the causal relationships in the graph structure, unlike other policies that do not estimate the graph structure. As we see, the time-averaged expected regret of SEM-UCB tends to zero. This matches with our theoretical results in Section 3.4. Note that, the benchmark policies exhibit a suboptimal regret performance as they have to deal with non-identically distributed random variables $\mathbf{y}_t$ [1].

3.5.2 Real-world application

We evaluate our proposed algorithm on the Covid-19 outbreak dataset of daily new infected cases during the pandemic in different regions within Italy.[2] The dataset fits in

[1] More experiments are presented in Section A.2 of the appendix

[2] https://github.com/pcm-dpc/COVID-19

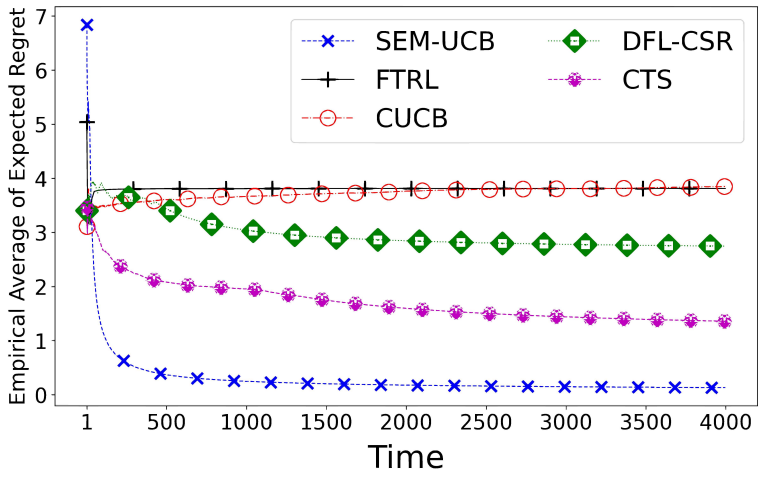


Figure 3.2: Time-averaged expected regret of different policies.

our framework as the daily new cases in each region results from the causal spread of Covid-19 among the regions in a country [107] and the region-specific characteristics [63]. As the regions differ in their regional characteristics, such as socio-economic and geographical characteristics, each region has a specific exposure risk of Covid-19 infection. To be consistent with our terminology in Section 3.2, at each time (day) t, we use the *overall reward* $\mathbf{y}_t[i]$ to refer to the *overall daily new cases* in region i and use the *instantaneous reward* $\mathbf{b}_t[i]$ to refer to the *region-specific daily new cases* in region i. Naturally, the overall daily new cases includes the region-specific daily new cases of Covid-19 infection.

Governments around the world strive to track the spread of Covid-19 and find the regions that are contributing the most to the total number of daily new cases in the country [28]. By the end of this experiment, we address this critical problem and highlight that our algorithm is capable of finding the optimal candidate regions for political interventions in order to contain the spread of a contagious disease such as Covid-19.

Data preparation. We focus on the recorded daily new cases from 10 August to 15 October, 2020, for $N = 21$ regions within Italy. The Covid-19 dataset only provides us with the overall daily new cases of each region. Hence, in order to apply our algorithm, we need to infer the distribution of region-specific daily new cases for each region. In the following, we describe this process and further pre-processing of the Covid-19 dataset.

According to [30], for the time period from 18 May to 3 June, 2020, all places for work and leisure activities were opened and travelling within regions was permitted while travelling between regions was forbidden. Consequently, during this period, there are no causal effects on the overall daily new cases of each region from other regions. In addition, according to google mobility data [118], during 4 weeks prior to 18 May the mobility was increasing within the regions while travel ban between the regions was still

imposed. Hence, we use this expanded period to estimate the underlying distributions of the region-specific daily new cases using a kernel density estimation. Finally, considering that the daily recorded data noticeably fluctuates, a 7-day moving average was applied to the signals.

We create the region-specific daily new cases for each region by sampling from the estimated distributions. Below, we present the results of applying our algorithm on the pre-processed Covid-19 dataset. Since the data only contains the reported overall daily new cases for a limited time period, care should be exercised in interpreting the results. However, by providing more relevant data, our proposed framework helps towards more accurate detection of the regions that contribute the most to the development of Covid-19.

Learning the structural dependencies. Our algorithm learns the ground truth adjacency matrix $\mathbf{A}$ using Equation (3.10). As for the choice of regularization function in Equation (3.10), we employ Directed Total Variation (DTV) which is a novel application of the Graph Directed Variation (GDV) function [128]. DTV regularization function is defined as

$$\mathrm{g}(\mathbf{A}) = \lambda \sum_{i,j=1,\dots,N} \mathbf{A}[i,j] \sum_{k=1,\dots,t} \left[\mathbf{Y}[i,k] - \mathbf{Y}[j,k]\right]^{+}, \tag{3.16}$$

$$[y]^{+} = \max\{y, 0\}. \tag{3.17}$$

The regularization function addresses the smoothness of the entire observations $\mathbf{Y}$ over the underlying directed graph. To be more realistic, since the causal spread of the disease might create cycles, we additionally include cyclic graphs in the search space of the optimization problem in Equation (3.10).

We perform cross-validation technique to tune the regularization parameter λ. As mentioned before, we work on a limited time period with $T = 66$ days. Thus, we split the data into train and validation sets in 10:1 ratio. More specifically, we split the data into 6 subsets of 11 consecutive days. In each subset, one day is chosen uniformly at random to be included in the validation set while the remaining 10 days are added to the train set. We calculate the prediction error at each time t by

$$Error(t) = \frac{1}{NK(t)} \sum_{i \in \mathcal{K}(t)} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1, \tag{3.18}$$

where $\mathcal{K}(t)$ is the validation set at time t with cardinality $K(t) = |\mathcal{K}(t)|$ and $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the validation data and the corresponding predicted value using the estimated graph for day i, respectively. Figure 3.3 compares the ground truth overall daily new cases and the predicted overall daily new cases using the estimated graph on 4 different days of the Covid-19 outbreak in our validation data. Due to space limitation, we use abbreviations for region names. Table A.1 of the appendix lists the abbreviations together with the original names of the regions. We observe that our proposed framework is capable to estimate the data for each region efficiently, that helps the agent to improve its decision-

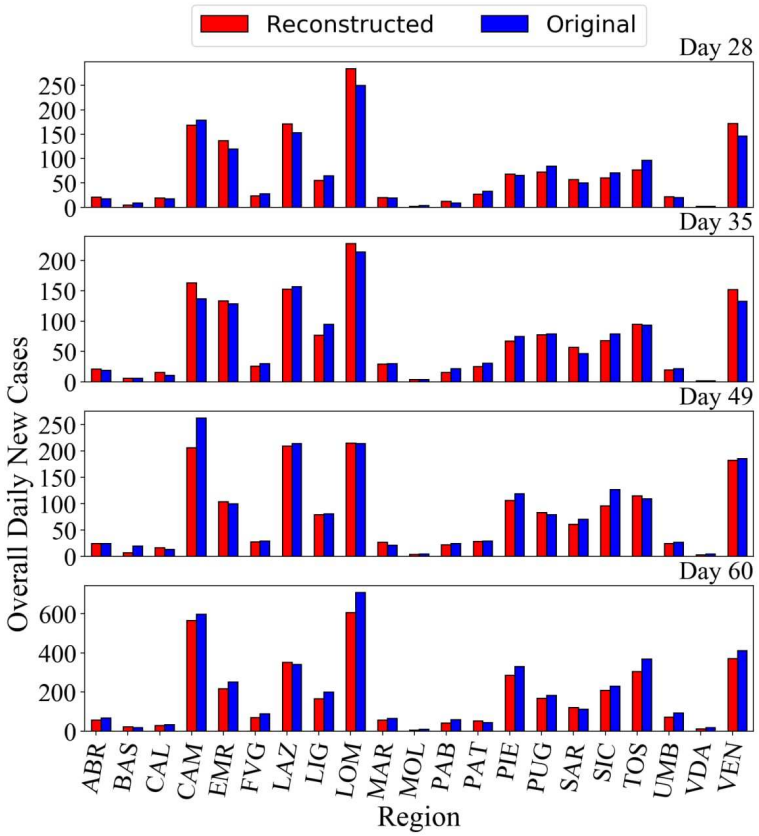


Figure 3.3: Original overall daily new cases and the corresponding predicted values for different days in the validation set.

making process in a real-world scenario.

Learning regions with the highest contribution. In Figure 3.4, we show the decision-making process of the agent over time by following the SEM-UCB policy. Dark rectangles represent the 6 selected regions at each day (time). Based on our framework, we represent the selected regions by our algorithm as those with biggest contributions to the development of Covid-19 during the time interval considered in our experiment. More specifically, we find the regions of Lombardia, Emilia-Romagna, Lazio, Veneto, Piemonte, and Liguria as the ones that contribute the most to the spread of Covid-19 during that period in Italy.

We emphasize that, due to the causal effects among the regions, contribution of each region to the spread of Covid-19 differs from its overall daily cases of infection. Thus, the set of regions with the highest contribution does not necessarily equal to the set of regions with the highest total number of daily cases. This is a key aspect of our problem formulation that is addressed by SEM-UCB in Figure 3.4. We elaborate more on this

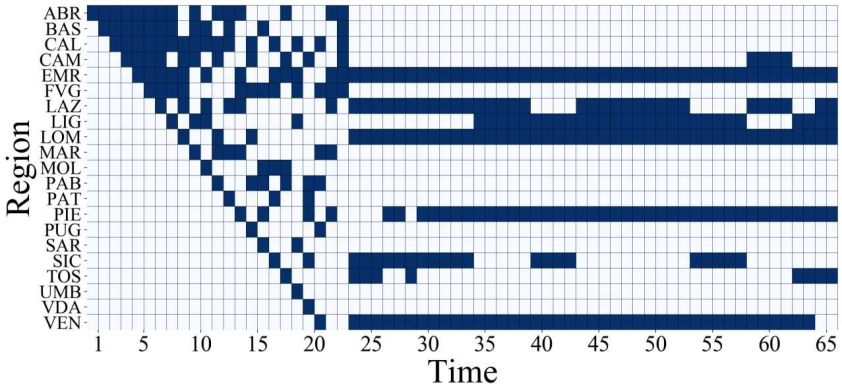


Figure 3.4: Selected regions on each day.

fact in Section A.2 of the appendix.

3.6 Conclusion

In this paper, we developed a combinatorial semi-bandit framework with causally related rewards, where we modelled the causal relations by a directed graph in a structural equation model. We developed a decision-making policy, namely SEM-UCB, that learns the structural dependencies to improve the decision-making process. We proved that SEM-UCB achieves a sublinear regret bound in time. Our framework is applicable in a number of contexts such as network data analysis of biological networks or financial markets. We applied our method to analyze the development of Covid-19. The experiments showed that SEM-UCB outperforms several state-of-the-art combinatorial semi-bandit algorithms. Future research directions would be to extend the current framework to deal with piece-wise stationary environments where the causal graph and/or the expected instantaneous rewards of the base arms undergo abrupt changes over time.

Chapter 4

Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards

We study the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In our nonstationary environment, variations in the base arms' distributions, causal relationships between rewards, or both, change the reward generation process. In such an environment, an optimal decision-maker must follow both sources of change and adapt accordingly. The problem becomes aggravated in the combinatorial semi-bandit setting, where the decision-maker only observes the outcome of the selected bundle of arms. The core of our proposed policy is the Upper Confidence Bound (UCB) algorithm. We assume the agent relies on an adaptive approach to overcome the challenge. More specifically, it employs a change-point detector based on the Generalized Likelihood Ratio (GLR) test. Besides, we introduce the notion of *group restart* as a new alternative restarting strategy in the decision making process in structured environments. Finally, our algorithm integrates a mechanism to trace the variations of the underlying graph structure, which captures the causal relationships between the rewards in the bandit setting. Theoretically, we establish a regret upper bound that reflects the effects of the number of structural- and distribution changes on the performance. The outcome of our numerical experiments in real-world scenarios exhibits applicability and superior performance of our proposal compared to the state-of-the-art benchmarks.

4.1 Introduction

Multi-armed bandit (MAB) [126] is a class of sequential learning- and optimization problems. In the seminal MAB problem, the decision-maker (agent) selects one of the K available arms, where each arm returns a reward drawn from a time-invariant, unknown distribution. The agent maximizes the total expected reward over the gambling horizon

by using an effective decision-making strategy that maps the historical actions and outcomes to future actions. That is equivalent to minimizing the total expected *regret*, which is the difference between the reward of the applied policy and that of the optimal policy in hindsight. Indeed, the MAB challenge boils down to the exploration-exploitation dilemma, where the agent decides between accumulating immediate rewards on the one side and obtaining information that might result in a larger reward only in the future on the other side. Due to its wide variety, the MAB framework is a potential candidate as a mathematical tool for tackling many real-world problems, for example, resource allocation in networks [105], recommender systems [95], and clinical trials [11].

The combinatorial multi-armed bandit (CMAB) problem is an extension of the seminal MAB. Instead of only one arm in each round, the agent chooses a number of them, i.e., it takes a combinatorial action. That results in exponential growth of the decision set by increasing the number of arms. Consequently, the conventional MAB methods such as UCB1 [8] become inefficient or inapplicable. In CMAB, we refer to each original arm as a base arm, and any subset of the base arms is a *super arm*. Sometimes, the agent observes the reward of all base arms inside the super arm; In some other cases, the agent observes only one reward. The former type of feedback is a *semi-bandit feedback*, whereas the latter is a *bandit feedback*. The bandit problem becomes aggravated when a statistical structure influences the reward generation processes so that besides the excessively-large action set, the player deals with the structural relationships to decide optimally. We focus on combinatorial semi-bandit (CSB) problem with causally related rewards.

The seminal settings of CMAB- or CSB problems do not assume any statistical or probabilistic relationship between the base arms; Nevertheless, in several application domains, the potential dependency between the random variables can be abstracted by a structure. Despite being neglected for a long time, different types of the MAB problem with probabilistic or statistical relationships between the base arms, referred to as *structured bandits* receive increasing attention from the research community in the past few years. For example, the papers [43, 161, 84] assume that some arms may probabilistically be triggered based on the outcome of other arms. In [90], prior knowledge about the causal structure that affects the rewards is available. The authors in [116] introduce a causally structured CSB problem and use a directed acyclic graph to model the causal structure that influences the reward generation process. Their algorithm does not need *a prior* knowledge concerning the structural relationships as it can learn the structure from the collected data. All of the works mentioned above study a stationary setting.

Unlike the stationary stochastic setting, in many real-world scenarios, the reward distributions of base arms change over time in an evolving environment. For example, in recommender systems, the behavioral feedback of users is time-variant. It is possible to address the nonstationary behaviors of rapidly-varying environments using the adversarial bandit framework [9]. However, in some cases, the environment changes slowly and less frequently. In such scenarios, policies designed for stationary or adversarial bandits are sub-optimal. Generally speaking, there are two main approaches in modeling this

type of nonstationarity in bandit problems; the *switching case* (abruptly changing) [173] and the *dynamic case* (smoothly changing) [44, 150]. For the switching case, the reward distributions of base arms remain unchanged for certain intervals. The environment then varies if the distributions of a subset of base arms change instantly. The point where distributions change is a *change-point* (or breakpoint) and an interval between any two consecutive change-point is a *stationary segment* [173]. In contrast, in the dynamic case, the base arms' mean rewards evolve slowly instead of abruptly changing at one point, and the variation is bounded by a variation budget [22]. In this paper, we focus on the switching case, also referred to as *piecewise* stationary bandit model [23]. We measure the decision-making performance using the notion of *piecewise stationary regret*, i.e., the regret w.r.t. an oracle that knows the best action in each stationary segment.

In a piecewise stationary structured bandit problem, the reward generation processes might vary by changing the base arms' reward distributions and the structural relationships between the variables. Although such a model has remained unaddressed in the MAB literature, it accommodates several real-world applications. Those include financial markets, where not only the *investors' stock purchasing behavior* but also the *causal effects amongst the stock prices* can be time-varying [134]. In such scenarios, an optimal investor follows both sources of change and adapts accordingly. While the availability of prior knowledge about the structural relationships is a strong and unrealistic assumption, inferring such structural relationships from the collected partial feedback in the bandit setting is also challenging. We study a piecewise stationary structured CSB problem, where the causal relationships between the rewards and the distributions of the base arms evolve. In order to model the structural relationships, we rely on *Structural Equation Models* [61].

In general, there are two main approaches to follow the piecewise stationary behaviour of base arms distributions; *passively adaptive* approach [58, 159] and *actively adaptive* approach [23, 33, 66, 101, 173, 45]. Methods of the former category are unaware of the change-points and rely on their understanding of the optimal action based on the most recent observations. On the contrary, methods of the latter category use a change detection algorithm to follow the distributions' changes and decide accordingly [23]. Some studies show the superior performance of actively adaptive approaches [110]. Clearly, the performance of actively adaptive approaches rely significantly on the ability of the agent in handling the breakpoints. Current actively adaptive algorithms incorporate either *global restart* or *local restart* to restart learning the expected value of the instantaneous rewards of the base arms. The former method resets learning the expected values of all arms after detecting a change in one of them. The latter restarts learning only for those arms undergoing a change. These approaches suffer from a drawback as they ignore possible relationships amongst arms' distributions in making a decision upon restarting process. There are main reasons for introducing a new restarting strategy for bandit algorithms in structured piecewise stationary environments. Firstly, social networks, as one of the main target applications for bandit algorithms, exhibit large modularity measures [114, 26]. Secondly, in some real-world scenarios changes within a network are not completely in-

dependent, but they are rather the result of the local spread of a change-seed within the network structure through mechanisms such as contagion [53], social influence [53], or diffusion [144], e.g. media-based marketing campaigns, or rumor diffusion over social networks [132]. In this regard, we introduce the notion of *group restart* where we restart the set of arms that are in the same group, upon detecting a change in any of them. We elaborate more on this in the following sections. We show the superior adaptation capabilities of this approach over local and global restarts in our experiments and discuss the effects of this approach over the upper bound of regret in theory.

In this work, we introduce a piecewise stationary CSB problem with causally related rewards. Our framework accommodates the changes in the base arms' reward distributions and also in the causal relationships between the rewards. We provide an actively adaptive approach to tackle the problem. We introduce a novel alternative restarting strategy, namely *group restart*, that can be used in the adaptation of stationary bandit algorithms to the piecewise stationary environments. We highlight the importance of using the knowledge of relationships amongst arms' distributions in our group restart strategy. We achieve this by showing its effects on the regret of the algorithm in dealing with the costly effects of both *restarts* and *delays of change point detectors*. Our algorithm uses a UCB-based policy for learning the expected rewards of the base arms and a GLR change-point detector. Furthermore, we integrate a mechanism in our algorithm to follow the changes of the causal graph structure that models the causal relationships between the rewards in the bandit setting. We provide the theoretical analysis of the regret upper bound for our algorithm. Our regret bound reflects the effects of both the number of causal graph changes and the number of distribution changes. Our numerical experiments using synthetic- and real-world data establish the advantage of our algorithm compared to the benchmarks.

In Section 4.2, we introduce the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In Section 4.3, we develop our decision-making policy, namely, PS-SEM-UCB-Gr. Section 4.4 presents the theoretical analysis of the regret performance of PS-SEM-UCB-Gr. Section 4.5 includes the numerical experiments. Section 4.6 concludes the paper with some suggestions for future works.

4.2 Problem setup

In a **piece-wise stationary combinatorial semi-bandit (PSCSB)** problem with causally related rewards, a change from one stationary segment to the other results from varying (i) base arms' reward distributions or (ii) the causal relationships between rewards. The intervals with fixed reward distributions and static causal graph are distribution- and graph stationary segments, respectively. The change-points of both segment types appear randomly. We use $\mathcal{K} = \{1, \ldots, \mathrm{K}\}$ to represent the set of K base arms, $\mathcal{D} \subseteq 2^{\mathcal{K}}$ the set of all super arms, and $\mathcal{T} = \{1, \ldots, T\}$ a sequence of T time-steps. Besides, $\theta_{k,t}$ is the distribution of the instantaneous reward of arm k at time t with mean $\mu_{k,t}$ and bounded

support within $[0,1]$. The vector $\boldsymbol{\mu}_t = [\mu_{1,t}, \ldots, \mu_{K,t}]$ is the expected values of the instantaneous rewards of all base arms at time t. Additionally, $\mathcal{A}_t$ is the underlying graph that shows the causal relations between the base arms' rewards. We use $\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t)$ to denote the agent's expected reward from the decision vector $\mathbf{x}_t$ given $\boldsymbol{\mu}_t$ and $\mathcal{A}_t$. Consequently, we characterize a PSCSB with the tuple $(\mathcal{K}, \mathcal{D}, \mathcal{T}, \left\{\theta_{k,t}\right\}_{k\in\mathcal{K}, t\in\mathcal{T}}, \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t))$. Vector of base arms' instantaneous rewards at time t is represented by $\mathbf{b}_t = [b_{1,t}, \ldots, b_{K,t}] \in [0,1]^K$ and it follows a piece-wise independent and identically distributed (i.i.d.) model in each distribution stationary segment. A change to the distribution stationary segment of the environment corresponds to a change in at least one arm's reward distribution. Our setting assumes that the agent is given the meta information regarding the grouping (clustering) of arms such that arms within the same group tend to have their instantaneous rewards' distributions changed together. We use g to denote a group of arms. $K_g = |g|$ is used to show the cardinality of the group g. g_k represents the group to which arm k belongs. We use G to denote the set of all groups, $G = \left\{g^{(1)}, \ldots, g^{(\zeta)}\right\}$, with $|G| = \zeta$ and $\bigcup_{i\in[\zeta]} g^{(i)} = \mathcal{K}, \forall i, j \in [\zeta], g^{(i)} \cap g^{(j)} = \varnothing$ where $[\zeta] = \left\{1, \ldots, \zeta\right\}$. N_Θ is used to denote the number of distribution stationary segments of the environment. We define the total number of distribution stationary segments for group g as

$$N_g = 1 + \sum_{t=1}^{T-1} \mathbb{1}\left\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\right\}. \tag{4.1}$$

Hence, the total number of stationary segments for all groups is $N_G = \sum_{g\in G} N_g$. This clarifies that N_G can change depending on the way the grouping is performed. At each time t, the agent selects a *decision vector* $\mathbf{x}_t = [x_{1,t}, \ldots, x_{K,t}] \in \{0,1\}^K$. We use $\mathcal{I}_t \subset \mathcal{K}$ to denote the set of chosen base arms $I_t \in \mathcal{K}$ at round t. We have $x_{k,t} = 1$ if the base arm k is in the super arm $\mathcal{I}_t$ at time t, otherwise $x_{k,t} = 0$. The agent selects at most m base arms at each time step. Hence, we define the set of all feasible decision vectors as $\mathcal{X} = \left\{\mathbf{x} \mid \mathbf{x} \in \{0,1\}^K \wedge \|\mathbf{x}\|_0 \leq m\right\}$ where $\|\cdot\|_0$ determines the number of non-zero elements in a vector and the parameter m is pre-determined. The causal relationships in the environment are modelled using a directed graph. More precisely, we consider an unknown piecewise static sparse Directed Acyclic Graph (DAG), $\mathcal{A}_t = (\mathcal{V}, \mathcal{E}_t, \mathbf{W}_t)$. $\mathcal{V}$ represents the set of K vertices, i.e., $|\mathcal{V}| = K$, $\mathcal{E}_t$ and $\mathbf{W}_t$ denote the edge set and the weighted adjacency matrix at time t, respectively. We allow the edge set $\mathcal{E}_t$ to change arbitrarily every time the causal graph structure changes. However, the set of vertices $\mathcal{V}$ stays unchanged across time. This implies that the adjacency matrix $\mathbf{W}_t$ changes only in the elements $\mathbf{W}_t[i,j], \forall i, j \in \mathcal{K}, \forall t \in \mathcal{T}$, as far as the underlying graph structure remains a DAG without self-loop, i.e., $\mathbf{W}_t[i,i] = 0, \forall i \in \mathcal{K}, \forall t \in \mathcal{T}$. $N_{\mathbf{W}}$ represents the number of graph stationary segments. An error-free piecewise static Structural Equation Model (SEM) [61] is used to model the generation of reward in the environment. At each time t, $\mathbf{z}_t = [z_{1,t}, \ldots, z_{K,t}]$ is used to represent the exogenous input vector while

$\mathbf{y}_t = [y_{1,t}, \ldots, y_{K,t}]$ denotes the endogenous output vector of the SEM. We write,

$$\mathbf{z}_t = \text{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{4.2}$$

where $\text{diag}(\cdot)$ represents a diagonal matrix. This implies that the exogenous input $\mathbf{z}_t$ contains the semi-bandit feedback in the decision-making problem. We define the k^{th} element of the endogenous output vector $\mathbf{y}_t$ at any time t as

$$y_{k,t} = \sum_{j=1}^{K} \mathbf{W}_t[k,j] y_{j,t} + z_{k,t}, \quad \forall k \in \mathcal{K}, \tag{4.3}$$

At each time t, the endogenous output $y_{k,t}$ represents the *overall reward* of base arm $k \in \mathcal{K}$. The element $\mathbf{W}[k,j]$ represents the causal effect of the overall reward of base arm j on the overall reward of base arm k. Therefore, the overall rewards of base arms are causally related while the instantaneous reward of arm k only directly contributes to the overall reward of arm k. It is important to distinguish between the relationships amongst arms' distributions and the causal relationships amongst the overall rewards. The first one only explains the prior information regarding the groupings of arms, while the second one is used in the mathematical formulation of the problem.

The adjacency matrices $\mathbf{W}_t, \forall t \in \mathcal{T}$ are unknown a priori and $\mathbf{W}_t[i,j] \geq 0, \forall i,j \in \mathcal{K}, \forall t \in \mathcal{T}$. The matrix form of Equation (4.3) at time t is given as

$$\mathbf{y}_t = \mathbf{W}_t \mathbf{y}_t + \mathbf{z}_t. \tag{4.4}$$

As a result, we write $\mathbf{y}_t = (\mathbf{I} - \mathbf{W}_t)^{-1} \text{diag}(\mathbf{b}_t)\mathbf{x}_t$ by solving Equation (4.4) for $\mathbf{y}_t$, where $\mathbf{I}$ is the identity matrix. We assume that the agent is able to observe both the instantaneous semi-bandit feedback vector $\mathbf{z}_t$ and the overall reward feedback vector $\mathbf{y}_t$. The *payoff* received by the agent upon choosing the decision vector $\mathbf{x}_t$ is defined as

$$r_t(\mathbf{x}_t) = \mathbf{c}^\top \mathbf{y}_t = \mathbf{c}^\top (\mathbf{I} - \mathbf{W}_t)^{-1} \text{diag}(\mathbf{b}_t)\mathbf{x}_t, \tag{4.5}$$

where $\mathbf{c} = [c_1, \ldots, c_K] \in \{0,1\}^K$ is pre-determined. The agent is interested in the output y_k in the causal network if $c_k = 1$, and $c_k = 0$ otherwise. Since the graph $\mathcal{A}_t$ is a DAG, the adjacency matrix $\mathbf{W}_t$ is nilpotent. This property guarantees that the matrix $(\mathbf{I} - \mathbf{W}_t)$ is invertible. Given a decision vector $\mathbf{x}_t \in \mathcal{X}$, the expected payoff at time t is calculated as

$$\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x}_t) = \mathbb{E}\left[r_t(\mathbf{X}) | \mathbf{X} = \mathbf{x}_t\right], \tag{4.6}$$

where the expectation concerns the randomness in the reward generating process. We denote by $\mathbf{x}_t^* = \underset{\mathbf{x} \in \mathcal{X}}{\text{argmax}}\ \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\mathbf{x})$ the decision vector with maximum expected reward at

time t. The agent minimizes the cumulative piecewise stationary regret defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(\psi_{\boldsymbol{\mu}_t,\mathcal{A}_t}(\mathbf{x}_t^*) - \psi_{\boldsymbol{\mu}_t,\mathcal{A}_t}(\mathbf{x}_t))\right]. \tag{4.7}$$

4.3 Decision-making strategy

In this section, we develop a solution to the formulated problem. We first introduce the group restart strategy, and the online graph learning. Afterward, we present our decision-making policy, namely, PS-SEM-UCB-Gr.

4.3.1 Group restart strategy.

Restarting process plays a key role in the decision making strategy in piecewise stationary bandit algorithms. Upon taking the global restart strategy, the agent's regret increases due to the costly effects of restarting of all arms. Moreover, by taking local restart strategy, delays of change point detectors for different arms can make the algorithm to incur linear regret in some intervals. One way to address these issues is in the structured environments where changes are not always completely independent and having side information w.r.t. relationships between arms' distributions can be helpful in making decisions upon restarts. There are certain research directions in MAB literature where relationships amongst the arms are considered. For instance, in [153], it is assumed that each item that the algorithm recommends is a node of a known graph and the expected rating of the neighboring nodes are similar. Furthermore, in [59], it is suggested that the nodes of the graph can be clustered according to some a priori unknown clustering and the arms within the same cluster exhibit similar behaviours. Also, in [169], the relationship between the users is captured by an underlying graph and user preferences are assumed to have smooth signals on the graph. In such settings, it is natural to anticipate that if an arm's expected reward is changed, then due to the relationships of the arms, the set of arms that are closely connected to it go through changes as well. Consequently, we propose *group restart* strategy as an efficient alternative in structured environments where grouping information might be either available in advance or learned from the data [59, 96]. As the result of our theoretical analysis, we show that a structure-based grouping in group restart strategy can help to reduce the regret upper bound compared to local and global restarts.

4.3.2 Piece-wise static graph learning.

Considering the required knowledge of $\mathbf{W}_t$ in finding the optimal decision vector, we propose an online graph learning framework that uses the collected feedback $\mathbf{y}_t$ and $\mathbf{z}_t$ and allows for modelling both the random and the smooth transitions of the causal graph.

At each time t, we stack the feedback, from the last graph-change-point up to the current time, as consecutive columns in $\mathbf{Z}_t$ and $\mathbf{Y}_t$, hence, $\mathbf{Y}_t = \mathbf{W}_t\mathbf{Y}_t + \mathbf{Z}_t$. We use the collected feedback history, $\mathbf{Y}_t$ and $\mathbf{Z}_t$, as the input to a parametric graph learning algorithm for a static SEM [61]. Formally, the adjacency matrix at time t is the solution to the following optimization problem:

$$\begin{aligned} \hat{\mathbf{W}}_t = \underset{\mathbf{W}\in\mathbb{R}^{K\times K}}{\operatorname{argmin}} \quad & \|\mathbf{Y}_t - \mathbf{W}\mathbf{Y}_t - \mathbf{Z}_t\|_{\mathrm{F}}^2 + \lambda_1 \|\mathbf{W}\|_1 \\ \text{s.t.} \quad & \mathbf{W}[k,k] = 0, \forall k \in \mathcal{K} \end{aligned} \tag{4.8}$$

where $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm of matrices. The symbol $\|\cdot\|_1$ denotes the L^1-norm of the matrices and it is used to impose sparsity over the estimated adjacency matrix $\hat{\mathbf{W}}_t$. We use the notation $\hat{\mathbf{W}}^{(i)}$ to represent the estimated adjacency matrix for the i^{th} static graph. In order to impose slow topological variations across time, from one static graph to the next, one may add a second regularization term $\lambda_2 \left\|\hat{\mathbf{W}}^{(i+1)} - \hat{\mathbf{W}}^{(i)}\right\|_1$ in Equation (4.8) and have a form of the optimization problem in Equation (4.8) that stays convex. This second regularization allows the algorithm to penalize deviation of the current graph estimate from the predecessor, hence implementing a transfer of knowledge that is gained from the previous segment.

4.3.3 PS-SEM-UCB-Gr algorithm

In this section, we describe our decision-making policy. Its core is the Upper Confidence Bound policy. Besides, we use two previously-proposed methods, namely *group restart*, and *piece-wise static graph learning*. Finally, we integrate a mechanism for detecting the changes to the adjacency matrix of the causal graph. Each time the algorithm decides to infer the new adjacency matrix, it starts a subroutine inside the main algorithm to obtain K data samples by interacting with the new environment. It is crucial that the new dataset satisfies the conditions for the precise inference and unique identification of the new graph adjacency matrix [20, 116]. We refer to this subroutine as *Graph Learning Data Generation* (GLDG). For these K rounds, PS-SEM-UCB-Gr picks K columns of an *initialization matrix*, namely, $\mathbf{Init} \in \{0,1\}^{K\times K}$ in a sequential way where **Init** is created as described in Section 3.3.2. Based on the discussion above, we assume that there are at least $K+1$ rounds between any two consecutive changes in the graph. That guarantees sufficient time to infer the new ground truth graph after every change. We refer to the rounds inside a GLDG phase as *graph initialization* rounds and the rest as *normal* rounds. In every round, the GLR change-point detectors and the UCB index developments are working. The input parameters of PS-SEM-UCB-Gr include the number of steps (T), number of arms (K), uniform exploration probability $p \in (0,1)$, and δ as the confidence level of the GLR change-point detector. The policy uses the parameter τ' to perform the uniform forced exploration over all base arms in Algorithm 3. The forced uniform exploration guarantees that the GLR change-point detectors receive suf-

Initialization: $\forall k \in \mathcal{K}$, $n_{k,0} \leftarrow 0$, $\hat{\mu}_{k,0} \leftarrow 0$, $\tau_k = 0$, $t = 1$, $\tau' = 0$, $flag = 1$.
Get $G = \left\{ g^{(1)}, \ldots, g^{(\zeta)} \right\}$
while $t \leq T$ **do**
  if $flag = 1$ **then**
    Run ***GLDG***.
  end if
  if $\Omega \neq \varnothing$ **then**
    Pick $a \in \Omega$, Randomly choose $\mathcal{I}_t$ with $a \in \mathcal{I}_t$.
    Remove a from Ω.
  else
    Solve Eq. (4.11) for $\mathbf{x}_t$.
  end if
  Play $\mathcal{I}_t$, receive reward $r(\mathbf{x}_t)$, $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$.
  for all $I_t \in \mathcal{I}_t$ **do**
    update: $\hat{\mu}_{I_t,t}$ using Eq. (4.9), $n_{I_t,t}$ using Eq. (4.10).
    if $\mathbf{GLR}(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **then**
      $\forall k \in g_{I_t}$: $n_{k,t} \leftarrow 0$, $\hat{\mu}_{k,t} \leftarrow 0$, $\tau_k \leftarrow t$.
      $\tau' \leftarrow t$, $\Omega \leftarrow \Omega \cup g_{I_t}$.
    end if
  end for
  if $\exists\, c \in \mathbb{N} : t - \tau' = c \left\lfloor \frac{K}{p} \right\rfloor$ **then**
    $\Omega = \bigcup_{i \in [\zeta]} g^{(i)}$
  end if
  for all $k \in \mathcal{K}$ **do**
    if $n_{k,t} \neq 0$ **then**
      $U_{k,t} \leftarrow \hat{\mu}_{k,t} + \sqrt{\frac{(m+1)\log(t-\tau_k)}{n_{k,t}}}$
    end if
  end for
  $[\mathbf{Y}_t] \leftarrow [\mathbf{Y}_{t-1}, \mathbf{y}_t]$, $[\mathbf{Z}_t] \leftarrow [\mathbf{Z}_{t-1}, \mathbf{z}_t]$
  if $\left\| \mathbf{y}_t - \hat{\mathbf{W}}_{t-1}\mathbf{y}_t - \mathbf{z}_t \right\|_2^2 > \varepsilon$ **then**
    $flag = 1$, $\mathbf{Y}_t = []$, $\mathbf{Z}_t = []$
  else
    Solve Eq. (4.8) to get $\hat{\mathbf{W}}_t$.
  end if
  $t \leftarrow t + 1$
end while

3: PS-SEM-UCB-Gr: Piecewise Stationary - Structural Equation Model - Upper Confidence Bound - Group Restart

1: Create an initialization matrix **init**, $\mathbf{Y}_0 = []$, $\mathbf{Z}_0 = []$.
2: **for** $t' = 1, 2, \ldots, K$ **do**
3: $\quad \mathbf{x}_t := \mathbf{init}[:, t']$
4: $\quad$ Play $\mathcal{I}_t$, receive reward $r(\mathbf{x}_t)$, $s_{I_t, n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$.
5: $\quad$ **for all** $I_t \in \mathcal{I}_t$ **do**
6: $\quad\quad$ update: $\hat{\mu}_{I_t,t}$ using Eq. (4.9), $n_{I_t,t}$ using Eq. (4.10).
7: $\quad\quad$ **if GLR**$(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **then**
8: $\quad\quad\quad \forall k \in g_{I_t}$: $n_{k,t} \leftarrow 0$, $\hat{\mu}_{k,t} \leftarrow 0$, $\tau_k \leftarrow t$.
9: $\quad\quad\quad \tau' \leftarrow t$, $\Omega \leftarrow \Omega \cup g_{I_t}$.
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: $\quad$ **for all** $k \in \mathcal{K}$ **do**
13: $\quad\quad$ **if** $n_{k,t} \neq 0$ **then**
14: $\quad\quad\quad U_{k,t} \leftarrow \hat{\mu}_{k,t} + \sqrt{\frac{(m+1)\log(t-\tau_k)}{n_{k,t}}}$
15: $\quad\quad$ **end if**
16: $\quad$ **end for**
17: $\quad [\mathbf{Y}_t] \leftarrow [\mathbf{Y}_{t-1}, \mathbf{y}_t]$, $[\mathbf{Z}_t] \leftarrow [\mathbf{Z}_{t-1}, \mathbf{z}_t]$, $t \leftarrow t+1$
18: **end for**
19: Solve Eq. (4.8) to get $\hat{\mathbf{W}}_{t-1}$.
20: $flag = 0$

2: Graph Learning Data Generation (GLDG)

ficient samples. Considering that we are using group restart, UCB developments of arms from different groups might have different resetting times. Therefore, the policy uses $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_K]$ to manage the restarting times of UCB indices. The variable *flag* is used to call the GLDG subroutine. For any arm k, the empirical average of the instantaneous rewards at any time $t = t_1$ w.r.t. its last restarting time at $t = \tau_k$ yields

$$\hat{\mu}_{k,t_1} = \frac{\sum_{t=\tau_k+1}^{t_1} z_{k,t}}{n_{k,t_1}}, \tag{4.9}$$

where n_{k,t_1} is the number of times that the base arm k is observed up to time $t = t_1$ since its last restart at $t = \tau_k$. Formally,

$$n_{k,t_1} = \sum_{t=\tau_k+1}^{t_1} x_{k,t}. \tag{4.10}$$

The set Ω holds the index of those arms whose UCB developments are being restarted or that are candidates for forced exploration. After the graph initialization period, in each round, PS-SEM-UCB-Gr first checks the set Ω, in Line 7, otherwise the agent plays

the next super arm according to the result of the combinatorial optimization in Line 11. The combinatorial optimization uses the current UCB indices and the last estimate of the causal graph. We denote the UCB index of base arm k at time t as $U_{k,t}$ such that we have the UCB indices of all base arms in the vector $\mathbf{U}_t = [U_{1,t}, \ldots, U_{K,t}]$. Therefore, the combinatorial optimization for finding the best decision vector yields

$$\mathbf{x}_t = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}} \quad \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{U}_{t-1})\mathbf{x}. \tag{4.11}$$

Let $\mathbf{M}^\top = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{U}_{t-1})$. The elements of $\hat{\mathbf{W}}_{t-1}$, $\mathbf{c}$, and $\mathbf{U}_{t-1}$ are non-negative, then the optimization problem Equation (4.11) can be solved by finding a subset of elements in $\mathbf{M}$ such that $\mathbf{x} \in \mathcal{X}$. Therefore, it is solvable by using an efficient sorting algorithm that ranks the elements of $\mathbf{M}$. Consequently, the agent plays $\mathbf{x}_t$, collects the reward in Line 13 according to Equation (4.5), and updates the vectors $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{n}_t$ in Line 15. The notation $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}$ in Line 13 implies that the collected feedback $z_{I_t,t}, \forall I_t \in \mathcal{I}_t$ is the sample number $n_{I_t,t}$ in the sequence of samples for arm I_t since its last restart at $t = \tau_{I_t}$. We use the **GLR** change-point detector [23] defined as

$$\begin{aligned}\mathbf{GLR}(s_1, \ldots, s_n; \delta) = \mathbb{1}\{\sup_{\alpha\in[1,n-1]} [\alpha \times \mathrm{kl}(\hat{\beta}_{1:\alpha}, \hat{\beta}_{1:n}) + \\ + (n-\alpha) \times \mathrm{kl}(\hat{\beta}_{\alpha+1:n}, \hat{\beta}_{1:n})] \geq \gamma(n,\delta)\},\end{aligned} \tag{4.12}$$

where $\hat{\beta}_{\alpha:\alpha'}$ is the mean of the observations between α and α', $\mathrm{kl}(x,y) = x\log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$ is the binary relative entropy between any two Bernoulli distributions with means x and y. The function $\gamma(n,\delta)$ is the threshold function for the GLR test. Theoretically, we choose this threshold function following Lemma 2 in [23]. However, in all our numerical experiments, we follow "Practical considerations" in [23], and select $\gamma(n,\delta) = \ln(\frac{3n\sqrt{n}}{\delta})$. If $\mathbf{GLR}(s_1, \ldots, s_n; \delta) = 1$, the algorithm applies group restarts in Line 17. The algorithm updates the UCB indices in Line 26. In Line 30, the graph-change detection mechanism uses two vectors of $\mathbf{z}_t$ and $\mathbf{y}_t$ to test the validity of the last estimate of the graph adjacency matrix, $\hat{\mathbf{W}}_{t-1}$. If the error value for the graph-change detection formulation exceeds ε, then the algorithm notifies that the previous estimate of the graph structure is no longer valid. Consequently, in Line 31, we have flag $= 1$, and the previously collected sets of feedback in $\mathbf{Z}_t$ and $\mathbf{Y}_t$ are dropped. Parameter ε represents the error we accept in the graph estimation process. In this paper, we take $\varepsilon = 0$. However, assuming $\varepsilon \neq 0$, the effects of ε should be considered in the regret analysis. In case the collected feedback vectors $\mathbf{y}_t$ and $\mathbf{z}_t$ satisfy the SEM formulation for $\hat{\mathbf{W}}_{t-1}$, in Line 29, PS-SEM-UCB-Gr uses the newly updated matrices $\mathbf{Y}_t$ and $\mathbf{Z}_t$, to improve the adjacency matrix estimation. It is important to notice that the algorithm does not restart the UCB development upon detecting a graph-change. It also does not restart the graph learning following any distribution-change detection.

4.4 Theoretical analysis

In this section, we deliver the analysis for the expected regret of PS-SEM-UCB-Gr algorithm. We perform the regret analysis according to any grouping of arms, with local and global restarts as special cases. We denote the maximum delay across all detected changes as d. We divide the time line into stationary segments of base arm distributions. The graph changes will be treated separately as they do not affect the UCB developments and only contribute to the regret in terms of a constant, based on the graph-learning phase. We also define the suboptimality gaps in our setting as the reward difference between the optimal decision vector $\mathbf{x}^*$ and an arbitrary decision vector $\mathbf{x}$: $\Delta_t(\mathbf{x}) = \psi_t(\mathbf{x}_t^*) - \psi_t(\mathbf{x})$, where $\psi_t(\mathbf{x})$ is the mean reward of $\mathbf{x}$, with only subscript t used to parameterize it for better readability. The largest gap is denoted as $\Delta_{\max} = \max_t \max_{\mathbf{x}:\psi_t(\mathbf{x})<\psi_t(\mathbf{x}_t^*)} \Delta_t(\mathbf{x})$, and the smallest $\Delta_{\min} = \min_t \min_{\mathbf{x}:\psi_t(\mathbf{x})<\psi_t(\mathbf{x}_t^*)} \Delta_t(\mathbf{x})$. As it is essential for estimating the total regret bound, we deliver the regret for the stationary case, with an improvement over the work of [116], as our result does not scale with the number of layers in the causal graph but only with the total number of arms;

Lemma 1. *Let* $\omega_t^T = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_{t+1})$ *and* $\omega_{\max} = \max_t \max_k \omega_{k,t}, k \in \mathcal{K}$. *In the stationary case* $(N_\Theta = 1 \wedge N_{\mathbf{W}} = 1)$ *of the PS-SEM-UCB-Gr algorithm, the upper regret bound is given as:*

$$\mathcal{R}(T) \leq \left[\frac{4\omega_{\max}^2 m^2 (m+1) K \log(T)}{\Delta_{\min}^2} + \frac{\pi^2}{3} mK + K \right] \Delta_{\max},$$

with $\Delta_{\max}$ *as the largest suboptimality gap and* $\Delta_{\min}$ *smallest suboptimality gap.*

The adapted proof is given in the supplementary materials. The following Theorem 2 states a bound on the regret in the non-stationary case of our proposed decision-making policy for any grouping of arms.

Theorem 2. *Let* $\omega_t^T = \mathbf{c}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_{t+1})$ *and* $\omega_{\max} = \max_t \max_k \omega_{k,t}, k \in \mathcal{K}$. *The expected regret of the PS-SEM-UCB-Gr policy is upper bounded as:*

$$\mathcal{R}(T) \leq \sum_{g \in G} \left[N_g K_g R_0(T) + (\delta T + 1 + \frac{\pi^2 m}{3}) N_g K_g \Delta_{\max} \right] + \\ + \left(Tp + dN_G + \delta T(K + N_G) + N_{\mathbf{W}} K \right) \Delta_{\max},$$

with $R_0(T) = \frac{4\omega_{\max}^2 m^2 (m+1) \log(T)}{\Delta_{\min}^2} \Delta_{\max}$.

proof. See Section B of the appendix.

This is the general regret upper bound that reflects the importance of grouping of arms. We are able to retrieve the bounds according to the given grouping of arms. We assumed

the knowledge of groupings of base arms based on structural relationships between arms' distributions.

Following local restart strategy, $G = G_{\text{local}}$, we have $K_g = 1$, $\forall g \in G_{\text{local}}$ and $|G_{\text{local}}| = K$, thus $\sum_g N_g K_g = N_{G_{\text{local}}}$. If we follow global restart strategy, $G = G_{\text{global}}$, we have $K_g = K$, $\forall g \in G_{\text{global}}$ and $|G_{\text{global}}| = 1$, thus $\sum_g N_g K_g = K N_{G_{\text{global}}}$. It is important to note that the number of restarts differs for local and global restart strategies, since $N_{G_{\text{global}}} \leq N_{G_{\text{local}}}$. In the following, we compare the performance of our approach with local restarts and global restarts on the amount of regret increase in the distribution stationary segment after a breakpoint.

Remark 4. *We rewrite the regret upper bound in Theorem 2 as* $R(T) \leq \Sigma_{g\in G}[C_1 N_g K_g + C_2 N_{\text{g}}] + C_3$ *where* C_1, C_2, C_3 *are independent of the grouping of arms. Let us assume that the breakpoint* ν *happens from* t *to* $t+1$ *with change to* $\Re_\nu$ *arm distributions that belong to* η_ν *groups (clusters). The increase of the regret value within the stationary segment after breakpoint* ν *can be written as* $\Delta R(\nu) \leq C_1 \Sigma_{g\in G} K_g \mathbb{1}\left\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\right\} + C_2 \Sigma_{g\in G} \mathbb{1}\left\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\right\}$. *Consequently, we have the followings;*

- *In the case of Local restart, we have* $\Delta R(\nu) \leq C_1 \Re_\nu + C_2 \Re_\nu$.
- *In the case of Global restart, we have* $\Delta R(\nu) \leq C_1 K + C_2$.
- *If the total number of arms inside the* η_ν *groups is* $\Re_\nu$*, for Group restart, we have* $\Delta R(\nu) \leq C_1 \Re_\nu + C_2 \eta_\nu$.
- *If in the* η_ν *groups, there are collectively* s *arms whose distributions did not change at* ν*, in this case, for Group restart we have* $\Delta R(\nu) \leq C_1(\Re_\nu + s) + C_2 \eta_\nu$.

In the above, the first term, scaling with C_1, is the regret due to number of restarted arms, while the second term, scaling with C_2, is affected by the delays. These results clarify the idea behind using a group restart strategy, especially in cases where the $\Re_\nu$ changed arms are from a small number of η_ν clusters. Intuitively, in networks with high modularity measures, we can expect to have smaller number for s and a better performance for the group restart strategy.

By the following corollary, through fine-tuning the hyper-parameters δ and p and with the assumption of the prior knowledge of N_G, we can achieve a sub-linear regret bound;

Corollary 1. *Let* $\Delta_{\min}^{change} = \min_i \max_{k\in\mathcal{K}} |\mu_{k,i} - \mu_{k,i-1}|$. *By choosing* $\delta = \frac{1}{T}$ *and* $p = \sqrt{\frac{N_G K \log T}{T}}$*, the regret is upper-bounded by the following,*

$$\mathcal{O}\left(\left(\frac{\sum_{g\in G} N_g K_g \log T}{\Delta_{\min}} + \frac{\sqrt{N_G K T \log T}}{\left(\Delta_{\min}^{change}\right)^2} + N_{\mathbf{W}} K\right) \Delta_{\max}\right)$$

Our regret bound shows an improvement in comparison to the result of [173] in terms of the dependency of total restarts N_G, even though our algorithm does not require the prior knowledge of the causal graphs. In the absence of graph-changes, the respective contribution to the regret stems solely from the very first initialization, i.e., $N_{\mathbf{W}} = 1$.

4.5 Experimental analysis

In this section, we evaluate the performance of our proposed decision-making policy using synthetic- and real-world datasets by comparing it with the following state-of-the-art combinatorial semi-bandit algorithms as benchmarks; **CTS** [73] is a Thompson sampling-based algorithm for stationary environments; **GLR-CUCB** [173] is a UCB-based algorithm for piecewise stationary environments. It employs a GLR change-point detector and uses a global restart strategy. We implemented the same algorithm with local restarts and group restarts, GLR-CUCB-Lo and GLR-CUCB-Gr, respectively; **CUCB-SW** [44] is an algorithm that uses a sliding window to follow the base arms' distribution changes while developing the corresponding UCB indices; **Orc-R** is PS-SEM-UCB-Gr with the Oracle-Restart. This algorithm is given the prior information w.r.t. all distribution change-points and it only restarts the groups where a change is detected. In addition, we implement the PS-SEM-UCB-Gl and PS-SEM-UCB-Lo that are working based on global restart and local restart strategy, respectively.

All three algorithms *CTS*, *GLR-CUCB*, and *CUCB-SW* require access to the exact- or to an approximation oracle that solves the combinatorial optimization in Equation (4.11); that is, they need prior knowledge of the ground truth causal graph at any time. Such a strong assumption renders them inapplicable in the absence of such prior knowledge. For a fair comparison, we apply all benchmarks to the instantaneous rewards feedback vector $\mathbf{z}_t$ at each time t. We implemented the exact optimization oracles for *CTS*, *GLR-CUCB*, *GLR-CUCB-Gr*, *GLR-CUCB-Lo*, and *CUCB-SW*.

4.5.1 Synthetic dataset

In the following, we describe the synthetic dataset used in the experiments. It has 4 graph-change-points and 4 distribution-change-points. For all different graph structures, we have $K = 18$ nodes. We draw the elements of the adjacency matrices $\mathbf{W}_t$ from a uniform distribution over $[0.1, 0.9]$. The edge density of the ground truth adjacency matrices is 0.15. The $K = 18$ arms are divided into 3 groups of 6 arms. We select $m = 4$ in this experiment and $T = 25000$. At each time t, the vector of instantaneous rewards $\mathbf{b}_t$ follows a multivariate normal distribution with the support in $[0, 1]^{18}$ and a spherical covariance matrix. In appendices section of the thesis, Figure B.1 visualizes the expected values of base arms' rewards across time, and Figure B.2 presents the visualization of optimal super arm across time. As shown in Section 4.2, the reward generation process follows the SEM in Equation (4.3). All distribution-stationary-segments of the environ-

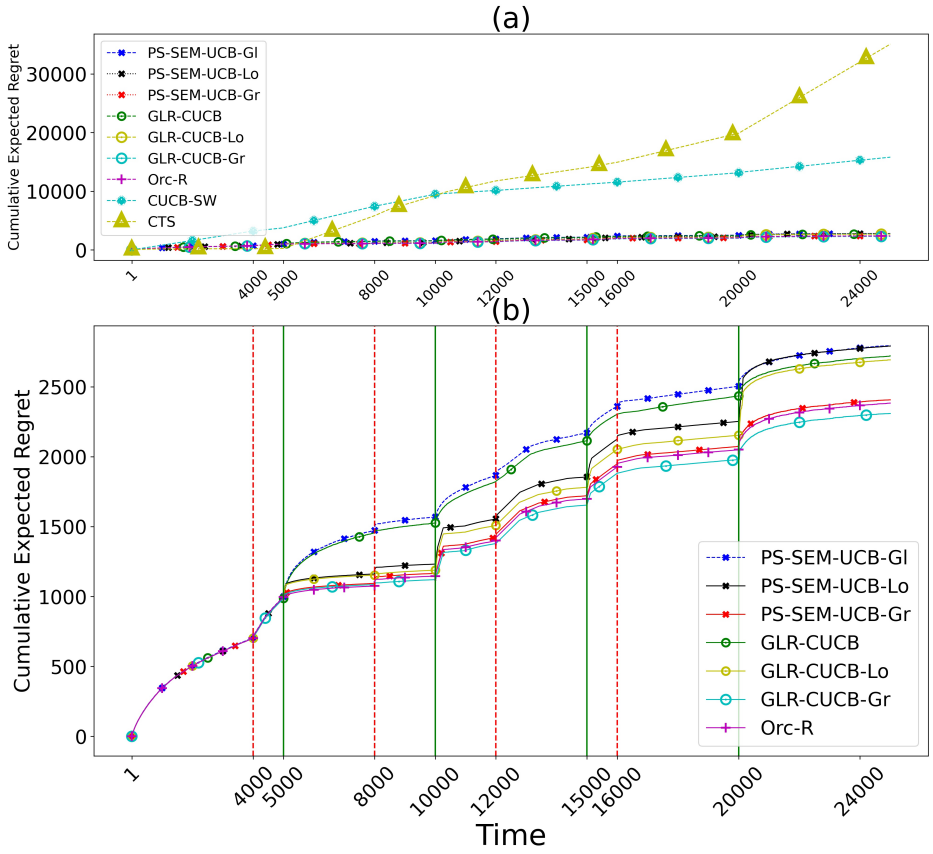


Figure 4.1: Cumulative Expected Regret.

ment have the same lengths. The regularization parameter λ_1 is tuned by grid search over $[0.0001, 10000]$. We evaluate the estimated adjacency matrix at each time t by using the mean squared error defined as $\text{MSE} = \frac{1}{K^2}\left\|\mathbf{W}_t - \hat{\mathbf{W}}_t\right\|_{\text{F}}^2$. Figure 4.1-a shows the poor performance of *CUCB-SW*, and *CTS*, compared to other algorithms. In Figure 4.1-b, we highlight the differences in the performance of *Orc-R*, PS-SEM-UCB, *GLR-CUCB* under various restarting strategies. One can observe the better performance of PS-SEM-UCB-Gr compared to *GLR-CUCB*. That happens although PS-SEM-UCB-Gr does not require prior knowledge of the distribution-change-points and the causal graphs. The effects of different restarting strategies can be observed as well. Global restarts adds to the regret significantly by restarting the entire set of base arms. On the opposite, local restart suffers from the delay on those breakpoints where the number of changed distributions is large. In Figure 4.1-b, each vertical green solid line represents the time of a distribution-change, and each vertical red dashed line represents a graph-change.

4.5.2 Real-world application

In this section, we provide the results of applying our algorithm, to the Covid-19 outbreak dataset of daily new infected cases during the pandemic in different regions within Italy.[1] The goal is to find a subset of regions with the highest contribution to the spread of the virus in the country in a non-stationary period. We use the *overall reward* y_i for the *overall daily new cases* in region i. Besides, we use the *instantaneous reward* b_i for the *region-specific daily new cases* in region i. The data of the period from 3^{rd} July 2020, to 10^{th} October 2020 was used. We pre-process the dataset following [116]; nevertheless, we use a 14-day moving average instead of a 7-day moving average. Instead of the L^1-norm in Equation (4.8), we use the Directed Total Variation (DTV) $\sum_{i,j\in\mathcal{K}} \mathbf{W}[i,j] \sum_{h=1,\dots,t} \left[\mathbf{Y}[i,h] - \mathbf{Y}[j,h]\right]^+$ regularizer [128], where $[y]^+ = \max\{y,0\}$. Since the causal spread of the disease might create cycles, we allow cyclic graphs as the result of the optimization problem in Equation (4.8). Considering that the ground truth graphs are not available, we use a cross-validation technique to tune the regularization parameter λ_1. We split the data into 10 subsets of 10 consecutive days. In each subset, one day is chosen uniformly at random to be included in the validation set, while the remaining 9 days are added to the train set. We calculate the prediction error at each time t by $Error(t) = \frac{1}{K|\mathfrak{v}(t)|} \sum_{i\in\mathfrak{v}(t)} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1$ where $\mathfrak{v}(t)$ is the validation set at time t with cardinality $|\mathfrak{v}(t)|$. Besides, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are respectively the validation data, and the corresponding predicted value using the estimated graph for day i. Figure 4.2 compares the ground truth overall daily new cases and the predicted total daily new cases using the estimated graph in 3 days of the Covid-19 outbreak in our validation data.[2] According to Figure 4.2, our algorithm estimates the data for each region efficiently. This helps the agent to find the optimal decision vector. Regarding that the benchmarks need the prior knowledge of the causal graph, this real-world application highlights the drawbacks of the benchmarks. Considering the impacts of geographical factors on Covid-19 cases [160], we divide the country into 4 clusters, using *graph-based clustering* library of *Python*, based on Euclidean distances between regional capitals. In Figure 4.3, we show the regions that PS-SEM-UCB-Gr selects over time. On each day, the selected regions are highlighted by dark rectangles. PS-SEM-UCB-Gr finds changes in the distribution of the region-specific daily new cases of different regions belonging to each group. Consequently, it restarts the UCB procedure for all the groups within the period $t = 58$ and $t = 79$. Due to space limitations, the details about the groupings and their change-detection times are mentioned in the supplementary. We see that selected subsets of regions before and after the restart of the algorithm are different due to newly calculated UCB indices after the restarts. This shows how the main contributors to the spread of the virus changed from one stationary segment to the next.

[1] https://github.com/pcm-dpc/COVID-19

[2] Due to space limitations, we use abbreviations for region names. Table A.1 in the appendix lists the abbreviations together with the original names of the regions.

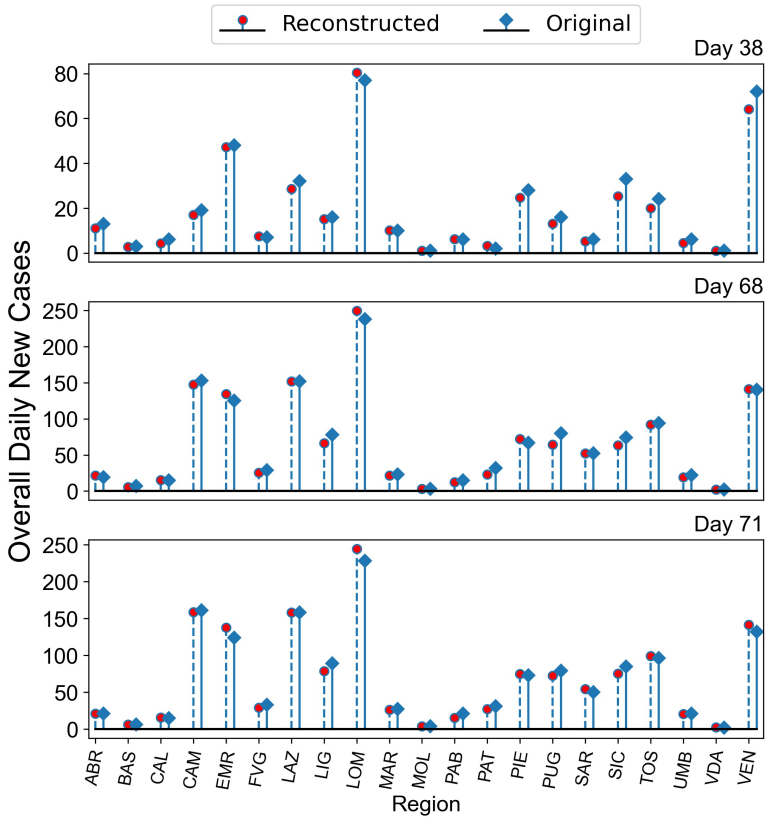


Figure 4.2: Original and reconstructed daily new cases.

4.6 Conclusion

In this paper, we developed a piecewise stationary combinatorial semi-bandit framework with causally related rewards. We developed a decision-making policy that follows distribution- and causal graph changes to adapt the decisions. We introduced a new alternative for the restarting process of bandit algorithms in structured environments under piecewise stationary settings. We proved that PS-SEM-UCB-Gr achieves a sublinear regret bound. The experiments showed the superior performance of PS-SEM-UCB-Gr compared to several state-of-the-art combinatorial algorithms. Our regret analysis clarifies the effects of global and local restarts as special cases of group restarts. It clarifies the importance of using relationships amongst base arms' distributions for the purpose of grouping of arms to minimize the regret incurred by the restarting process in group restarts. As for future research direction, we aim at studying our problem under the presence of noise in the SEM.

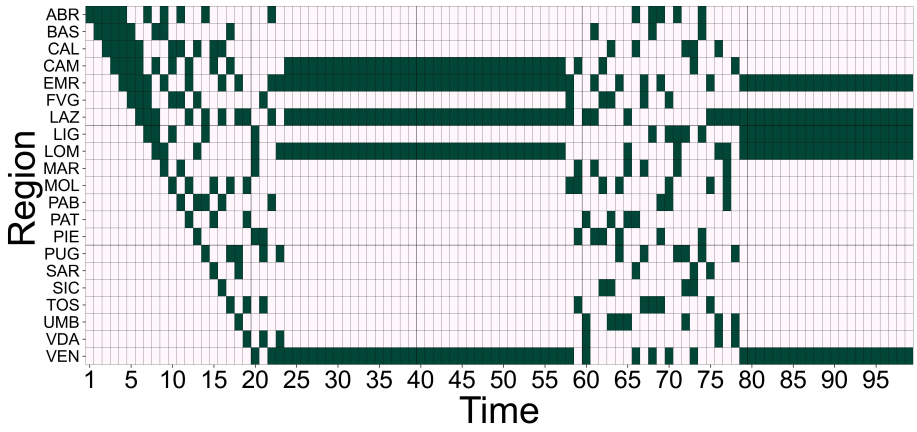


Figure 4.3: Selected regions on each day of the experiment.

Chapter 5

Clusters Agnostic Network Lasso Bandits

We consider a multi-task contextual bandit setting, where the learner is given a graph encoding relations between the bandit tasks. The tasks' preference vectors are assumed to be piecewise constant over the graph, forming clusters. At every round, we estimate the preference vectors by solving an online network lasso problem with a suitably chosen, time-dependent regularization parameter. We establish a novel oracle inequality relying on a convenient restricted eigenvalue assumption. Our theoretical findings highlight the importance of dense intra-cluster connections and sparse inter-cluster ones. That results in a sublinear regret bound significantly lower than its counterpart in the independent task learning setting. Finally, we support our theoretical findings by experimental evaluation against graph bandit multi-task learning and online clustering of bandits algorithms.

5.1 Introduction

Online commercial websites aim to recommend their products to their customers properly, and the performance of these recommendations depends on the knowledge of users' preferences. Unlike traditional collaborative-filtering-based methods [139], such knowledge is initially unavailable. Therefore, the online recommender systems need to recommend various items to the users and observe their ratings to *explore* their preferences. At the same time, the recommender system should be able to recommend items that attract users' attention and receive high ratings by *exploiting* the learned knowledge. The contextual bandit frameworks [95] have been popularly used to formalize and address this exploration-exploitation trade-off.

However, the classical form of contextual bandits [95, 47, 2] ignores the availability of social networks amongst users and solves the problem for each user separately. Consequently, such algorithms have some drawbacks when applied to problems with a large number of users. First, such a large number hinders their computational efficiency. Sec-

ond, the partial feedback of the bandit settings exposes the algorithms to having weak estimations and impairing their decision-making ability [169]. Consequently, to improve bandit algorithms' performance for large-scale applications, structural assumptions that link the different users are usually integrated within bandit algorithms [38, 59, 97, 68].

Cesa-Bianchi *et al.* [38] and Yang *et al.* [169] integrate the prior knowledge of social networks into their contextual bandit algorithms. Both papers propose UCB-style algorithms and exhibit the importance of using the social network graph to achieve lower regrets using Laplacian regularization. The latter regularization promotes smoothness among the preference vectors of users, allowing the transfer of the collected information between them. However, the Laplacian regularization does not account for the smoothness heterogeneity introduced by a piecewise constant behavior over the graph [162]. On the other hand, algorithms of online clustering of bandits [59, 97] tackle such a piecewise constant behavior by explicitly estimating user clusters. However, their clustering can cause overconfidence in the constructed clusters, potentially leading to error accumulation.

In this paper, we assume access to a graph encoding relations between bandit tasks, and that the task parameter vectors are piecewise constant over the graph. We propose an algorithm that integrates the prior knowledge of the piecewise constant structure to update tasks rather than finding the clusters explicitly. That way, we mitigate the limitations mentioned above: the piecewise constant smoothness is naturally integrated into our regularizer, and we do not estimate the clusters so our algorithm does not suffer from overconfidence drawbacks.

More precisely, we provide the following contributions

- We analyze an instance of the Network Lasso problem [65], estimating every vertex's preference vector using data generated during the interaction between users and the bandit. We provide the first oracle inequality in this setting and link it to fundamental quantities characterizing the relation between the graph and the true preference vectors of the users. Our result relies on our novel restricted eigenvalue (RE) condition, which we assume for our setting. This result is of independent interest and can be applied to i.i.d. data as a special case.

- We prove that the empirical multi-task Gram matrix of the data inherits the RE condition from its true counterpart. Both this result and the previous one depend on the sparsity of inter-cluster connections and the density of intra-cluster ones.

- We provide a regret upper bound for our setting. Our bound highlights the advantage of our algorithm in high dimensional settings, and for large graphs.

- We support our theoretical findings by extensive numerical experiments on simulated data that prove the advantage of our algorithm over other related approaches.

The rest of the chapter is organized as follows. Section 5.2 discusses the relation of our work to the literature. We formulate our problem and state some of our assumptions

in Section 5.3, then present our bandit algorithm in Section 5.4. We analyze the problem theoretically in Section 5.5 and demonstrate its practical interest experimentally in Section 5.6.

5.2 Related work

Lasso contextual bandits. To address the high dimensional setting for linear bandits, several multi-armed bandit papers solve a LASSO [147] problem under different assumptions [18, 81, 119, 6]. They all rely on a previously established compatibility or RE condition [29], that they adapt to the non-i.i.d case resulting from the context selection procedure across rounds. Such assumptions were also used in the multi-task setting by Cella and Pontil [34] with a Group Lasso regularization [171], and to impose a low-rank structure on the task preference vectors in [36]. In our case, we establish a novel oracle inequality, rather than only generalize an existing one to the non-i.i.d setting, with a newly introduced RE assumption, which can be of independent interest.

Clustering of bandits. Gentile *et al.* [59] introduced sequential clustering of bandits with the CLUB algorithm. The latter starts with a fully connected graph, and then an iterative graph learning process is performed, where edges between users are deleted if their preference vectors are significantly different. As a result, any connected component is seen as a cluster and only one recommendation per cluster is developed. The SCLUB algorithm of Li *et al.* [97] generalizes CLUB via including merging operations in addition to splitting. In contrast to these approaches, Nguyen and Lauw [115] groups users via K-means clustering, and Cheng *et al.* [46] rely on hedonic games for online clustering of bandits. Furthermore, Yang and Toni [168] make use of community detection techniques on graphs to find user clusters. Gentile *et al.* [60] study the clustering of the contextual bandit problem where their proposed algorithm, named CAB, adaptively matches user preferences in the face of constantly evolving items. Our work fundamentally differs from the previous ones on two aspects. First, we assume access to a graph encoding relations between users, which is more informative than a complete graph. Second, we do not keep track of a model for each cluster, but rather we integrate a prior over the graph via a graph total variation regularizer that enforces a piecewise constant behavior for the estimated preference vectors.

Multi-task learning. Several contributions assume that the bandit tasks share some underlying structure. In [34], task preference vectors are assumed to be sparse and to share their sparsity support, implying that they lie in a low-dimensional subspace with dimensions aligning with the canonical basis vectors. This idea is further generalized in [36], where the tasks are assumed to be confined to an arbitrary unknown low-dimensional subspace. That work improves upon [71] by not requiring the knowledge of the small dimension of the task space. It can be considered to solve our problem if the number of

clusters is smaller than the dimension, resulting in a low-rank structure. However, our work does not rely on any assumption between the number of clusters and the dimension. The underlying structure linking tasks can also be a graph encoding relations between them [38, 168], which is our case. However, while they assume smoothness as a prior, we assume piecewise constant behavior.

5.3 Problem setup

We consider a linear bandit setting, with a finite number of tasks representing users in a recommendation system for example. For each task the agent has to choose among K arms, each associated to a d-dimensional context vector. All interactions over a horizon of T time steps. We further assume that we have access to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ representing the tasks and edge set $\mathcal{E}$ encoding the relationships between them. We identify the vertex set $\mathcal{V}$ with the set of vertex indices $[\![|\mathcal{V}|]\!]$. Thus, we consider $\mathcal{E}$ to be a subset of $\mathcal{V}^2$, where every edge $(m,n) \in \mathcal{E}$ has weight $w_{mn} > 0$, with $m < n$. The tasks' preference vectors are denoted by $\{\boldsymbol{\theta}_m\}_{m\in\mathcal{V}} \subset \mathbb{R}^d$ verifying $\|\boldsymbol{\theta}_m\| \leq 1 \ \forall m \in \mathcal{V}$, which we concatenate as row vectors into matrix $\boldsymbol{\Theta} \in \mathbb{R}^{|\mathcal{V}|\times d}$. The latter represents a graph vector signal, assumed to be piecewise constant over $\mathcal{G}$.

At a round $t \in \mathbb{N}^\star$, a user $m(t) \in \mathcal{V}$ is selected uniformly at random and served an arm with context vector $\mathbf{x}(t)$ from a finite action set $\mathcal{A}(t) \subset \mathbb{R}^d$ with size K, depending on their estimated preference vector $\hat{\boldsymbol{\theta}}_{m(t)}(t) \in \mathbb{R}^d$. We assume the expected reward to be linear, with an additive, σ-sub-Gaussian noise conditionally on the past. Formally, denoting by $\mathcal{F}_0$ the trivial sigma-algebra, and for all $t \geq 1$, by $\mathcal{F}_t$ the sigma-algebra generated by history set $\{m(1), \mathbf{x}(1), y(1), \cdots, m(t), \mathbf{x}(t), y(t), m(t+1)\}$, the received reward $y(t)$ is given by $y(t) = \left\langle \boldsymbol{\theta}_{m(t)}, \mathbf{x}(t) \right\rangle + \eta(t)$, where $\eta(t)$ is $\mathcal{F}_t-$measurable and

$$\mathbb{E}\left[\eta(t)|\mathcal{F}_{t-1}\right] = 0, \qquad \mathbb{E}\left[\exp(s\eta(t))|\mathcal{F}_{t-1}\right] \leq \exp(\frac{1}{2}\sigma^2 s^2) \quad \forall t \geq 1, \forall s \in \mathbb{R}. \tag{5.1}$$

At the end of a round t, all preference vectors are updated into a new estimation $\hat{\boldsymbol{\Theta}}(t)$ while leveraging the structure of graph $\mathcal{G}$, formally by solving the following optimization problem:

$$\hat{\boldsymbol{\Theta}}(t) = \underset{\tilde{\boldsymbol{\Theta}}\in\mathbb{R}^{|\mathcal{V}|\times d}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \left(\left\langle \tilde{\boldsymbol{\theta}}_{m(\tau)}, \mathbf{x}(\tau) \right\rangle - y(\tau) \right)^2 + \alpha(t) \sum_{(m,n)\in\mathcal{E}} w_{mn} \left\| \tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_n \right\|, \tag{5.2}$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors. The performance of our policy is assessed by the expected regret over the T interaction rounds for all tasks:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \max_{\tilde{\mathbf{x}}\in\mathcal{A}(t)} \left\langle \boldsymbol{\theta}_{m(t)}, \tilde{\mathbf{x}} \right\rangle - \left\langle \boldsymbol{\theta}_{m(t)}, \mathbf{x}(t) \right\rangle \right]. \tag{5.3}$$

The Optimization problem in Equation (5.2) is an instance of the Network Lasso [65]. Several instances of the same type were studied by Jung *et al.* [76], Jung and Vesselinova [75], Jung [74]. The objective is characterized by its second term which, while being just the Laplacian regularization without squaring the norms, promotes a piecewise constant behavior rather than smoothness. For real-valued signals ($d = 1$), this regularization has been extensively studied for image and graph signal denoising, for the problem of trend filtering on graphs [162]. According to [162], that regularization better adapts to the heterogeneity of smoothness of the signal and induces a cluster structure in the data: similar users will not only have similar models but the same model, which offers a compression of the overall model over the graph. Note that our setting is cluster agnostic; our algorithm does not aim to learn the cluster structure explicitly but to exploit it implicitly using the total variation semi-norm as regularization. The latter's strength is controlled via a time-dependent regularization coefficient $\alpha(t)$, which we will express later in the analysis.

We formalize our assumption on the context generation as follows.

Assumption 1 (i.i.d action sets)**.** *Context sets $\{\mathcal{A}(t)\}_{t=1}^{T}$ are generated i.i.d. from a distribution p over $\mathbb{R}^{K\times d}$, such that $\|\mathbf{x}\| \leq 1 \forall\, \mathbf{x} \in \mathcal{A}(t)\ \forall t \geq 1$.*

In addition to the i.i.d assumption, we assume more regularity.

Assumption 2 (Relaxed symmetry and balanced covariance)**.** *There exists a constant $\nu \geq 1$ such that for all $\mathbf{X} \in \mathbb{R}^{K\times d}$, $p(-\mathbf{X}) \leq \nu p(\mathbf{X})$. Furthermore, there exists $\omega > 0$, such that for any permutation $(a_1, \cdots, a_K)$ of $[K]$, for any $i \in \{2, \cdots, K-1\}$, $\mathbf{w} \in \mathbb{R}^d$, we have*

$$\mathbb{E}\left[\mathbf{x}_{a_i}\mathbf{x}_{a_i}^{\top}[\mathbf{w}^{\top}\mathbf{x}_{a_1} < \cdots < \mathbf{w}^{\top}\mathbf{x}_{a_K}]\right] \preccurlyeq \omega\mathbb{E}\left[(\mathbf{x}_{a_1}\mathbf{x}_{a_1}^{\top} + \mathbf{x}_{a_K}\mathbf{x}_{a_K}^{\top})[\mathbf{w}^{\top}\mathbf{x}_{a_1} < \cdots < \mathbf{w}^{\top}\mathbf{x}_{a_K}]\right],$$

where $\mathbf{M} \preccurlyeq \mathbf{N}$ means that $\mathbf{N} - \mathbf{M}$ is a PSD matrix.

This assumption was introduced in [119], and has already been used in a multi-task setting by Cella *et al.* [36]. Parameter ν controls the skewness, as $\nu = 1$ corresponds to a symmetric distribution. ω decreases with increasing positive correlation between arms. It verifies $\omega = O(1)$ for multi-variate Gaussians and uniform distributions over the unit sphere [119]. The piecewise constant behavior of the graph signal $\boldsymbol{\Theta}$ is formalized in the next assumption.

Assumption 3 (Piecewise constant signal)**.** *There exists a partition $\mathcal{P}$ of $\mathcal{V}$, such that for any cluster $\mathcal{C} \in \mathcal{P}$, signal $\boldsymbol{\Theta}$ is constant on $\mathcal{C}$, and the graph obtained by taking the vertices in $\mathcal{C}$ and the edges linking them is connected.*

Assumption 3 basically states that the true preference vectors are clustered and that the given graph induces the cluster structure. It is required for our approach to be beneficial, as we will detail in the analysis section. For the sake of clarity, we defer the statement of other technical assumptions to Section 5.5.

5.4 Decision-making strategy

Our policy in Algorithm 4 follows a greedy arm selection rule in a multi-task setting, in the same vein as those presented in [119, 36]. Indeed, as pointed out in [119], exploration is implicitly incorporated into regularization parameter $\alpha(t)$'s time dependence. It has the following expression

$$\alpha(t) := \frac{\alpha_0 \sigma}{t} \sqrt{t + \sqrt{2 \sum_{m \in \mathcal{V}} \left|\mathcal{T}_m(t)\right|^2 \log \frac{1}{\delta(t)}} + 2 \max_{m \in \mathcal{V}} \left|\mathcal{T}_m(t)\right| \log \frac{1}{\delta(t)}}, \tag{5.4}$$

where the set of time steps a task m has been selected up to time t is denoted by $\mathcal{T}_m(t)$. At each time step the network Lasso problem is solved via the primal-dual algorithm [74].

Input $T, \alpha_0 > 0, \mathcal{G}$, function δ
Initialization : $\hat{\mathbf{\Theta}}(0) = \mathbf{0} \in \mathbb{R}^{|\mathcal{V}| \times d}$
for $t \in [1, T]$ **do**
  Draw a user $m(t) \in \mathcal{V}$ uniformly at random.
  Observe context set $\mathcal{A}(t)$.
  Select $\mathbf{x}(t) \in \arg\max_{\tilde{\mathbf{x}} \in \mathcal{A}(t)} \left\langle \hat{\boldsymbol{\theta}}_{m(t-1)}, \tilde{\mathbf{x}} \right\rangle$, breaking ties arbitrarily.
  Receive payoff $y(t)$
  Update $\alpha(t)$ via Equation (5.4)
  Update $\hat{\mathbf{\Theta}}(t)$ via solving the network Lasso problem Equation (5.2)
end for

4: Network Lasso Policy

5.5 Theoretical analysis

This section provides the main steps of the analysis. One of the paper's contribution lies in finding an oracle inequality of the network lasso problem given a restricted eigenvalue condition holding for the true multi-task Gram matrix. In this regard, the next major challenge and contribution is to show that the empirical multi-task Gram matrix, estimated in the algorithm, satisfies the restricted eigenvalue condition. We start by proving an oracle inequality for the estimation error of $\mathbf{\Theta}$. Then, we prove that the latter assumption holds with high probability given that the true multi-task Gram matrix satisfies it. We end this section by establishing a regret bound for our algorithm.

5.5.1 Notations and technical assumptions

We provide additional notations required for the analysis. We denote by $\partial\mathcal{P}$ the set of all edges in $\mathcal{E}$ connecting vertices from different clusters from partition $\mathcal{P}$ (Assumption 3), and we call it the boundary of $\mathcal{P}$. Thus, $\partial\mathcal{P}^c$, the complementary set of $\partial\mathcal{P}$, is formed by edges connecting vertices of the same cluster. The total weight of the boundary, *i.e.* the sum of its edges' weights, is referred to as $w(\partial\mathcal{P})$. Given a signal $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}$, we denote by $\overline{\mathbf{Z}}_{\mathcal{P}}$ the signal obtained by setting row vectors of $\mathbf{Z}$ to their mean-per-cluster value w.r.t. $\mathcal{P}$. For any edge subset $I \subseteq \mathcal{E}$, we denote the following norms: $\|\cdot\|_F$ as the Frobenius norm and $\|\boldsymbol{\Theta}\|_I \coloneqq \sum_{(m,n)\in I} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\|$ as the total variation semi-norm of $\boldsymbol{\Theta} \in \mathbb{R}^{|\mathcal{V}|\times d}$ over I. Thus, the regularization term of the problem in Equation (5.2) is equal to $\|\boldsymbol{\Theta}\|_{\mathcal{E}}$. Also, we define the incidence matrix $\mathbf{B}_I \subset \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$restricted to $I \subseteq \mathcal{E}$ to be null except at rows with index $i \in I$ corresponding to edge (m,n), where it equals $w_{mn}(\mathbf{e}_m - \mathbf{e}_n)$, where $\mathbf{e}_m$ is the m^{th} canonical basis vector of $\mathbb{R}^{|\mathcal{V}|}$. We define $\mathbf{A}_{\mathcal{V}}(t) \coloneqq \operatorname{diag}\left(\mathbf{X}_1(t)^\top\mathbf{X}_1(t), \ldots, \mathbf{X}_{|\mathcal{V}|}(t)^\top\mathbf{X}_{|\mathcal{V}|}(t)\right) \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$, and subsequently the empirical multi-task Gram matrix up to time step t is given by $\frac{1}{t}\mathbf{A}_{\mathcal{V}}(t)$. The following definition introduces quantities related to the clusters defined by partition $\mathcal{P}$, with crucial roles that we will elucidate throughout the analysis.

Definition 1 (Cluster content constants)**.** *Let $\mathcal{C} \in \mathcal{P}$ be a cluster.*

- *We denote by $\partial_v\mathcal{C}$ the inner boundary of $\mathcal{C}$,* i.e. *the vertices of $\mathcal{C}$ that are connected to its complementary. We define the inner isoperimetric ratio of $\mathcal{C}$ as $\iota_{\mathcal{G}}(\mathcal{C}) \coloneqq \frac{|\partial_v\mathcal{C}|}{|\mathcal{C}|}$.*
- *By abuse of notation, we denote as $\mathbf{B}_{\mathcal{C}}$ the incidence matrix restricted to edges linking vertices of $\mathcal{C}$, its associated Laplacian matrix by $\mathbf{L}_{\mathcal{C}} \coloneqq \mathbf{B}_{\mathcal{C}}^\top\mathbf{B}_{\mathcal{C}}$, and its pseudo-inverse by $\mathbf{L}_{\mathcal{C}}^\dagger$. The topological centrality index of node $m \in \mathcal{C}$ w.r.t $\mathcal{C}$ is equal to $(\mathbf{L}_{\mathcal{C}}^\dagger)_{mm}^{-1}$. We define the topological centrality index of $\mathcal{C}$ by $c_{\mathcal{G}}(\mathcal{C}) \coloneqq \min_{m\in\mathcal{C}}(\mathbf{L}_{\mathcal{C}}^\dagger)_{mm}^{-1}$.*

The inner isoperimetric ratio of a cluster measures how many "interior" nodes a cluster contains, in the sense that they are not connected to its complementary. It is at most equal to the isoperimetric ratio for weightless graphs as the size of the inner boundary is at most equal to that of the edge boundary, the latter being connected to the algebraic connectivity via the Cheeger inequality [39].

The topological centrality index measures the overall connectedness of a vertex in a network and indicates how robust a node is to edge failures [124]. Also, it can be tied to electricity spreading in a network according to [154]. We refer the interested reader to the two previously mentioned works for a detailed account of the properties of the topological centrality index. In the appendix, we show that for binary weights graphs the minimum topological centrality index is at least equal to the algebraic connectivity theoretically and experimentally, where we showcase that the difference between the two can be significant.

To proceed, we will need the following definition that introduces several notations to reduce the clutter.

Definition 2 (Restricted Eigenvalue (RE) condition and norm)**.** *A PSD matrix* $\mathbf{M} \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$ *verifies the RE condition with constants* $\kappa \geq 1$ *and* $\phi > 0$ *if*

$$\phi^2 \|\mathbf{Z}\|_{\mathrm{RE}} \leq \mathrm{vec}(\mathbf{Z}^\top)^\top \mathbf{M}\, \mathrm{vec}(\mathbf{Z}^\top) \quad \forall \mathbf{Z} \in \mathcal{S}, \tag{5.5}$$

where $\mathcal{S}$ *is the cone defined by:*

$$\mathcal{S} := \{\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}; a_1(\mathcal{G},\mathbf{\Theta})\|\mathbf{Z}\|_{\partial\mathcal{P}^c} \leq a_2(\mathcal{G},\mathbf{\Theta})\left\|\overline{\mathbf{Z}}_{\mathcal{P}}\right\|_F\},$$
$$a_1(\mathcal{G},\mathbf{\Theta}) := 1 - \frac{\frac{1}{\alpha_0} + 2\kappa w(\partial\mathcal{P})}{\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}}, \quad a_2(\mathcal{G},\mathbf{\Theta}) := \frac{1}{\alpha_0} + \sqrt{2}\kappa w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})},$$

and the RE semi-norm is defined by $\|\mathbf{Z}\|_{\mathrm{RE}} := \left\|\overline{\mathbf{Z}}_{\mathcal{P}}\right\|_F$.

For our main results, we cover the case of $\kappa \geq 1$ but treat the more general case $\kappa > 0$ in the proofs in the supplementary material. For such a simplification to be valid, we assume that $\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} > 2w(\partial\mathcal{P})$.

To explain the RE condition, if we had $\mathcal{S} = \mathbb{R}^{|\mathcal{V}|\times d}$ and $\|\cdot\|_{\mathrm{RE}} = \|\cdot\|_F$, then $\mathbf{M}$ would be invertible with minimum eigenvalue at least ϕ^2. In comparison, our requirement is weaker since it holds only for signals $\mathbf{Z} \in \mathcal{S}$ and for the $\|\cdot\|_{\mathrm{RE}}$ semi-norm. It has the same form as the compatibility assumption for the Lasso problem in [29, 119] or the restricted strong convexity assumption [36].

We further make the following assumption on the true multi-task Gram matrix:

Assumption 4 (RE condition for the true multi-task Gram matrix)**.** *For* $k \in [K]$*, let* $\mathbf{\Sigma}_k := \mathbb{E}\left[\mathbf{x}_k\mathbf{x}_k^\top\right]$ *be the Gram matrix of the* k^{th} *context vector's marginal distribution, let* $\mathbf{\Sigma}_{\mathcal{V}}$ *be the true multi-task Gram matrix of the context vector generating distribution, given by*

$$\mathbf{\Sigma}_{\mathcal{V}} := \mathbf{I}_{|\mathcal{V}|} \otimes \overline{\mathbf{\Sigma}}, \quad \textit{where} \quad \overline{\mathbf{\Sigma}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{\Sigma}_k. \tag{5.6}$$

We assume that $\mathbf{\Sigma}_{\mathcal{V}}$ *verifies RE condition (Definition 2) with some problem dependent constants* $\kappa \in \left[1, \frac{1}{2w(\partial\mathcal{P})} \min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}\right)$ *and* $\phi > 0$.

This assumption is common to several Lasso-like bandit problems [119, 6, 36]. We will later show that it can be transferred to the empirical multi-task Gram matrix.

5.5.2 Oracle inequality

This section is dedicated to provide a bound on the estimation error of the Network Lasso problem given in Equation (5.2) at a particular step t of Algorithm 4. We assume fixed design, meaning that the context vectors are given and fixed, and we are not concerned by their randomness (due to the context generating distribution), nor by the randomness of their number for each user (due to random selection at each time step).

For a time step t, we deliver the oracle inequality controlling the deviation between the estimated preference vectors $\hat{\mathbf{\Theta}}(t)$ and the true ones $\mathbf{\Theta}$.

Theorem 3 (Oracle inequality)**.** *Assume that the RE assumption holds for the empirical multi-task Gram matrix* $\frac{1}{t}\mathbf{A}_{\mathcal{V}}(t)$ *with constants* $\kappa \in \left[1, \frac{1}{2w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}\right)$ *and* $\phi > 0$*. Suppose that* $\max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt$ *for some* $b > 0$*. Then, with a probability at least* $1-\delta(t)$*, we have*

$$\left\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}(t)\right\|_F \leq 2\frac{\sigma}{\phi^2\sqrt{t}} f(\mathcal{G},\mathbf{\Theta}) \sqrt{1 + 2b\sqrt{|\mathcal{V}|\log\frac{1}{\delta(t)}} + 2b\log\frac{1}{\delta(t)}},$$

where

$$f(\mathcal{G},\mathbf{\Theta}) := \alpha_0 a_2(\mathcal{G},\mathbf{\Theta}) \left(\frac{a_2(\mathcal{G},\mathbf{\Theta})}{a_1(\mathcal{G},\mathbf{\Theta}) \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} + 1\right).$$

The proof relies on decomposing the estimation error signal into a sum of two terms. The first term amounts to taking its mean per cluster, that is, every node within the same cluster is mapped to the mean estimation error of its cluster. The second term is proven to be related to the incidence matrices of each cluster. The probabilistic statement comes from a high probability bound on the Euclidean norm of an empirical vector process associated with our problem, using a generalization of the Hanson-Wright inequality to the subgaussian case [70, Theorem 2.1]. Compared to the bound of Jung [74, Theorem 1], we bound a norm of the estimation error rather than just the total variation semi-norm. Besides, due to the expressions of $a_1(\mathbf{\Theta},\mathcal{G})$ and $a_2(\mathbf{\Theta},\mathcal{G})$, the bound significantly decreases with the products $w(\partial\mathcal{P}) \min_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota(\mathcal{C})}$ and $w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-\frac{1}{2}}$, which are small enough for dense intra-cluster edge links and sparse inter-cluster ones. The bound on the oracle inequality clearly grows with κ, thus it is most beneficial if κ is close to 1.

5.5.3 RE condition for the empirical multi-task Gram matrix

To establish the oracle inequality, we assumed that the RE condition holds for the empirical multi-task Gram matrix. In this section, we prove that this holds with high probability.

To this end, we use the same strategy as in [119, 36]. We prove that on the one hand, the empirical multi-task Gram matrix inherits the RE condition from its adapted counterpart since it concentrates around it. On the other hand, we show that the adapted Gram matrix verifies the RE condition due to Assumptions 1, 2 and 4.

Theorem 4 (RE condition holding for the empirical multi-task Gram matrix)**.** *Under Assumptions 2 and 4, let $t \geq 1$, and let κ, ϕ be the constants from Assumption 4. Assume that $\max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt$. Then, for any $\gamma \in \left(0, \left(1+\frac{a_2(\mathcal{G},\boldsymbol{\Theta})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^{-2}\right)$, the empirical multi-task Gram matrix $\frac{1}{t}\mathbf{A}_{\mathcal{V}}(t)$ verifies the RE condition with constants κ and $\hat{\phi}$, where*

$$\hat{\phi} = \tilde{\phi}\sqrt{1-\gamma\left(1+\frac{a_2(\mathcal{G},\boldsymbol{\Theta})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^2}, \tag{5.7}$$

with a probability at least equal to $1-6d|\mathcal{V}|\exp(\frac{-3\gamma^2\tilde{\phi}^4(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C})\wedge\tilde{c}_{\mathcal{G}}(\mathcal{C})^2)t}{6b+2\sqrt{2}\gamma\tilde{\phi}^2})$,

where $\tilde{\phi} \coloneqq \frac{\phi}{\sqrt{2\nu\omega}}$ and $\tilde{c}_{\mathcal{G}}(\mathcal{C}) \coloneqq c_{\mathcal{G}}(\mathcal{C}) \wedge |\mathcal{C}| \quad \forall \mathcal{C} \in \mathcal{P}$.

The proof follows the same approach as in [119, 36]; we prove that the RE condition transfers from the true multi-task Gram matrix to its adapted counterpart $\mathbf{V}_{\mathcal{V}}(t)$, defined as follows:

$$\mathbf{V}_{\mathcal{V}}(t) = \operatorname{diag}\left(\mathbf{V}_1(t), \cdots, \mathbf{V}_{|\mathcal{V}|}(t)\right), \tag{5.8}$$

where

$$\mathbf{V}_m(t) = \frac{1}{t}\sum_{\tau\in\mathcal{T}_m(t)} \mathbb{E}\left[\mathbf{x}(\tau)\mathbf{x}(\tau)^\top | \mathcal{F}_{\tau-1}\right]. \tag{5.9}$$

This transfer relies on the work of Oh *et al.* [119, lemma 10]. The other step of the proof is showing that the empirical multi-task Gram matrix and $\mathbf{V}_{\mathcal{V}}(t)$ become close to each other with high probability after sufficiently many time steps, in the sense of a matrix norm induced by the RE semi-norm and the restriction to set $\mathcal{S}$ (Definition 2). The bound showcases a dependence on $\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) \wedge |\mathcal{C}|$, which is of the same order as $|\mathcal{C}|$ for a fully connected cluster with vertices $\mathcal{C}$. It is also clear that the probability of satisfying the RE condition increases with a higher minimum centrality of a cluster.

5.5.4 Regret bound

To bound the regret, we bound the expected instantaneous regret for each round $t \geq 1$. This bound relies on the oracle inequality holding and the RE condition being satisfied for the empirical Gram matrix, both with high probability. Thanks to Theorem 3 and Theorem 4, these two conditions are ensured.

Theorem 5. *Let the mean horizon per node be $\overline{T} = \frac{T}{|\mathcal{V}|}$. Under Assumptions 1 to 4, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

$$\mathcal{R}(\overline{T}) \leq \mathcal{O}\left(\frac{\nu\omega f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\phi^2}\left(\sqrt{|\mathcal{V}|}+\sqrt{\log(\overline{T}|\mathcal{V}|)}+\sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)\right.$$
$$\left.+\frac{1}{A}\log(d|\mathcal{V}|)+\sqrt{|\mathcal{V}|}\right).$$

$$\text{with } A = \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}}+\sqrt{2}\gamma} \text{ and } \gamma = \frac{1}{2}\left(1+\frac{a_2(\mathcal{G},\boldsymbol{\Theta})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^{-2}.$$

Our regret is mainly formed of two parts. The first one is the sublinear time-dependent term and represents the bulk of horizon dependence. Interestingly, it decreases as the topological centrality index grows with the graph size, which proves the importance of intra-cluster high connectivity.

The second significant term comes from ensuring the RE condition for the empirical multi-task Gram matrix, and can be interpreted as the number of time steps necessary for it to hold, as pointed out by Oh *et al.* [119]. It has a logarithmic dependence in the graph size and in the dimension, which is a characteristic of regret bound of the "lasso typ". Also noteworthy is that the regret grows with $\log(d)$ only in the time-independent term, making our policy useful in high-dimensional settings.

Corollary 2. *Let $h_1 \coloneqq \sqrt{2}\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}$, $h_2 \coloneqq \min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\kappa w(\partial\mathcal{P})$. If we set $\alpha_0 = \frac{1}{h_2}\left(1+\sqrt{1+\frac{h_2}{h_1}}\right)$ then $f(\boldsymbol{\Theta},\mathcal{G})$ is minimized. Assume further that*

$$\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})} \gg \kappa w(\partial\mathcal{P})$$

and that $\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})} \leq 1$. If $|\mathcal{V}| \gg \log T$ and $|\mathcal{V}| = \mathcal{O}(T)$, the expected regret can be simplified as follows:

$$\mathcal{R}(\overline{T}) = \mathcal{O}\left(\left(\frac{\nu\omega\left(1+\sqrt{\frac{\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}}\right)}{\phi^2}+1\right)\sqrt{\overline{T}|\mathcal{V}|}+\frac{1}{A}\log(d|\mathcal{V}|)\right)$$

The simplified bound in Corollary 2 exhibits the typical multi-task learning dependency $\sqrt{\overline{T}|\mathcal{V}|}$ rather than the independent task learning case $|\mathcal{V}|\sqrt{\overline{T}}$.

5.6 Experimental analysis

We compare our algorithm with $\alpha_0 = 1$ to several baselines of the literature. On the one hand, we consider baselines relying on a given graph, GOBLin [38] and GraphUCB [169] that use the Laplacian to smooth the preference vectors. On the other hand, we compare to clustering of bandits baselines, namely CLUB [59], SCLUB [97] and LOCB [16]. We provided CLUB with graph $\mathcal{G}$ rather than a fully connected graph for a fair comparison. We also include the trace norm bandit algorithm ([36]), which is relevant when the number of clusters is smaller than d (we explain this point in the appendix). As a sanity check, we compare to the independent task learning case with LinUCB (LinUcbITL) where each task is solved independently. The graph used is weightless and generated using a stochastic block model to ensure a cluster structure, where an edge is constructed with probability p within clusters and q between clusters.

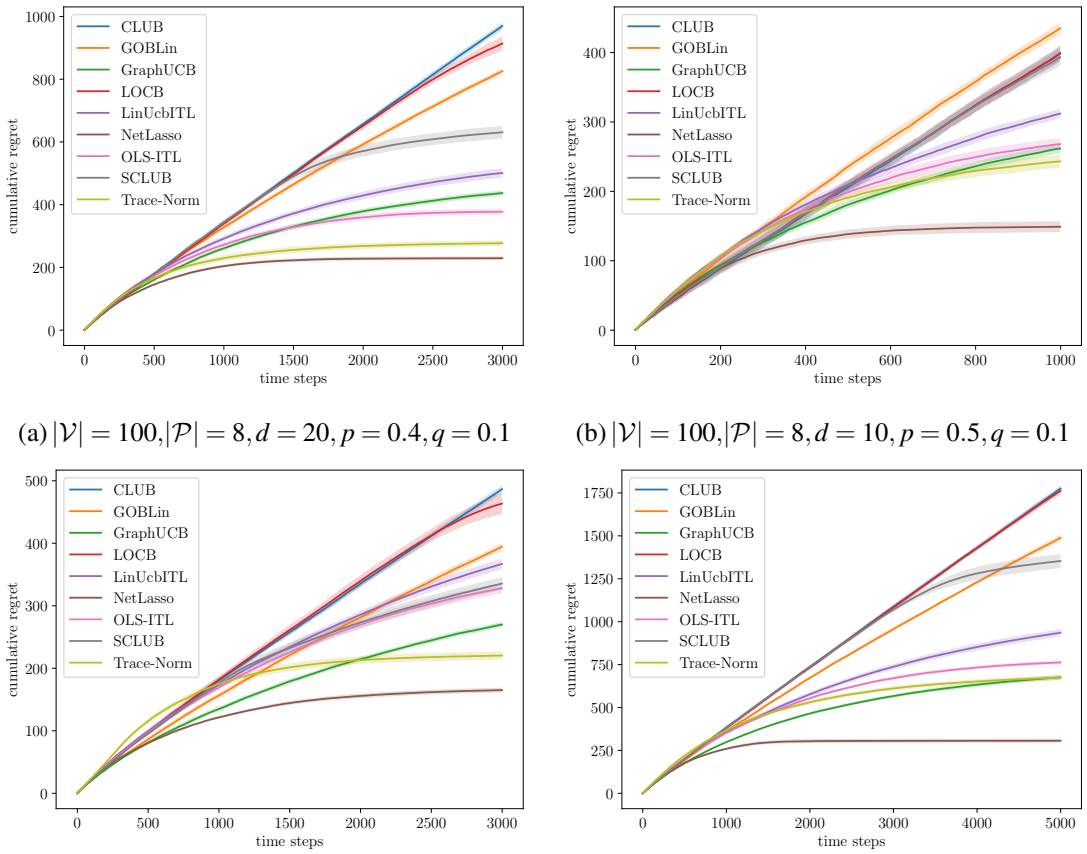
Experimentally, we found that normalizing the weights as $w_{mn} = (\deg(m)\deg(n))^{-\frac{1}{2}}$, where $\deg(m)$ denotes the degree of node m, yields significantly better results. Indeed, such a normalization makes the algorithm focus more on edges between low-degree nodes, which improves the propagation of the collected information within the graph.

Our results clearly showcase an improvement compared to the other baselines. Apart from the oracle that has complete knowledge of all clusters from the beginning, our policy performs significantly better than the rest beyond the error margins, covering one standard deviation at ten repetitions. We provide results for up to $|\mathcal{V}| = 200$ nodes showing the effective transfer of knowledge within the graph.

5.7 Conclusion

In this work, we proposed a multi-task bandit framework that solves the case where the task preference vectors are piecewise constant over a graph. To this end, we used the Network Lasso policy to estimate the task parameters, which bypasses explicit clustering procedures. We established a sublinear regret bound and proved a novel oracle inequality that relies on the small size of the boundary and the high value of the topological centrality index of each node within its cluster. Our experimental evaluations highlight the advantage of our method, especially when either the number of dimensions or nodes increases.

Due to the technical similarity of our problem with the Lasso, a natural extension would be to extend it to a thresholded approach, in the same vein as [6]. Another possible extension would be to use regularization with higher order total variation terms that



(a) $|\mathcal{V}| = 100, |\mathcal{P}| = 8, d = 20, p = 0.4, q = 0.1$

(b) $|\mathcal{V}| = 100, |\mathcal{P}| = 8, d = 10, p = 0.5, q = 0.1$

(c) $|\mathcal{V}| = 50, |\mathcal{P}| = 5, d = 80, p = 0.8, q = 0.2$

(d) $|\mathcal{V}| = 200, |\mathcal{P}| = 10, d = 20, p = 0.5, q = 0.05$

Figure 5.1: Synthetic data experiments showing the cumulative regret of Network Lasso Policy as a function of time-steps compared to other baselines, for different choices of $|\mathcal{V}|, |\mathcal{P}|, d, p$ and q.

impose a piecewise polynomial signal on a graph, as explained for scalar signals in Wang *et al.* [162], Ortelli and van de Geer [121].

Chapter 6

Online Influence Maximization with Semi-Bandit Feedback under Corruptions

In this work, we investigate the online influence maximization in social networks. Most prior research studies on online influence maximization assume that the nodes are fully cooperative and act according to their stochastically generated influence probabilities on others. In contrast, we study the online influence maximization problem in the presence of some corrupted nodes whose damaging effects diffuse throughout the network. We propose a novel bandit algorithm, CW-IMLinUCB, which robustly learns and finds the optimal seed set in the presence of corrupted users. Theoretical analyses establish that the regret performance of our proposed algorithm is better than the state-of-the-art online influence maximization algorithms. Extensive empirical evaluations on synthetic and real-world datasets also show the superior performance of our proposed algorithm.

6.1 Introduction

In the last decade, social networks have played critical roles in the analysis and optimization of data in epidemiology, marketing, and economics [53, 38, 152], as a result of information propagation or diffusion that is inherent in such networks. That has led to developing frameworks such as Influence Maximization (IM), where companies aim to select a fixed number of customers that greatly influence others, called seeds or source nodes, to receive reimbursement in return for advertising their products [164, 157, 137]. The companies aim at maximizing the influence spread given a limited budget.

Online social networks enable frequent information collection and updating of information regarding users' connections and interactions, which simplifies IM. In most solutions, a social network is modeled as a graph, and users as nodes. The edges represent the users' relations and the edge weights represent influence probabilities between users.

Influence propagates through the network under a specific diffusion model. The independent cascade (IC) model and linear threshold (LT) model are the two most widely used models [166, 98]. In the IC model, an adopter user has an activation probability, i.e., to convince each neighbor to adopt the product. The activation probabilities between pairs of users are independent. In the LT model, a user adopts the product only if the aggregated influence from its neighbors reaches a threshold [79]. The IC model is particularly well-known and frequently studied, especially in the context of online influence maximization [164, 166]. Therefore, in this paper, we focus on the IM problem within the framework of the IC model.

In the offline IM problem, the network structure and edge weights are known in advance [79]. However, in real applications, even if the network topology is accessible, the influence probabilities are unknown a priori. That highlights the importance of online influence maximization (OIM) problem [157, 164, 166, 92]. In OIM, the activation probability is unknown and needs to be estimated by a learner through directly interacting with the network.

The researchers have studied the OIM problem from many perspectives [164, 157, 166, 98, 176]; Nevertheless, they mostly assume that all users in the social networks are fully cooperative and influence others voluntarily and automatically, which ignores the adverse effects of potentially corrupted users/nodes as a critical factor. However, in real-world applications, malicious users trick the system with disputed behaviors. Even if not selected as seeds, they can spread corruption effects throughout the system by disrupting the information flow. Hence, it is imperative to develop an algorithm to address this challenge in influence maximization.

Before describing our contributions, we motivate our settings with several examples. Nowadays, customers heavily lean on online reviews to guide their purchasing decisions. These reviews extend beyond traditional online shopping platforms like Amazon. They also play a significant role in invisible marketing, where brands collaborate with influencers to integrate the products into their content. However, not all reviews are persuasive; sometimes, they can have a subtle counterproductive effect [102, 54, 51], e.g., when one uses humor or puns to subtly highlight a product's potential weaknesses, thus reducing the followers' enthusiasm to purchase that product. Another example is when the influencers adopt the comparative method for recommendation, which draws the customers' attention to similar products. Overemphasis of the products is another instance [31]. Such behaviors, malicious or not, impact the activation probabilities. In all these cases, most activation probabilities still follow a predictable pattern, whereas a fraction of them are corrupted under arbitrary patterns and are not identically distributed over time.

The proposed corruption-robust IM is not a straightforward extension of the previous work on corruption-robust bandit algorithms. While the concept of corruption-robust bandit algorithms is not new in the research of linear bandits, its application within the online influence maximization remains unexplored. Although each user in OIM can represent an arm in the bandit setting, it cannot be directly generalized to combinatorial

setting since the seeds are not selected in isolation as the users mutually affect each other according to the social network model. Besides, in OIM, the reward is not a linear function of the outcomes obtained from each selected seed and has a more complicated structure. It involves the cascading feedback model and limited feedback information (binary feedback). Furthermore, IM introduces a unique aspect where the impact of corruption can also propagate throughout the entire network, which makes the problem even more challenging. Additionally, the offline IM with given graph and activation probability information is an NP-hard problem.

In this work, we develop a novel OIM framework within a social network with several corrupted users to fill the gap of OIM problem under corruption. We summarize our main contributions as follows.

- We propose a novel algorithm for corruption-robust OIM, titled CW-IMLinUCB, which builds on an OIM algorithm with a corruption-robust linear bandit algorithm [67]. By integrating the weighted regression into an OIM algorithm, our proposal alleviates the problems arising from inaccurate estimations caused by corrupted users.

- We theoretically demonstrate that our proposed CW-IMLinUCB algorithm achieves the regret guarantee $O(dBE^*\sqrt{T}\log(nT) + BE^*E^cCd\log(nT))$ while being robust to malicious behaviors.

- Extensive experiments on synthetic and real-world datasets show the superior performance of our algorithm compared with the existing methods.

The rest of this paper is organized as follows: Section 6.2 summarizes the related work. In Section 6.3, we formulate the online influence maximization problem under corruption. Section 6.4 introduces our proposed algorithm CW-IMLinUCB. Section 6.5 presents the theoretical analysis of CW-IMLinUCB and Section 6.6 demonstrates the experimental results. Finally, Section 6.7 concludes this work.

6.2 Related work

Our work is related to IM and bandits with adversarial corruptions.

6.2.1 IM related work

Influence maximization is first investigated as an algorithmic problem in [125]. Reference [79] formulates the influence maximization problem as a discrete optimization problem and proves the problem to be NP-hard. A greedy approximation algorithm is proposed and shown to be effective for both IC and LT models. The efficiency of this greedy algorithm is further improved in [40]. Besides, reference [57] extends the IM

problem to competitive influence maximization across multiple social events. In [80], the proposed approach solves the IM problem by identifying community bridge nodes and select them as seed set. Aforementioned work considers the offline IM problem setting, where the network structure and edge weights are known in advance [79]. However, in realistic scenarios, even if the topology of the social network might be known, via Facebook, or Twitter, etc, the influence probabilities are unknown apriori. This highlights the importance of online influence maximization (OIM) problem [156, 157, 164, 166, 92].

The framework of OIM problem can be formulated as a variation of combinatorial multi-armed bandit (CMAB) problem, where the learning agent selects several base arms, defined as super arm at each round and tries to maximize its cumulative reward [42, 43]. References [42, 43] first use the CMAB framework to solve the OIM problem and develop an 'Upper Confidence Bound (UCB)'-like algorithm based on the IC model and analyze the regret bound. In [164], the authors consider the OIM problem with an independent cascade semi-bandit (ICSB) model. They propose the linear generalization model of the activation probability and prove regret bounds assuming edge-semi bandit feedback. Reference [157] suggests a different parameterization for the IM problem concerning pairwise reachability probabilities regardless of the underlying diffusion models. In [166], the authors factorize the activation probability on the edges into two latent factors. They use an IC model to estimate the influence parameters at the node level. There are a few works investigating OIM under different diffusion models. Reference [98] presents the OIM problem under the LT model. Wu *et al.* [167] address the non-stationarity in an evolving underlying social network whose nodes and edges change over time. Reference [176] introduce competitive concept into OIM problem and extend the classical IC model to multi-item diffusion model. Authors in [82] study the OIM problem under a decreasing cascade model, which is also a variation of IC model with consideration of market saturation. Similar to [164, 166, 176, 82], we use an IC model for influence propagation under edge-level feedback, while the corruption in diffusion is also considered.

6.2.2 Bandits with corruption related work

Reference [104] extends the classic stochastic multi-armed bandit problem by allowing for corrupted feedback and developing a decision-making strategy whose regret is proportional to the total corruption at each round. In [64], the authors propose an algorithm for a similar setting, whose regret is the summation of two terms: a corruption-independent term that matches the regret of the seminal multi-armed bandit algorithm and a time-independent term that is linear in the total corruption. For the corrupted stochastic linear bandit setting, Li *et al.* [100] present an algorithm with an instance-dependent regret bound. For the same problem, the algorithm in [25] achieves a regret with a corruption term that is linear in the total corruption. Zhao *et al.* [172] develop a variance-aware algorithm based on the 'Optimism in the Face of Uncertainty Linear bandit (OFUL)' algorithm [2]. Reference [67] also proposes a computationally efficient

algorithm based on OFUL, by incorporating a weighted ridge regression that prevents using the contexts whose rewards might be corrupted. Wang *et al.* [163] extend the work of [67] by considering the online clustering bandits problem with corrupted users. Besides, an algorithm is proposed to identify such users. In our work, we focus on the weighted ridge regression approach utilized in [67, 163].

6.3 Problem setup

In this section, we formulate an OIM problem under corruption with bandit feedback, illustrated in Figure 6.1. In such a problem, the final corruption effect depends on the position of the corrupted users. Indeed, a higher probability of activating corrupted users increases the corruption level in the system. Sometimes, even a tiny perturbation by a corrupted node in some time intervals changes the seed set entirely, thereby failing a corruption-agnostic agent. That aspect differs fundamentally from previous work on corruption [67, 163].

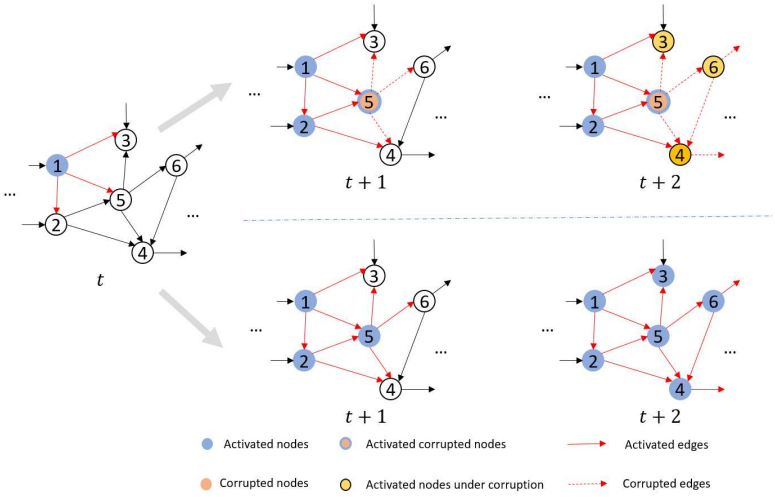


Figure 6.1: Online influence maximization under corruption.
Edges between users represent potential pathways for influence propagation. At t, when User 1 is activated, all of its out-edges trigger the activation of connected users. However, if User 5 is a corrupted user with unpredictable behavior, the outcomes at $t+1$ can be different. With some unknown probability, User 5 may behave normally (depicted in blue) or act adversarially (depicted in orange). User 5's corrupted behavior does not only perturb the influence diffusion when one selects it as a seed but also interferes with the influence diffusion when Users 1 or 2 are seeds, as user 5 lies within their diffusion pathways.

6.3.1 Notations

We use boldface lowercase letters and boldface uppercase letters to represent vectors and matrices respectively. For example, $\|\boldsymbol{x}\|_p$ denotes the $p-$norm of a vector $\boldsymbol{x}$. For a symmetric positive semi-definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, the weighted $2-$norm of vector $\boldsymbol{x} \in \mathbb{R}^d$ is defined by $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}$. The inner product is denoted by $\langle\cdot,\cdot\rangle$ and the weighted inner-product $\langle \boldsymbol{x},\boldsymbol{y}\rangle_{\boldsymbol{A}} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}$. Table 6.1 summarizes the definitions and notations.

Table 6.1: Notation

Problem-specific notations	
n	Number of users
m	Number of edges
$\mathcal{G}$	Graph that models the social network
$\mathcal{V}$	User set
$\mathcal{E}$	Edge set
$\mathcal{S}_t$	Seed set selected at t
K	Budget of seed set
$p(e)$	Activation probability of edge e
d	Dimension of feature vectors
T	Total number of rounds
$\boldsymbol{x}_e$	Feature vector of edge e
$\boldsymbol{\theta}$	Unknown feature vector
$c_{u,t}$	corruption level of node u at t
C_u	Total corruption budget of node u
C	The maximum corruption budget
$\omega_{e,t}$	Weight coefficient of edge e
$\boldsymbol{M}_t$	Gram matrix
$\boldsymbol{b}_t$	Vector that summarizes the past propagations
$\hat{\boldsymbol{\theta}}$	Estimation of unknown feature vector $\boldsymbol{\theta}$

6.3.2 Influence maximization

In the influence maximization (IM) problem, a directed graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ is utilized to model the social network. $\mathcal{V} = \{1,2,\ldots,n\}$ is the set of users (nodes) and $\mathcal{E}$ is the set of edges with cardinality $m = |\mathcal{E}|$. Each edge $e \in \mathcal{E}$ is associated with an activation probability $p(e) \in [0,1]$. For example, an edge $e = (u,v)$ represents that user v follows user u on some social media and $p(u,v)$ represents the probability that user v (receiving node) will be activated/influenced by user u (giving node). Denote $\boldsymbol{P} = (p(e_1),\ldots,p(e_m))$ to be

the activation probability vector. For a given seed set $\mathcal{S} \subseteq \mathcal{V}$ with activation probability $\boldsymbol{P}$, the expected number of influenced users under the diffusion model D is $f_{D,\boldsymbol{P}}(\mathcal{S})$. By definition, the users/nodes in $\mathcal{S}$ are always influenced.

Given $\mathcal{G}$ and a budget K on the number of seeds to be selected, IM aims to find a seed set that maximizes the influence spread. Formally,

$$\mathcal{S}^{\text{opt}} = \arg\max_{|\mathcal{S}| \leq K} f_{D,\boldsymbol{P}}(\mathcal{S}). \tag{6.1}$$

The IM problem is NP-hard [164, 157, 166], but approximation algorithms exist [41, 142]. In this paper, we refer to such algorithms as *oracles*, which take a graph, size of the seed set, and activation probabilities of all edges as inputs and output an appropriate set of seeds. Define $\mathcal{S}^{\text{opt}}$ as the optimal solution of the problem and $\mathcal{S}^* = \text{ORACLE}(\mathcal{G}, K, \boldsymbol{P})$ as the (possibly random) solution of an oracle ORACLE. It serves as an (α, γ)-approximation of $\mathcal{S}^{\text{opt}}$, $\alpha, \gamma \in [0,1]$, where $f_{D,\boldsymbol{P}}(\mathcal{S}^*) \geq \gamma f_{D,\boldsymbol{P}}(\mathcal{S}^{\text{opt}})$ with probability at least α [43]. This further implies that $\mathbb{E}[f_{D,\boldsymbol{P}}(\mathcal{S}^*)] \geq \alpha\gamma f_{D,\boldsymbol{P}}(\mathcal{S}^{\text{opt}})$. Besides, if $\alpha = \gamma = 1$, the oracle is exact.

The OIM problem is approachable within CMAB framework. In such a model, the users and any seed set represent the arms and a super arm, respectively. We assume that the expected influence spread (expected reward) satisfies the following assumptions, which are standard assumptions also in combinatorial bandit problems [161, 166].

Assumption 5 (Monotonicity)**.** *The expected reward of playing any super arm $\mathcal{S} \in \bar{\mathcal{S}}$ is monotonically non-decreasing with respect to the expectation vector, i.e., if for all $i \in [m]$, $p(e_i) \leq p'(e_i)$, we have $f_{D,\boldsymbol{P}}(\mathcal{S}) \leq f_{D,\boldsymbol{P}'}(\mathcal{S})$, for all $\mathcal{S} \in \bar{\mathcal{S}}$ with $\bar{\mathcal{S}}$ being the set of all candidate super arms.*

Assumption 6 (1-Norm Bounded Smoothness)**.** *A combinatorial multi-armed bandit with probabilisitically triggered arms (CMAB-T) satisfy 1-norm bounded smoothness, if there exists a bounded smoothness constant $B \in \mathbb{R}^+$ such that for any two distributions with expectation vectors $\boldsymbol{P}$ and $\boldsymbol{P}'$ and any action $\mathcal{S}$, we have $|f_{D,\boldsymbol{P}} - f_{D,\boldsymbol{P}'}| \leq B\sum_{i \in \tilde{S}} |p(e_i) - p'(e_i)|$, where $\tilde{S}$ is the set of edges (arms) that are triggered by $\mathcal{S}$.*

Remark 5. *For the OIM problems, the 1-Norm Bounded Smoothness (Assumption 6) holds with smoothness constant $B = \tilde{n}$, where $\tilde{n}$ is the largest number of nodes any node can reach in the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ [161].*

6.3.3 Online influence maximization under corruption

In real-world applications, the activation probability vector $\boldsymbol{P}$ is unknown and shall be learned via interaction with the network: At each round t, the learner/agent firstly chooses a seed set $\mathcal{S} \subseteq \mathcal{V}$ with cardinality K based on its prior information and past observations. It then uses the feedback from the observed influence spread to refine the

estimation of $\boldsymbol{P}$. The learner aims to maximize the influence spread through this repeated process. Multi-armed bandit framework, especially the linear bandit model, is widely used to solve the OIM problem [164, 157, 166].

Similar to [164], we assume each edge $e \in \mathcal{E}$ is associated with a known feature vector $\boldsymbol{x}_e \in \mathbb{R}^d$ and an unknown coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the dimension of the feature vector. Previous works assume that for all $e \in \mathcal{E}$, $p(e)$ is *well-approximated* by $\boldsymbol{x}_e^T \boldsymbol{\theta}$. We assume malicious users can occasionally corrupt the diffusion process to mislead the agent into selecting sub-optimal seed sets. At each round t, if user u is malicious, it can corrupt all the connected out edges with activation probabilities to its neighbors by $c_{u,t}$. Formally, the behavior of a corrupted user satisfies the following assumption.

Assumption 7. *For all $u, v \in \mathcal{V}$ with $e = (u,v) \in \mathcal{E}$, let $p(e)$ be the probability that user v can be activated by u at round t. For a normal user, the activation probability $p(e)$ of any out-edge e can always be well-approximated as*

$$p(e) = \boldsymbol{x}_e^T \boldsymbol{\theta}, \tag{6.2}$$

whereas for any corrupted user $u \in \mathcal{V}$, the activation probability of its out-edge e at t is given by

$$p_t(e) = \boldsymbol{x}_e^T \boldsymbol{\theta} + c_{u,t}. \tag{6.3}$$

In real-world applications, the activation probabilities of corrupted users' out-edges are often well-approximated by $\boldsymbol{x}_e^T \boldsymbol{\theta}$, similar to normal users; Nevertheless, a small fraction of them can be adversarially corrupted at time step t with level $c_{u,t}$. Therefore, since $c_{u,t}$ can become zero at some time intervals, learning the ground truth $p(e)$ is challenging. Similar to [164, 166], we assume an independent cascade diffusion model. In addition, below, we define the edge semi-bandit feedback.

Definition 3 (Edge semi-bandit feedback)**.** *In edge semi-bandit feedback, or edge level bandit feedback, the agent observes the influenced edge; That is, at any round t, the agent observes an edge $e = (u,v)$ if and only if its starting node u is activated.*

The performance measure for the learning algorithm is the *expected regret*, which is the difference between the optimal influence under perfect knowledge and the realized influence spread by the algorithm. Since computing the optimal seed set is NP-hard even under the perfect knowledge, similar to [43, 164, 157, 166, 176], we measure the performance of the algorithm by scaled cumulative regret defined as follows.

$$R^{\alpha\gamma}(T) = T \cdot f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma}\mathbb{E}[\sum_{t=1}^{T} f_{D,\boldsymbol{P}}(\mathcal{S}_t)], \tag{6.4}$$

where $\alpha\gamma \in (0,1)$.

We assume that the feature vector $\boldsymbol{x}$ and $\boldsymbol{\theta}$ satisfy the following assumption on the bandit model.

Assumption 8. *For any edge $e \in \mathcal{E}$, the feature vector $\boldsymbol{x}_e$ satisfies $\|\boldsymbol{x}_e\| \leq 1$. The unknown coefficient feature vector $\boldsymbol{\theta}$ satisfies $\|\boldsymbol{\theta}\| \leq \Theta$ where Θ is a constant. For the normalized feature vector, $\Theta = 1$.*

To measure the level of adversarial corruptions, we define the *corruption level (total corruption budget)* as $C_u = \sum_{t=1}^{T} |c_{u,t}|$ and $C = \max_{u \in \mathcal{V}} C_u$ is the maximum corruption level.

Remark 6. *The presence of corruption in activation probabilities does not affect the fulfillment of Assumptions 5 and 6. Assumptions 5 and 6 are originally proposed in [161] with a multi-armed bandit framework, where each edge is linked to an activation probability, without making any assumptions on how this probability is approximated using edge features. Consequently, they remain valid irrespective of the activation probability approximation model.*

Remark 7. *Our definition of corruption level is an extension of the definition in [25, 67]. We extend the original definition from one corrupted user setting to multiple corrupted users by considering the worst-case with the maximum corruption level.*

6.4 Decision-making strategy

In this section, we propose Confidence Weighted Influence Maximization Linear UCB (CW-IMLinUCB) algorithm, which robustly learns the activation probability over the directed graph from corrupted feedback. Algorithm 5 presents the pseudocode.

The inputs of CW-IMLinUCB are the network topology $\mathcal{G}$, the seed set cardinality K, the optimization algorithm ORACLE, the feature vectors $\boldsymbol{x}_e \in \mathbb{R}^d$, $\forall e \in \mathcal{E}$ and three algorithm hyper-parameters $\lambda, \sigma, \beta > 0$. The value of σ is proportional to the noise in the observations and hence controls the learning rate [157]. For each time step t, we define the Gram matrix $\boldsymbol{M}_t \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_t \in \mathbb{R}^d$ as the vector summarizing the past propagations. Besides, $\hat{\boldsymbol{\theta}}_t$ refers to the estimation of the unknown coefficient vector at time step t. $\boldsymbol{M}_t$ and $\boldsymbol{b}_t$ are sufficient statistics to compute $\hat{\boldsymbol{\theta}}_t$ and estimate the activation probability $p(e)$. The parameter β is utilized in forming the upper confidence bound (UCB) to consider the tradeoff between mean and variance, thus controls the *degree of optimism* of the algorithm [157].

At each time step t, CW-IMLinUCB firstly uses the estimated UCB of the activation probability from last time step to compute the seed set $\mathcal{S}_t$ based on the given optimization algorithm ORACLE (Line 4). Then the algorithm receives the edge semi-bandit feedback. $\tilde{\mathcal{E}}_t$ refers to the set including all the observed edges at time step t and $\boldsymbol{y}_t$ is an m-dimensional vector with $y_t(e_i) = y_t((u,v)) = \mathbb{1}\{v \text{ is activated via edge } e_i \text{ at time step } t\}$, $\forall i \in \{1, 2, \ldots, m\}$, which records the activation result. Afterwards, it updates $\boldsymbol{M}$, $\boldsymbol{b}$ and

Input: Graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, seed set cardinality K, oracle ORACLE, edge feature vector $\boldsymbol{x}_e$, $\forall e \in \mathcal{E}$, algorithm parameters $\lambda, \sigma, \beta > 0$.
Initialization

- $\boldsymbol{b}_0 = \mathbf{0} \in \mathbb{R}^d$ and $\boldsymbol{M}_0 = I \in \mathbb{R}^{d \times d}$;
- $\hat{\boldsymbol{\theta}} = \mathbf{0} \in \mathbb{R}^d$ and $\hat{p}_0(e) = 1$, for all $e \in \mathcal{E}$;

for $t = 1, 2, \cdots, T$ **do**
  Choose $\mathcal{S}_t \leftarrow \text{ORACLE}(\mathcal{G}, K, \hat{\boldsymbol{P}}_{t-1})$ where $\hat{\boldsymbol{P}}_{t-1} = \{\hat{p}_{t-1}(e)\}_{e \in \mathcal{E}}$
  Observe the edge level semi-bandit feedback $\boldsymbol{y}_t \in \mathbb{R}^m$
  for $e \in \mathcal{E}$ **do**
    if $e \in \tilde{\mathcal{E}}_t$ **then**
      {weighted regression}
      $\omega_{e,t} = \min\{1, \lambda / \|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}\}$
      $\boldsymbol{b}_t \leftarrow \boldsymbol{b}_{t-1} + \omega_{e,t} \boldsymbol{x}_e y_t(e)$
      $\boldsymbol{M}_t \leftarrow \boldsymbol{M}_{t-1} + \sigma^{-2} \omega_{e,t} \boldsymbol{x}_e \boldsymbol{x}_e^T$
    else
      $\boldsymbol{b}_t \leftarrow \boldsymbol{b}_{t-1}$
      $\boldsymbol{M}_t \leftarrow \boldsymbol{M}_{t-1}$
    end if
  end for
  $\hat{\boldsymbol{\theta}}_t \leftarrow \sigma^{-2} \boldsymbol{M}_t^{-1} \boldsymbol{b}_t$
  $\hat{p}_t(e) = \mathbb{P}_{[0,1]}(\hat{\boldsymbol{\theta}}_t^T \boldsymbol{x}_e + \beta \|\boldsymbol{x}_e\|_{\boldsymbol{M}_t^{-1}})$, for all $e \in \mathcal{E}$
end for

5: CW-IMLinUCB

the UCB of the activation probability for each edge, where $\mathbb{P}_{[0,1]}(\cdot)$ denotes the Euclidean projection onto the nearest point in the interval $[0, 1]$. The algorithm utilizes the updated activation probability estimation in the seed set selection of the next round.

Specially, different from previous works [156, 157, 166], which directly apply the classical OFUL algorithm with ridge regression [2] to estimate the unknown feature vector, our algorithm assigns each edge e a weight factor $\omega_{e,t}$. More precisely, the previous works estimate $\boldsymbol{\theta}$ by online ridge regression over all past observations, i.e.,

$$\boldsymbol{\theta}_t \leftarrow \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 + \sum_{\tau=1}^{t} \sum_{e \in \tilde{\mathcal{E}}_\tau} \sigma^{-2} (\boldsymbol{\theta}^T \boldsymbol{x}_e - y_\tau(e))^2.$$

However, in the presence of corruption, the previous algorithms that rely on the upper confidence bound parameter β without accounting for corruption [2, 164] will experience

a deterioration in regret performance, which will lead to a term $O(C\sqrt{T})$ in the regret, i.e., the regret bound is C times worse than the regret without corruption [67].

To overcome this difficulty, inspired by [67], we use the weighted ridge regression to estimate $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}}_t \leftarrow \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 + \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,\tau}\sigma^{-2}(\boldsymbol{\theta}^T \boldsymbol{x}_e - y_\tau(e))^2,$$

where its closed-form solution is $\hat{\boldsymbol{\theta}}_t = \sigma^{-2}\boldsymbol{M}_t^{-1}\boldsymbol{b}_t$, where

$$\boldsymbol{M}_t = I + \sigma^{-2} \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,\tau}\boldsymbol{x}_e\boldsymbol{x}_e^T,$$

and $\boldsymbol{b}_t = \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,\tau}\boldsymbol{x}_e y_\tau(e)$ (line 6 to line 16). We set the weight of sample at round t as $\omega_{e,t} = \min\{1, \frac{\lambda}{\|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}}\}$, where $\lambda > 0$ is a threshold coefficient to be determined later.

Remark 8. *The term $\|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}$ in line 8 of Algorithm 5 refers to the confidence radius. If $\|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}$ is large, CW-IMLinUCB assigns a small weight $\omega_{e,t}$ to avoid the potentially large regret caused by noise and adversarial corruption, while when $\|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}$ is small, it assigns a large weight $\omega_{e,t}$ (no more than 1) [67, 163]. Therefore, with carefully selected λ, our CW-IMLinUCB algorithm can get rid of the $O(C\sqrt{T})$ term caused by the corruption in the final regret compared to the OIM algorithm with original ridge regression [156].*

Remark 9. *CW-IMLinUCB has a storage complexity independent of t since only $\boldsymbol{M}_t$ and $\boldsymbol{b}_t$ need to be stored and updated. We need to emphasise that CW-IMLinUCB's computational efficiency relies heavily on the computational efficiency of* ORACLE. *Specifically, at each time step t, the computational complexity of CW-IMLinUCB from line 5 to line 17 is $O(md^2)$. Although updates of CW-IMLinUCB in line 8 and 16 involve matrix inversions, the operations do not incur high computational complexity since $\boldsymbol{M}_t$ only have sizes $d \times d$.*

6.5 Theoretical analysis

In this section, we derive a regret bound of CW-IMLinUCB under Assumptions 5-8. Notice that Assumptions 5-7 are standard for bandit analysis, and Assumption 8 can be satisfied by rescaling the feature vectors.

First we introduce the following definition.

Definition 4. *Assume that the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ includes l disconnected subgraphs $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$, …, $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$, which are in a descending order according to the number of*

nodes of each graph. E^* *is defined as the number of the edges containing in the first* $\min\{l,K\}$ *subgraphs [164],*

$$E^* = \sum_{i=1}^{\min\{l,K\}} |\mathcal{E}_i|, \tag{6.5}$$

and it is easy to obtain $E^* \leq |\mathcal{E}| = m$.

Furthermore, we introduce $\mathcal{D}_u(\mathcal{G}_i)$ *as the set containing all descendants of node* u *within graph* $\mathcal{G}_i$. *Let* $\mathcal{P}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}$ *denote the set containing all paths from node* u *to its descendant* $v \in \mathcal{D}_u(\mathcal{G}_i)$, *and* $\mathcal{E}_{u,d\in\mathcal{D}_u(\mathcal{G}_i)}$ *denote as the set collecting all edges within* $\mathcal{P}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}$. *In other words,* $\mathcal{E}_{u,d\in\mathcal{D}_u(\mathcal{G}_i)}$ *captures the edges forming paths to all descendants of node* u *within* $\mathcal{G}_i$. *For simplification, we call these edges as descendant edges.* E^c *is defined as*

$$E^c = \sum_{i=1}^{\min\{l,K\}} \max_u |\mathcal{E}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}|. \tag{6.6}$$

In words, E^c *is the summation of maximum count of descendant edges within the first* $\min\{l,K\}$ *subgraphs, with* $E^c \leq E^* \leq m$.

The following lemma defines the upper confidence bound parameter β.

Lemma 2. *For any* $0 < \delta < 1$ *and corruption budget* $C \geq 0$, *set the confidence radius* $\beta = \sigma^{-2}\sqrt{d\log(1+\frac{E^*T}{d}) + 2\log(\frac{1}{\delta})} + \sigma^{-2}\lambda E^c C + \Theta$ *then with probability at least* $1-\delta$, *for every round* t, *the good event* $\xi_{t-1} = \left\{ |\boldsymbol{x}_e^T(\hat{\boldsymbol{\theta}}_{\tau-1} - \boldsymbol{\theta})| \leq \beta\sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{\tau-1}^{-1}\boldsymbol{x}_e}, \forall e \in \mathcal{E}, \forall \tau \leq t \right\}$ *happens* $\forall t \in \{1,2,\ldots,T\}$. $E^c \leq E^* \leq |\mathcal{E}|$ *and the corresponding definitions are stated in Definition 4.*

Proof. See Appendix D.1. □

The following theorem states the regret bound of CW-IMLinUCB.

Theorem 6. *Assume that the activation probability of users satisfy Assumption 7. Besides,* ORACLE *is an* (α,γ)*-approximation algorithm. Let* Θ *be the known upper bound on* $\|\boldsymbol{\theta}\|$, $C \geq 0$ *is the corruption budget,* $\lambda = \frac{\sqrt{d}}{CE^c}$. *For any* $\sigma > 0$ *and any* $x_e \in \mathbb{R}^d$, $\forall e \in \mathcal{E}$, *if* β *satisfies*

$$\beta \geq \sigma^{-2}\sqrt{d\log(1+\frac{E^*T}{d}) + 2\log(nT)} + \sigma^{-2}\lambda E^c C + \Theta, \tag{6.7}$$

the $\alpha\gamma$*-scaled regret is upper bounded as*

$$R^{\alpha\gamma}(T) \leq O(dBE^*\sqrt{T}\log(nT) + BE^*E^c Cd\log(nT)). \tag{6.8}$$

Proof. See Appendix D.2. □

Remark 10. *If we consider the individual corruption budget C_u, by selecting*

$$\lambda = \frac{\sqrt{d}}{\sum_{i=1}^{\min\{l,K\}} \max_u |\mathcal{E}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}| C_u},$$

we can derive the tighter regret bound

$$O(dE^*\sqrt{T}\log(nT) + (\sum_{i=1}^{\min\{l.K\}} \max_u |\mathcal{E}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}|)C_u E^* d\log(nT)).$$

Definition 4 defines $\mathcal{E}_{u,v\in\mathcal{D}_u(\mathcal{G}_i)}$ and $\mathcal{D}_u(\mathcal{G}_i)$.

Remark 11. *The regret bound in Equation (6.8) is a topology-dependent bound. By Definition 4, E^c and E^* are less than m. Thus, the worst-case upper bound of the scaled regret yields $O(dmB\sqrt{T}\log(nT) + Bm^2Cd\log(nT))$.*

Remark 12. *When $C = 0$, i.e., no user is malicious, our setting reduces to the classic OIM problem. For unknown C, if a (potentially imprecise) estimation of it, namely, $\bar{C}$, is available, by selecting $\lambda = \sqrt{d}/(\bar{C}E^c)$, $\beta \geq \sigma^{-2}\sqrt{d\log(1+\frac{E^*T}{d}) + 2\log(nT)} + \sigma^{-2}\lambda E^c\bar{C} + \Theta$, we can distinguish the following cases:*

- *If $C \leq \bar{C}$, the scaled regret is bounded by*

$$O(dBE^*\sqrt{T}\log(nT) + BE^*E^c\bar{C}d\log(nT)).$$

- *If $C \geq \bar{C}$, the algorithm has a linear regret bound with respect to the time horizon, i.e., $O(T)$.*

In addition, if we set $\overline{C} = \sqrt{T}$, then when $0 < C \leq \sqrt{T}$, the regret is upper bounded by $O(dBE^*E^c\log(nT))$.

6.6 Experimental analysis

Unlike the previous work dealing with corruption concerning a single agent, our work delves into a unique aspect where the impact of corruption can propagate throughout the entire network in influence maximization. In this case, the number of corrupted users is not the sole determinant of the regret bound. Indeed, the placement of corrupted nodes within the network significantly influences the extent of regret. When a corrupted node occupies a pivotal position within the network, the resulting corruption effect can surpass that caused by numerous randomly selected nodes. In other words, even a few

number of corrupted users have the potential to disseminate the corruption effects across the entire network. In this section, we first highlight the importance of the position of the corrupted nodes in the network and compare the performance of our algorithm against benchmarks with a toy example. We evaluate CW-IMLinUCB on a carefully selected network topology (Figure 6.2) and validate our algorithm's performance under the dissemination of the corruption effects. Similar to previous work [156, 157, 83], we evaluate the performance of our algorithm using a randomly-generated synthetic- and real-world datasets. Besides, we compare the results to the following state-of-the-art bandit algorithms:

- **ε-greedy**: This algorithm learns the activation probability of each edge independently and uses ε-greedy [8] to balance exploitation and exploration.
- **CUCB** [161]: This algorithm learns the activation probability of each edge independently via a multi-armed bandit framework.
- **IMLinUCB** [164]: This algorithm learns under an edge-level bandit feedback and approximates the activation probability as the inner product of known edge features and one shared unknown feature among all the edges.
- **DILinUCB** [157]: This algorithm is model-dependent and approximates the activation probability as the inner product of the known target feature vector of the edge's ending node and weight vector of the edge's source node.
- **OIMLinUCB**: This algorithm is the variation of IMFB [166] and we assume the susceptibility vector $\boldsymbol{x}$ is known in advance, which is similar to DILinUCB and IMLinUCB.

Specially, for **DILinUCB** and **OIMLinUCB**, we also implement their variants with confidence weighted regression (**CW-DILinUCB** and **CW-OIMLinUCB**) to compare. Additionally, for all implemented algorithms, the DegreeDiscount algorithm [40] is used as the ORACLE. ALL the experimental results are the average of ten independent runs. In all plots, error bars indicate the standard deviations divided by $\sqrt{10}$.

6.6.1 Toy example

In the toy example experiment, we implement our algorithm in a ten-node network with a single corrupted user and select one seed. The network is an Erdős-Rényi graph and creates possible edges with probability 0.3. Figure 6.2 shows the network structure. We consider a variation of flip-$\boldsymbol{\theta}$ attack as the corrupted behavior of the corrupted users. Flip-θ attack simply flips the reward from $\boldsymbol{\theta}^T\boldsymbol{x}$ to $-\boldsymbol{\theta}^T\boldsymbol{x}$ [25, 67]. Considering activation probability acts as reward in online influence maximization, we add one constant to the reward and then make a flip-θ attack in our experimental setting. Thus, corrupted users trick the learning algorithm by changing the activation probability to become $p(e) =$

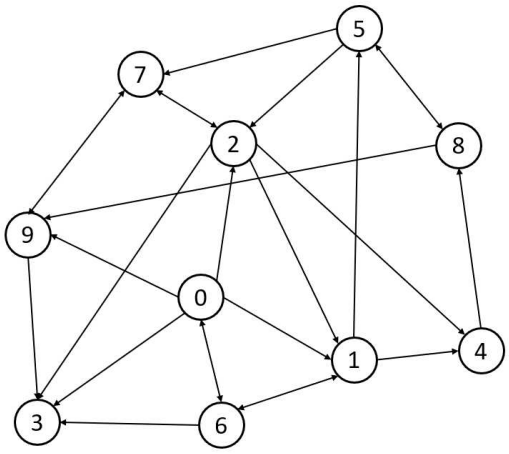


Figure 6.2: Network structure of Experiment I.

$\max(0, 0.05 - \boldsymbol{x}_e^T \boldsymbol{\theta})$ for the first $C_T = 100$ rounds. In the remaining rounds, the corrupted user acts normally. The activation probabilities in normal manner follow Equation (6.2) in Assumption 7. Each dimension of feature vectors $\boldsymbol{x}_e \in \mathbb{R}^{25}$, $\forall e \in \mathcal{E}$ and $\boldsymbol{\theta} \in \mathbb{R}^{25}$ is generated randomly from uniform distribution $U(0, 0.1)$ and then feature vectors are normalized. The average activation probability over edges is 0.175.

In Figure 6.2, Node 3 has no out-edges, whereas Node 1 has three in-edges and two out-edges. Therefore, the latter, if corrupted, can disseminate the corruption effect. Figure 6.3 shows the cumulative regret, where $\mathcal{U} = \{3\}$ and $\mathcal{U} = \{1\}$ in the legend indicate the corruption of Node 3 and Node 1, respectively.

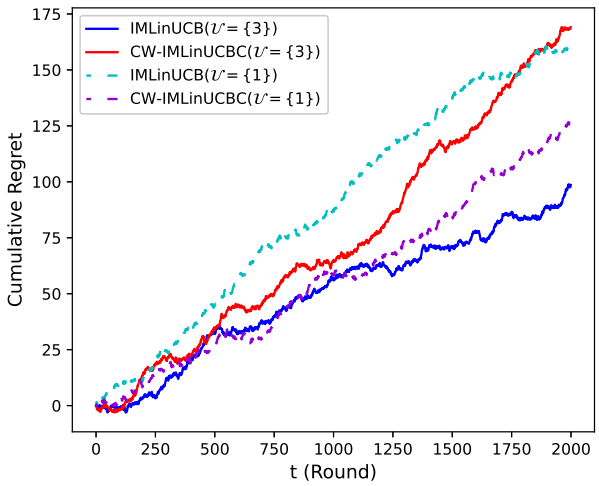


Figure 6.3: Effects of various corrupted user positions ($n = 10$).

According to Figure 6.3, when Node 3 is malicious, IMLinUCB performs better than CW-IMLinUCB. This is based on the fact that Node 3 cannot disseminate the corruption effect due to the network structure. Indeed, in this case, the extra term in the UCB of our corruption-robust algorithm adds an unnecessary exploration that degrades the performance w.r.t. the corruption-agnostic algorithm. In contrast, if Node 1 is corrupted, that term is vital, and our algorithm has a superior performance. More precisely, if Node 3 is malicious, the additional regret of our algorithm stems from the overestimation of corruption within the upper confidence bound. In β, the second term $\sigma^{-2}\lambda E^c C$ denotes the potential spread of corruption throughout the whole network with maximal corruption budget to account for the uncertainty. However, such a scenario occurs only when the crucial nodes in the most vital positions act in a corrupted manner permanently. Thus, the over-estimated corruption in β will increase exploration in the learning process. We emphasize that the selection of β rests on the assumption that the positions of corrupted users remain unknown. This assumption aligns with real-world scenarios with hidden corrupted users whose exact positions cannot be identified. By assuming unknown positions, the term $\sigma^{-2}\lambda E^c C$ in β remains indispensable.

6.6.2 Synthetic dataset

As described in Section 6.3, the positions of the corrupted users play a crucial role. To demonstrate this, we first implement our synthetic dataset on the network with $n = 50$ and $K = 2$.

The network is an Erdős-Rényi graph and creates possible edges with probability 0.3. It has in total $m = 687$ edges. The corrupted users trick the learning algorithm by changing the activation probability to become $p(e) = \max(0, 0.05 - \boldsymbol{x}_e^T \boldsymbol{\theta})$ for the first $C_T = 200$ rounds similar to the toy example. In the remaining rounds, the corrupted user acts normally. The activation probabilities in normal manner follow Equation (6.2) in Assumption 7. The generation of feature vectors $\boldsymbol{x}_e \in \mathbb{R}^{25}$, $\forall e \in \mathcal{E}$ and $\boldsymbol{\theta} \in \mathbb{R}^{25}$ follow the same setting as toy example. The average activation probability over edges is 0.0295.

We first apply IMLinUCB and CW-IMLinUCB algorithms to two different corrupted user sets, namely, $\mathcal{U} = \{0, 37\}$, selected at random, and $\mathcal{U} = \{34, 36\}$, both of which are directly connected to the optimal seed set $\{7, 43\}$ according to ORACLE.

Figure 6.4 shows the cumulative regret with different corrupted users. Intuitively, the positions of Nodes 34 and 36 are more crucial than Nodes 0 and 37, so the corruption in those positions would cause more adverse effects. In Figure 6.4, the performance of IMLinUCB verifies this point. When Nodes 34 and 36 are malicious, the regret of IMLinUCB is higher, and the selected seed set activates fewer nodes in the network compared to the situation when Nodes 34 and 36 are malicious. CW-IMLinUCB algorithm can overcome this difficulty. Thus, it outperforms IMLinUCB in both cases. In addition, in some rounds, the cumulative regret is below zero. That is because the algorithm calculates regret by performing the independent cascading process for both the optimal seed set and the seed set selected by our algorithm, then comparing the number of activated

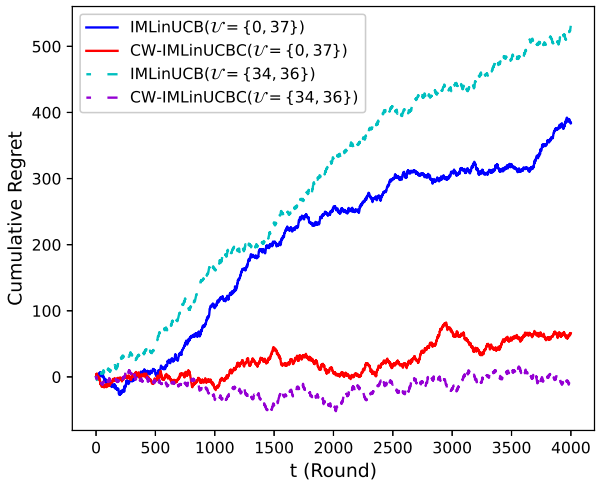


Figure 6.4: Effects of various corrupted user positions ($n = 50$).

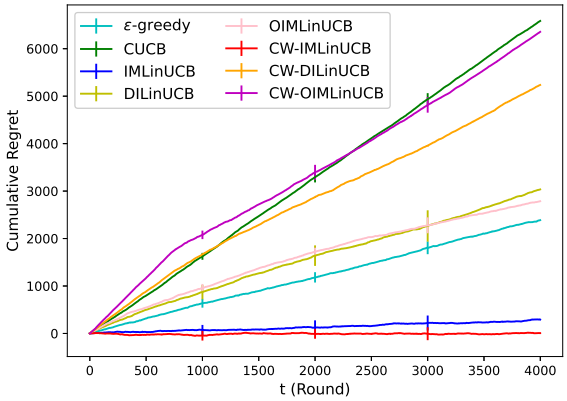
nodes at each time step. Due to the properties of independent cascading, the number of activated nodes can vary from round to round. Consequently, there is a possibility that the seed set selected by our algorithm might activate more nodes than the optimal seed set.

We also evaluate the performance of our proposed algorithm compared to the state-of-the-art algorithms under the same setting in previous experiment ($K = 2$, Nodes 0 and 37 are corrupted users). To guarantee that the activation probability is exactly $p(e = (u,v)) = \boldsymbol{x}_e^T \boldsymbol{\theta} = \boldsymbol{x}_v^T \boldsymbol{\theta}_u$ for all the implemented algorithms, we follow the setting in [166]: We first randomly sample $\boldsymbol{x}_v \in \mathbb{R}^{d_1}$ and $\boldsymbol{\theta}_u \in \mathbb{R}^{d_1}$ for DILinUCB and OIMLinUCB algorithms. Then, for each edge $e = (u,v) \in \mathcal{E}$, we take the outer product on $\boldsymbol{x}_v$ and $\boldsymbol{\theta}_u$ and reshape it into a $d = d_1 \times d_1$-dimensional vector, which is the edge feature vector $\boldsymbol{x}_e$ in the IMLinUCB algorithm with $d_1 = d_2 = 5$. Therefore, the IMLinUCB only needs to recognize the diagonal terms in the outer product. Figure 6.5a and 6.5b show the cumulative regret and the average reward, respectively.

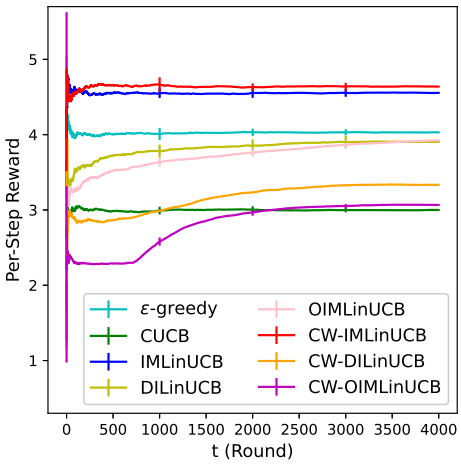
Figure 6.5 shows that our proposed algorithm has the lowest regret and the highest reward compared to all other methods. The algorithms CW-DILinUCB and CW-OIMLinUCB have higher regret compared to DILinUCB and OIMLinUCB. Although CW-DILinUCB and CW-OIMLinUCB integrate the weighted regression into the algorithm, their performance is not as good as expected.

6.6.3 Real-world application

We implement our algorithm in a subgraph of the Facebook network data [93]. The dataset has 4039 nodes and 88234 edges, and the subgraph includes the first $n = 300$



(a) Cumulative regret.



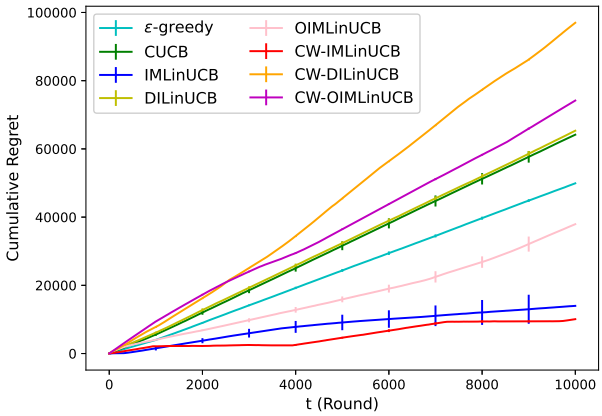
(b) Per-step reward.

Figure 6.5: Result of Experiment I.

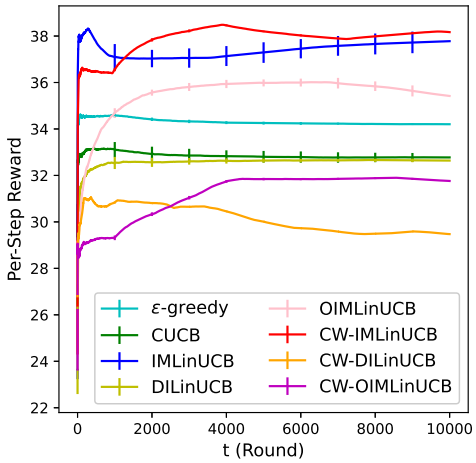
nodes and $|\mathcal{E}| = 2046$ edges. The average edge activation probability on this subgraph is 0.0497. Several authors use the Facebook dataset to evaluate the effectiveness of OIM algorithms [156, 157]. A challenge is the absence of information about edge activation probabilities, seed set sizes, and other parameters. To address that, we use the same approach in previous work [156], treating the sampled values as the ground truth. The data sampling process is as follows.

We first choose $K = 20$ for the seed set. There are $n_c = 20$ randomly-selected corrupted users in the network. The generation of feature vectors is the same as Experiment I in Section 6.6.2. The corrupted users also follow the previous setting with $C_T = 1000$. Figure 6.6a and 6.6b respectively show the cumulative regret and average reward, i.e.,

the number of per-step activated users.



(a) Cumulative regret.

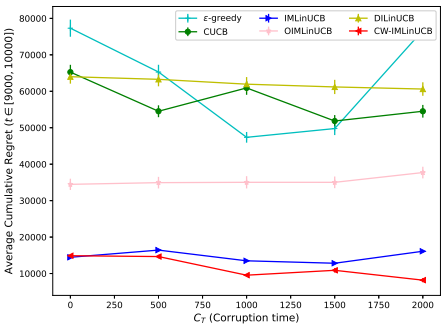


(b) Per-step reward.

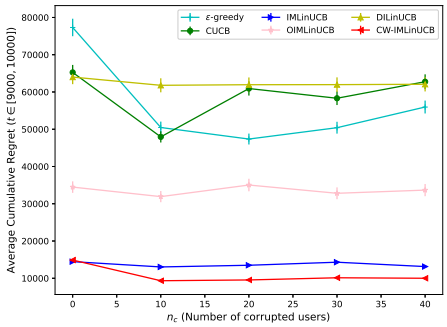
Figure 6.6: Result of Experiment II.

Moreover, we evaluate the performance of our proposed algorithm under different corruption levels. Similar to previous numerical simulations, the confidence weighted regression is not compatible to DILinUCB and OIMLinUCB algorithms, hence we omit the plot of CW-DILinUCB and CW-OIMLinUCB in the following experiments. Figure 6.7a shows the regret of all algorithms under different time horizons of the corruption C_T with a fixed set of $n_c = 20$ corrupted users. Figure 6.7b shows the performance of our algorithm when the number of corrupted users n_c changes while $C_T = 1000$ remains fixed. Particularly, when $n_c \geq 10$, the experiments share the same ten corrupted users. For each experiment, we add the randomly selected users to the previous corrupted set

of users. Both experiments have $K = 20$.



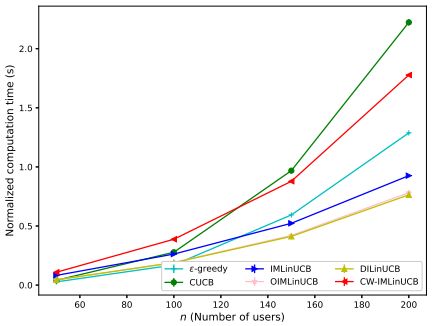
(a) Regret for different corruption time.



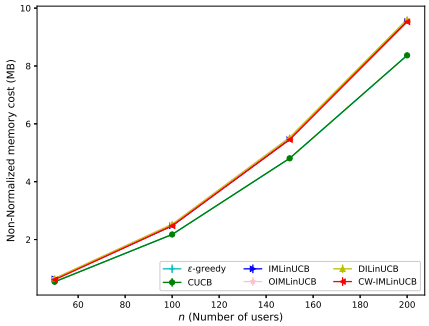
(b) Regret for different number of corrupted users

Figure 6.7: Result of different corruption levels.

In addition, we evaluate the time complexity and memory requirements of our proposed algorithm. In all experiments across various networks, we generated the network using the same process as in the experiments on the synthetic dataset. We conducted all the experiments with a horizon of $T = 5000$ and the corruption time $C_T = 200$. Figure 6.8 shows the results.



(a) Normalized computation time.



(b) Non-normalized memory cost

Figure 6.8: Complexity analysis under different networks.

From the experimental results, we conclude the followings:

- In all experiments, our proposed algorithm, CW-IMLinUCB, outperforms other methods.
- As Figure 6.6 shows, CW-IMLinUCB has several inflection points in the regret plot. That is because the algorithm uses the weighted regression, i.e., a small

weight for a large confidence radius, and vice versa. That prevents a potentially large regret caused by the corruption and consequently guarantees the superior performance of CW-IMLinUCB. Therefore, the abrupt change of the upper confidence bound caused by the weight change influences the seed set and reflects the inflection points in the plot.

- Compared to DILinUCB and OIMLinUCB, enhancing CW-DILinUCB and CW-OIMLinUCB with weighted ridge regression does not improve their performance compared to the vanilla version. The reason is that the weighted ridge regression framework is effective in dealing with the corruption effect under our proposed framework whereas its applicability to other frameworks in online influence maximization remains limited. Thus, our proposed structure is more compatible with IMLinUCB framework and exploits the weighted regression's strengths.

- According to Figure 6.7, when there is no corruption ($C_T = 0$ or $n_c = 0$), the performance of IMLinUCB is better than CW-IMLinUCB. The higher regret of CW-IMLinUCB here comes from the second $\sigma^{-2}\lambda E^c C$ of β. This term estimates the corruption effect within the upper confidence bound, however, in scenarios without corruption, it adds unnecessary exploration and degrades the performance of CW-IMLinUCB.

- According to Figure 6.7a, the benefits of CW-IMLinUCB become increasingly pronounced with longer corruption intervals when compared to state-of-the-art algorithms. The small fluctuation of the plot reflects the uncertainties in the diffusion process. This shows the robustness of CW-IMLinUCB algorithm in dealing with different corruption levels and its superiority compared to state-of-the-art algorithms.

- In Figure 6.7b, unlike Figure 6.7a, although CW-IMLinUCB has the lowest regret with different n_c, its regret does not significantly change among different experiments and just increased small amount with increased number of corrupted users. That is because the number of corrupted users is not the sole- or main determinant of the regret. Rather, it is the placement of corrupted nodes within the network that significantly influences the extent of regret. Merely increasing the number of corrupted nodes, without putting the corrupted node in influential positions, does not necessarily yield more corruption effect, thus the regret performance might remain steady.

- Compared with CW-IMLinUCB, there is no obvious trend in the regret plots of other algorithms in Figure 6.7. The reason is that other algorithms do not take corruption into account, which also increases uncertainty in the behavior of their learning processes in facing corruptions.

- As shown in Figure 6.8, the computation time and memory cost of CW-IMLinUCB increase with the network size, a trend observed across all state-of-the-art algorithms. The time cost of our proposed CW-IMLinUCB algorithm is slightly higher than that of other state-of-the-art algorithms. However, its memory cost is comparable to other algorithms with linear bandit structures, such as IMLinUCB, DILinUCB, and OIMLinUCB.

6.7 Conclusion

In this work, we study the OIM problem in presence of corrupted users. We propose an algorithm CW-IMLinUCB that integrates the weighted ridge regression into an OIM algorithm with bandit feedback. We present a theoretical guarantee for our proposed algorithm. Experiments on synthetic- and real-world datasets confirm the effectiveness and robustness of our proposed algorithm. According to our analysis, the position of the corrupted users can influence the regret bound. Currently, our regret bound only considers the worst-case that the corrupted users can easily disseminate the corruption effect across the network. In future work, one can consider to integrate corrupted user detection mechanisms into our framework and consider the corrupted user location dependent algorithm. Thus, the regret bound could be further tighter. Furthermore, developing an algorithm that competes with a dynamically corrupting algorithm is another interesting research direction.

Chapter 7

Thesis Conclusion

In this thesis, we made multiple efforts towards introducing novel frameworks and results in sequential decision making based on interconnected data. Our proposed frameworks are in the direction of stochastic stationary and nonstationary multi-armed bandit problems. In all these settings, we try to motivate the mathematical problem within the context of real-world applications. In addition, we present rigorous theoretical analysis of the provided solutions. In the following, we bring a brief recap of the contributions in each one of the Chapters 3, 4, 5, and 6. In addition, we also elaborate on the possible future works for the research that is described in each of these chapters.

7.1 Summary of contributions

In Chapter 3, we presented a completely novel combinatorial bandit setting with causally related rewards. The study incorporates the causal structure through a directed graph representation within a structural equation model (SEM). The proposed SEM-UCB algorithm effectively addresses the challenges posed by the causal dependencies among rewards by simultaneously learning the underlying graph structure and optimizing decision-making. Theoretical analysis confirmed a sublinear regret bound for SEM-UCB, ensuring its robustness in sequential decision-making problems. Empirical evaluations on both synthetic and real-world datasets, including the analysis of Covid-19 spread in Italy, demonstrated the superior performance of SEM-UCB compared to established benchmarks. The results highlight the framework's potential applications across diverse domains such as network data analysis in biology, financial markets, and epidemiology.

Chapter 4 presents a novel bandit framework in which the underlying structure of the data as well as the distributions of the base arms within a combinatorial bandit framework are subject to change. This allows for a more realistic modeling of the real-world problems and better applicability of bandit algorithms. The study presented a piecewise stationary combinatorial semi-bandit framework. The proposed PS-SEM-UCB-Gr

algorithm effectively adapts to changes in both the reward distributions and the underlying causal structures by employing a group restart strategy, change-point detection, and a piecewise static graph learning mechanism. The proposed group restart strategy balances the trade-off between minimizing unnecessary restarts and responding swiftly to detected changes, leveraging the relationships among base arms to improve decision-making efficiency. Theoretical analysis demonstrated a sublinear regret bound, reflecting the benefits of grouping strategies in structured environments. Also, we improved the theoretical analysis for the stationary setting that was studied in Chapter 3. Consequently, we found out that the asymptotic regret of the stationary setting does not depend on the size of the causal system. This is due to the fact that the agent is able to estimate the ground truth causal system structure from the initial exploration phase. Experimental results on synthetic and real-world datasets, showcased the algorithm's superior performance compared to existing benchmarks. The findings emphasize the importance of integrating structural relationships in bandit algorithms, particularly in dynamic environments.

The research in Chapter 5 introduces the Network Lasso Bandit framework, a significant advancement in multi-task contextual bandit problems where task preferences are represented as piecewise constant clusters on a graph. By leveraging the properties of the graph's structure and introducing the Network Lasso optimization, the proposed approach eliminates the need for explicit clustering. Instead, it directly incorporates a regularization term that enforces piecewise constant smoothness. Theoretical results, including novel oracle inequalities and sublinear regret bounds, demonstrate the framework's robustness and efficacy, especially in high-dimensional settings or when graph sizes are large. Empirical evaluations further validate the method's superiority over existing baselines.

The study in Chapter 6 addressed the challenging problem of Online Influence Maximization (OIM) in social networks, specifically under the presence of corrupted nodes that distort influence probabilities and disrupt information propagation. To tackle this, the CW-IMLinUCB algorithm was developed, which integrates weighted ridge regression into a contextual bandit framework, enhancing robustness against adversarial corruptions. Theoretical analysis demonstrated that the algorithm achieves a bounded regret under realistic assumptions, outperforming existing methods. Empirical evaluations on synthetic and real-world datasets validated the effectiveness and scalability of CW-IMLinUCB, showcasing its superior performance in diverse network configurations and corruption scenarios. The experiments emphasized the critical role of corrupted nodes' positions in determining their impact on network influence, underlining the algorithm's ability to adapt to varying corruption levels.

7.2 Future works

The performance of the algorithm in Chapter 3 relies on the initial exploration phase that serves to create proper data that allows for estimation and identifiability of the causal graph. Since the number of initial exploration rounds is equal to the number of base arms, if the number of base arms N grows such that it becomes comparable to the horizon T, then the regret of the agent will become closer to a linear regret. Hence, one possible future improvement for the setting in Chapter 3 is to improve the initial exploration phase. This indicates that one should study the possibility of establishing theoretical guarantees that allow for the identifiability of the adjacency matrix from a smaller number of rounds of initial exploration.

Another possible research focus can be to design another novel application for the provided framework of Chapter 3. Even though we believe that the Covid-19 data analysis is extremely interesting and inspiring, one could try to apply the provided framework in other realistic scenarios to show the applicability of the novel bandit setting. Furthermore, we designed both frameworks of Chapters 3 and 4 under the assumption of having a noise-free SEM. However, that is in reality just an approximation and one should study the same setting under the presence of noise. This makes the study more challenging since the theoretical guarantees for the identifiability of the causal system as well as for the upper bound of regret should be reconsidered. Moreover, another possibility for future research focus related to Chapter 3 is to assume to have non-linearities in the environment among the endogenous variables. In this case, the estimation process of the causal structure and its identifiability guarantees will be subject to change.

Regarding the possible future research directions for Chapter 5, considering the technical similarities between our problem setting and the Lasso framework, a logical direction for further exploration would be to adopt a thresholded methodology, similar to the approach presented by Ariu *et al.* [6]. Additionally, another promising avenue for extension involves integrating regularization with higher-order total variation terms. The goal would be to impose a piecewise polynomial structure on signals over a graph, as elaborated in the context of scalar signals by Wang *et al.* [162] and Ortelli and van de Geer [121].

As future works for the research in Chapter 6, one can consider to integrate mechanisms into our framework that try to detect the corrupted users and consequently design an algorithm that considers the detected corrupted-user's location in the decision making. Thus, the regret bound could be further tighter. Furthermore, developing an algorithm that competes with a dynamically corrupting algorithm is another interesting research direction. In this last case, the assumption is that the adversary is constantly changing the position of the corrupted nodes in the network.

Appendix A

Additional Material for Chapter 3

A.1 Proof of theorem 1

A.1.1 Notations

Before proceeding to the proof, in the following we introduce some important notations together with their definitions.

We define the *index set* of a decision vector $\mathbf{x} \in \mathcal{X}$ by $\mathcal{I}(\mathbf{x}) = \big\{i \mid \mathbf{x}[i] \neq 0, \forall i \in [N]\big\}$. The confidence bound of base arm i at time t is defined as $\mathbf{C}_t[i] = \sqrt{\frac{(s+1)\ln t}{\mathbf{m}_t[i]}}$. At each time t, we collect the empirical average of instantaneous rewards $\hat{\boldsymbol{\beta}}_t[i]$ and the calculated confidence bounds $\mathbf{C}_t[i]$ of all base arms $i \in [N]$ in vectors $\hat{\boldsymbol{\beta}}_t$ and $\mathbf{C}_t$, respectively. We have $\mathbf{E}_t = \hat{\boldsymbol{\beta}}_t + \mathbf{C}_t$. For ease of presentation, in the sequel, we use the following equivalence $\mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1}\text{diag}(\mathbf{E}_{t-1})\mathbf{x}_t = \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1}\text{diag}(\mathbf{x}_t)\mathbf{E}_{t-1}$. At each time t, we define the *selection index* for a decision vector $\mathbf{x} \in \mathcal{X}$ as $I_t(\mathbf{x}) = \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1}\text{diag}(\mathbf{x})\mathbf{E}_{t-1}$. To simplify the notation, sometimes we drop the time index t in $\mathbf{m}_t[i]$ and use $\mathbf{m}[i]$ to denote the number of times that the base arm i has been observed up to the current time instance.

For any $\mathbf{x} \in \mathcal{X}$, we use the counter $\mathcal{T}_\mathbf{x}(t)$ to represent the total number of times the decision vector $\mathbf{x}$ is selected up to time t. Finally, for each base arm $i \in [N]$, we define a counter $\mathcal{T}_i(t)$ which is updated as follows. At each time t after the initialization phase that a suboptimal decision vector $\mathbf{x}_t$ is selected, we have at least one base arm $i \in [N]$ such that $i = \underset{i \in \mathcal{I}(\mathbf{x}_t)}{\text{argmin}}\, \mathbf{m}_t[i]$. In this case, if the base arm i is unique, we increment $\mathcal{T}_i(t)$ by 1. If there are more than one such base arm, we break the tie and select one of them arbitrarily to increment its corresponding counter.

A.1.2 Auxiliary results

We use the following lemma in the proof of Theorem 1.

Lemma 1. [12] Let $z_1, z_2, \ldots, z_m$ be random variables and $z_i \in [0,1]$, $\forall i$. Moreover, $\mathbb{E}[z_t | z_1, \ldots, z_{t-1}] = \alpha$, for all $t = 1, \ldots, m$. Then, for all $D \geq 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} z_i - m\alpha\right| \geq D\right] \leq e^{-\frac{2D^2}{m}}. \tag{A.1}$$

A.1.3 Proof

We start by rewriting the expected regret as

$$\mathcal{R}_T(\mathcal{X}) = T\mu(\mathbf{x}^*) - \sum_{t=1}^{T} \mu(\mathbf{x}_t) = \sum_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})\mathbb{E}[\mathcal{T}_{\mathbf{x}}(T)].$$

Based on the definition of the counters $\mathcal{T}_i(t)$ for the base arms $i \in [N]$, at each time t that a suboptimal decision vector is selected, only one of such counters is incremented by 1. Thus, we have [56]

$$\mathbb{E}\left[\sum_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \mathcal{T}_{\mathbf{x}}(t)\right] = \mathbb{E}\left[\sum_{i=1}^{N} \mathcal{T}_i(t)\right], \tag{A.2}$$

which implies that

$$\sum_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \mathbb{E}\left[\mathcal{T}_{\mathbf{x}}(t)\right] = \sum_{i=1}^{N} \mathbb{E}\left[\mathcal{T}_i(t)\right]. \tag{A.3}$$

Therefore, we observe that

$$\mathcal{R}_T(\mathcal{X}) = \sum_{\mathbf{x}:\mu(\mathbf{x})<\mu(\mathbf{x}^*)} \Delta(\mathbf{x})\mathbb{E}[\mathcal{T}_{\mathbf{x}}(T)] \overset{(*)}{\leq} \Delta_{\max} \sum_{i=1}^{N} \mathbb{E}[\mathcal{T}_i(T)],$$

where $(*)$ follows from the definition of $\Delta_{\max}$.

Let $\mathbb{I}_i(t)$ denote the indicator function which is equal to 1 if $\mathcal{T}_i(t)$ is increased by 1 at time t, and is 0 otherwise. Therefore,

$$\mathcal{T}_i(T) = \sum_{t=N+1}^{T} \mathbb{1}\left\{\mathbb{I}_i(t) = 1\right\}. \tag{A.4}$$

If $\mathbb{I}_i(t) = 1$, it means that a suboptimal decision vector $\mathbf{x}_t$ is selected at time t. In this

case, $\mathbf{m}_t[i] = \min\left\{\mathbf{m}_t[j] | j \in \mathcal{I}(\mathbf{x}_t)\right\}$. Let $l = \left\lceil \frac{4(s+1)\ln T}{(\frac{\Delta_{\min}}{sw_{\max}})^2} \right\rceil$. Then,

$$
\begin{aligned}
\mathscr{T}_i(T) &= \sum_{t=N+1}^{T} \mathbb{1}\left\{\mathbb{I}_i(t) = 1\right\} \\
&\leq l + \sum_{t=N+1}^{T} \mathbb{1}\left\{\mathbb{I}_i(t) = 1 \ \&\ \mathscr{T}_i(t-1) \geq l\right\} \\
&\leq l + \sum_{t=N+1}^{T} \mathbb{1}\left\{I_t(\mathbf{x}^*) \leq I_t(\mathbf{x}_t) \ \&\ \mathscr{T}_i(t-1) \geq l\right\} \\
&= l + \sum_{t=N+1}^{T} \mathbf{1}\{\mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1}\mathrm{diag}(\mathbf{x}^*)\mathbf{E}_{t-1} \\
&\qquad \leq \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t-1})^{-1}\mathrm{diag}(\mathbf{x}_t)\mathbf{E}_{t-1} \ \&\ \mathscr{T}_i(t-1) \geq l\} \\
&= l + \sum_{t=N}^{T} \mathbf{1}\{\mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t})^{-1}\mathrm{diag}(\mathbf{x}^*)\mathbf{E}_{t} \\
&\qquad \leq \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_{t})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\mathbf{E}_{t} \ \&\ \mathscr{T}_i(t) \geq l\}.
\end{aligned} \tag{A.5}
$$

Based on the definition of $\mathscr{T}_i(t)$, we have $\mathscr{T}_i(t) \leq \mathbf{m}_t[i]$, $\forall i \in [N]$. Therefore, when $\mathscr{T}_i(t) \geq l$, the following holds [56].

$$
l \leq \mathscr{T}_i(t) \leq \mathbf{m}_t[j], \quad \forall j \in \mathcal{I}(\mathbf{x}_{t+1}). \tag{A.6}
$$

Let $\mathbf{v}_{t+1}^\top = \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_t)^{-1}\mathrm{diag}(\mathbf{x}^*)$ and $\mathbf{u}_{t+1}^\top = \mathbf{1}^\top(\mathbf{I} - \hat{\mathbf{A}}_t)^{-1}\mathrm{diag}(\mathbf{x}_{t+1})$. We order the elements in sets $\mathcal{I}(\mathbf{x}^*)$ and $\mathcal{I}(\mathbf{x}_{t+1})$ arbitrarily. In the following, our results are independent of the way we order these sets. Let v_k, $k = 1, \ldots, |\mathcal{I}(\mathbf{x}^*)| \leq s$, represent the kth element in $\mathcal{I}(\mathbf{x}^*)$ and u_k, $k = 1, \ldots, |\mathcal{I}(\mathbf{x}_{t+1})| \leq s$, represent the kth element in $\mathcal{I}(\mathbf{x}_{t+1})$.

Accordingly, we have

$$
\begin{aligned}
\mathscr{T}_i(T) &\leq l + \sum_{t=N}^{T} \mathbf{1}\left\{ \min_{0<\mathbf{m}[v_1],\ldots,\mathbf{m}[v_{|\mathcal{I}(\mathbf{x}^*)|}]\leq t} \sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|} \mathbf{v}_{t+1}^\top[v_j](\hat{\boldsymbol{\beta}}_t[v_j] + \mathbf{C}_t[v_j]) \leq \right. \\
&\left. \max_{l\leq\mathbf{m}[u_1],\ldots,\mathbf{m}[u_{|\mathcal{I}(\mathbf{x}_{t+1})|}]\leq t} \sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|} \mathbf{u}_{t+1}^\top[u_j](\hat{\boldsymbol{\beta}}_t[u_j] + \mathbf{C}_t[u_j]) \right\} \\
&\leq l + \sum_{t=1}^{\infty} \sum_{m_{v_1}=1}^{t} \cdots \sum_{m_{v_{|\mathcal{I}(\mathbf{x}^*)|}}=1}^{t} \sum_{m_{u_1}=l}^{t} \cdots \sum_{m_{u_{|\mathcal{I}(\mathbf{x}_{t+1})|}}=l}^{t}
\end{aligned}
$$

$$\mathbf{1}\left\{\sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|}\mathbf{v}_{t+1}^\top[v_j](\hat{\boldsymbol{\beta}}_t[v_j]+\mathbf{C}_t[v_j])\leq\sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|}\mathbf{u}_{t+1}^\top[u_j](\hat{\boldsymbol{\beta}}_t[u_j]+\mathbf{C}_t[u_j])\right\}.$$

We define the Event $\mathcal{P}$ as

$$\sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|}\mathbf{v}_{t+1}^\top[v_j](\hat{\boldsymbol{\beta}}_t[v_j]+\mathbf{C}_t[v_j])\leq\sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|}\mathbf{u}_{t+1}^\top[u_j](\hat{\boldsymbol{\beta}}_t[u_j]+\mathbf{C}_t[u_j]). \tag{A.7}$$

If the Event $\mathcal{P}$ in Equation (A.7) is true, it implies that at least one of the following events must be true.

$$\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t+\mathbf{C}_t)\leq\mathbf{1}^\top(\mathbf{I}-\mathbf{A})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}, \tag{A.8}$$

$$\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})(\hat{\boldsymbol{\beta}}_t-\mathbf{C}_t)\geq\mathbf{1}^\top(\mathbf{I}-\mathbf{A})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta}, \tag{A.9}$$

$$\mathbf{1}^\top(\mathbf{I}-\mathbf{A})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}<\mathbf{1}^\top(\mathbf{I}-\mathbf{A})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta}+2\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})\mathbf{C}_t. \tag{A.10}$$

First, we consider Equation (A.8). Based on our problem formulation and proposed solution, we know that matrices $\mathbf{A}$ and $\hat{\mathbf{A}}_t$ are nilpotent with index N. Thus, $\mathbf{A}^N=\mathbf{0}_{N\times N}$ and $\hat{\mathbf{A}}_t^N=\mathbf{0}_{N\times N}$. Hence, we can write the Taylor's series of $(\mathbf{I}-\mathbf{A})^{-1}$ and $(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}$ as

$$(\mathbf{I}-\mathbf{A})^{-1}=\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{N-1}, \tag{A.11}$$

and

$$(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}=\mathbf{I}+\hat{\mathbf{A}}_t+\hat{\mathbf{A}}_t^2+\cdots+\hat{\mathbf{A}}_t^{N-1}, \tag{A.12}$$

respectively. Substituting Equations (A.11) and (A.12) in (A.8) results in

$$\begin{aligned}&\mathbf{1}^\top(\mathbf{I}+\hat{\mathbf{A}}_t+\hat{\mathbf{A}}_t^2+\cdots+\hat{\mathbf{A}}_t^{N-1})\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t+\mathbf{C}_t)\\&\leq\mathbf{1}^\top(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{N-1})\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}.\end{aligned} \tag{A.13}$$

For $j=1,\ldots N$, we find the upper bound for

$$\mathbb{P}\left[\mathbf{1}^\top\hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t+\mathbf{C}_t)\leq\mathbf{1}^\top\mathbf{A}^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}\right]. \tag{A.14}$$

We consider the following Event $\mathcal{E}$.

$$\begin{aligned}&\mathbf{1}^\top\hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t+\mathbf{C}_t)+\mathbf{1}^\top\hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}\\&\leq\mathbf{1}^\top\hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}+\mathbf{1}^\top\mathbf{A}^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}.\end{aligned} \tag{A.15}$$

If $\mathcal{E}$ is true, then at least one of the following must hold.

$$\underbrace{\mathbf{1}^\top \hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t + \mathbf{C}_t) \leq \mathbf{1}^\top \hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}}_{\mathcal{I}}, \tag{A.16}$$

$$\underbrace{\mathbf{1}^\top \hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta} \leq \mathbf{1}^\top \mathbf{A}^{j-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\beta}}_{\mathcal{II}}. \tag{A.17}$$

Therefore, we have

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{I}] + \mathbb{P}[\mathcal{II}]. \tag{A.18}$$

Let $\mathbf{y}_t^\top = \mathbf{1}^\top \hat{\mathbf{A}}_t^{j-1}\text{diag}(\mathbf{x}^*)$. If Event $\mathcal{I}$ is true, then at least one of the following must hold.

$$\mathbf{y}_t^\top[v_1](\hat{\boldsymbol{\beta}}_t[v_1] + \mathbf{C}_t[v_1]) \leq \mathbf{y}_t^\top[v_1]\boldsymbol{\beta}[v_1], \tag{A.19}$$

$$\mathbf{y}_t^\top[v_2](\hat{\boldsymbol{\beta}}_t[v_2] + \mathbf{C}_t[v_2]) \leq \mathbf{y}_t^\top[v_2]\boldsymbol{\beta}[v_2], \tag{A.20}$$

$$\vdots$$

$$\mathbf{y}_t^\top[v_{|\mathcal{I}(\mathbf{x}^*)|}](\hat{\boldsymbol{\beta}}_t[v_{|\mathcal{I}(\mathbf{x}^*)|}] + \mathbf{C}_t[v_{|\mathcal{I}(\mathbf{x}^*)|}]) \leq \mathbf{y}_t^\top[v_{|\mathcal{I}(\mathbf{x}^*)|}]\boldsymbol{\beta}[v_{|\mathcal{I}(\mathbf{x}^*)|}]. \tag{A.21}$$

For $k = 1, \ldots, |\mathcal{I}(\mathbf{x}^*)|$, we have

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{y}_t^\top[v_k](\hat{\boldsymbol{\beta}}_t[v_k] + \mathbf{C}_t[v_k]) \leq \mathbf{y}_t^\top[v_k]\boldsymbol{\beta}[v_k]\right] \\
&\qquad \overset{(a)}{=} \mathbb{P}\left[\mathbf{m}_t[v_k](\hat{\boldsymbol{\beta}}_t[v_k] + \mathbf{C}_t[v_k]) \leq \mathbf{m}_t[v_k]\boldsymbol{\beta}[v_k]\right] \\
&\qquad \overset{(b)}{\leq} e^{-(2/\mathbf{m}_t[v_k])\mathbf{m}_t[v_k]^2\mathbf{C}_t[v_k]^2} \\
&\qquad \overset{(c)}{=} e^{-2(s+1)\ln t} \\
&\qquad = t^{-2(s+1)},
\end{aligned} \tag{A.22}$$

where (a) holds since $\mathbf{y}_t^\top[v_k] \geq 0$, $\forall k$, (b) follows from Lemma 1, and (c) results from the definition of $\mathbf{C}_t$. Hence, for Event $\mathcal{I}$, we conclude that

$$\mathbb{P}[\mathcal{I}] \leq |\mathcal{I}(\mathbf{x}^*)|t^{-2(s+1)} \leq st^{-2(s+1)}. \tag{A.23}$$

Now, we consider Event $\mathcal{II}$. Based on Theorem 1 in [21], we know that we can identify the adjacency matrix A uniquely by N samples gathered during the initialization period of our proposed algorithm. This means that with probability 1, after the time point $\theta = N < \infty$, $\hat{\mathbf{A}}_t = \mathbf{A}$ holds for all $t > \theta$. Therefore, for $t > N$, Event $\mathcal{II}$ holds with probability 1.

Combining the aforementioned results with Equation (A.18), we find the upper bound for Equation (A.14) as

$$\mathbb{P}\left[\mathbf{1}^\top \hat{\mathbf{A}}_t^{j-1}\mathrm{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t + \mathbf{C}_t) \leq \mathbf{1}^\top \mathbf{A}^{j-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta}\right] \leq st^{-2(s+1)}, \tag{A.24}$$

for each $j = 1,\ldots,N$. Since $\hat{\mathbf{A}}_t = \mathbf{A}$, $\forall t > N$ and the length of the largest path in the graph is p, we can rewrite Equations (A.11) and (A.12) as [52]

$$(\mathbf{I}-\mathbf{A})^{-1} = \mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^p, \tag{A.25}$$

and

$$(\mathbf{I}-\hat{\mathbf{A}}_t)^{-1} = \mathbf{I}+\hat{\mathbf{A}}_t+\hat{\mathbf{A}}_t^2+\cdots+\hat{\mathbf{A}}_t^p, \tag{A.26}$$

respectively. Therefore, by using Equations (A.25) and (A.26) in place of Equations (A.11) and (A.12), and based on Equation (A.24), the following holds for Equation (A.8).

$$\mathbb{P}\left[\mathbf{1}^\top (\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\mathrm{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\beta}}_t + \mathbf{C}_t) \leq \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta}\right] \leq s^p t^{-2p(s+1)}. \tag{A.27}$$

For Equation (A.9), we have similar results as follows.

$$\mathbb{P}\left[\mathbf{1}^\top (\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\mathrm{diag}(\mathbf{x}_{t+1})(\hat{\boldsymbol{\beta}}_t - \mathbf{C}_t) \geq \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta}\right] \leq s^p t^{-2p(s+1)}. \tag{A.28}$$

Finally, we consider Equation (A.10). We have

$$\begin{aligned}
&\mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - 2\mathbf{1}^\top (\mathbf{I}-\hat{\mathbf{A}}_t)^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\mathbf{C}_t \\
&\overset{(a)}{=} \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - 2\sum_{j:j\in\mathcal{I}(\mathbf{x}_{t+1})} \mathbf{w}_t^\top[j]\mathbf{C}_t[j] \\
&\overset{(b)}{=} \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - 2\sum_{j:j\in\mathcal{I}(\mathbf{x}_{t+1})} \mathbf{w}_t^\top[j]\sqrt{\frac{(s+1)\ln t}{\mathbf{m}_t[j]}} \\
&\overset{(c)}{\geq} \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - 2sw_{\max}\sqrt{\frac{(s+1)\ln T}{l}} \\
&\overset{(d)}{\geq} \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - \Delta_{\min} \\
&\overset{(e)}{\geq} \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\beta} - \mathbf{1}^\top (\mathbf{I}-\mathbf{A})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\beta} - \Delta(\mathbf{x}_{t+1}) = 0,
\end{aligned} \tag{A.29}$$

where in (a) and (c) we used the definition of $\mathbf{w}_t^\top$ and $w_{\max}$, respectively. Moreover, in (b) and (d), we substituted the value for $\mathbf{C}_t[j]$ and l, respectively. (e) follows from the

definition of $\Delta_{\min}$. Hence, we conclude that Equation (A.10) never happens. By using Equations (A.27), (A.28), and (A.29), we achieve the following.

$$\begin{aligned}
\mathbb{E}[\mathscr{T}_i(T)] &\leq \left\lceil \frac{4(s+1)\ln T}{(\frac{\Delta_{\min}}{s w_{\max}})^2} \right\rceil + \sum_{t=1}^{\infty} \left[\sum_{m_{v_1}=1}^{t} \cdots \sum_{m_{v_s}=1}^{t} \sum_{m_{u_1}=l}^{t} \cdots \sum_{m_{us}=l}^{t} 2s^p t^{-2p(s+1)} \right] \\
&\leq \frac{4w_{\max}^2 s^2 (s+1)\ln T}{\Delta_{\min}^2} + 1 + s^p \sum_{t=1}^{\infty} 2t^{-2} \\
&\leq \frac{4w_{\max}^2 s^2 (s+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} s^p. \qquad \text{(A.30)}
\end{aligned}$$

Therefore, the expected regret is upper bounded as

$$\begin{aligned}
\mathcal{R}_T(\mathcal{X}) &\leq \Delta_{\max} \sum_{i=1}^{N} \mathbb{E}[\mathscr{T}_i(T)] \\
&\leq \sum_{i=1}^{N} \left[\frac{4w_{\max}^2 s^2 (s+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} s^p \right] \Delta_{\max} \\
&\leq \left[\frac{4w_{\max}^2 s^2 (s+1) N \ln T}{\Delta_{\min}^2} + N + \frac{\pi^2}{3} s^p N \right] \Delta_{\max}. \qquad \text{(A.31)}
\end{aligned}$$

■

A.2 More on the experiments

A.2.1 Additional synthetic and real-data experiments

Synthetic data. Figure A.1 shows the cumulative expected regret for the experiment presented in Chapter 3 in Figure 3.2. Figure A.2 shows the performance of SEM-UCB against the benchmarks. SEM-UCB requires to spend the first N rounds in order to create the necessary data for learning the underlying causal graph. However, the ground truth causal graph is a priori given to the other algorithms except for the FTRL-Hybrid algorithm that does not consider the combinatorial optimization and directly uses the vectors of the overall rewards after each round. Figure A.2 shows that SEM-UCB can be categorized with DFL-CSR, CUCB, and CTS in terms of its performance even though it does not have the prior knowledge of the directed graph. The differences in the performance of the UCB-based algorithms of CUCB, DFL-CSR, and SEM-UCB come from the formation of the exploration terms of their corresponding UCB indices that are using different scales. DFL-CSR is apparently more aggressive since the exploration term is smaller, while the exploration term in the UCB indices of SEM-UCB is the biggest among these three algorithms.

Real data. As the governments try to contain the spread of Covid-19, they usually adopt restrictive measures such as quarantine over the regions that are showing the most number of overall daily new infections. As a result, they destructively ignore the effects of causal spread of the virus, meaning that they only focus on the overall daily new cases of regions without their causal effects on other regions. Therefore, we refer to this method of finding the best political interventions as the *naive approach*. Our goal is to show the superiority of our proposed algorithm over this *naive approach*. Similar to the experiments in our paper, we run SEM-UCB to find the 6 regions that are contributing the most to the total number of daily new cases in Italy. Figure A.4 compares the performance of our algorithm with that of the naive approach. The diagram shows

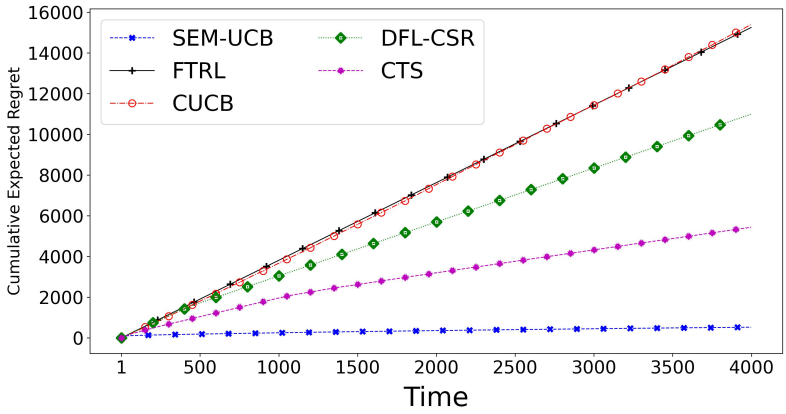


Figure A.1: Cumulative expected regret for the experiment in Figure 3.2.

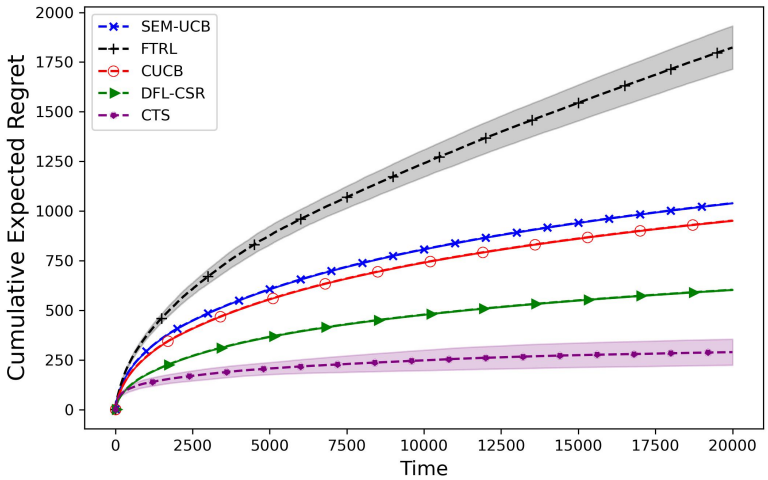


Figure A.2: Synthetic data experiment with the prior knowledge of the ground truth graph provided to CUCB, DFL-CSR, and CTS.

the ratio of the amount of contributions of the selected regions by the algorithms over the total number of daily new infections in the country for each day. As expected, after the initialization phase, SEM-UCB learns the underlying graph that influences the data. Consequently, it performs better with respect to the naive approach due to the fact that it takes the effects of causalities into account. We note that, due to such causal effects, it might be the case that a region with a lower number of overall daily cases contributes more than other regions with higher number of overall daily cases. This diagram provides the evidence that our framework can be highly effective in real-world applications such as analysis of the spread of Covid-19.

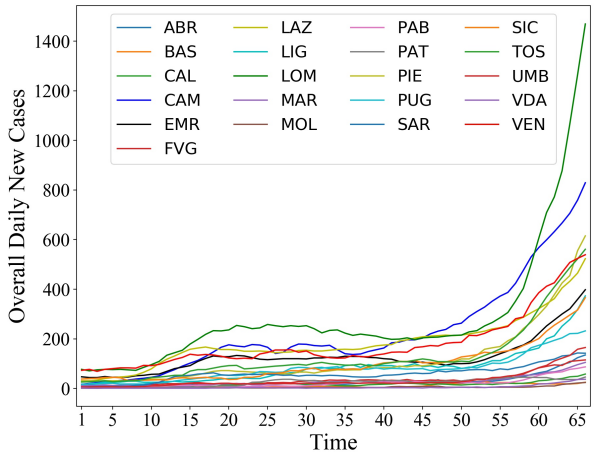


Figure A.3: Overall daily new cases of Covid-19 for different regions in Italy during the study period.

A.2.2 Abbreviations of regions in Italy

Table A.1 lists the abbreviations together with the original names of the 21 regions in Italy that we study in our numerical experiments.

Table A.1: List of regions in Italy and the corresponding abbreviations.

Abbreviation	Region Name
ABR	Abruzzo
BAS	Basilicata
CAL	Calabria
CAM	Campania
EMR	Emilia-Romagna
FVG	Friuli Venezia Giulia
LAZ	Lazio
LIG	Liguria
LOM	Lombardia
MAR	Marche
MOL	Molise
PAB	Provincia Autonoma di Bolzano
PAT	Provincia Autonoma di Trento
PIE	Piemonte
PUG	Puglia
SAR	Sardegna / Sardigna
SIC	Siciliana
TOS	Toscana
UMB	Umbria
VDA	Valle d'Aosta / Vallée d'Aoste
VEN	Veneto

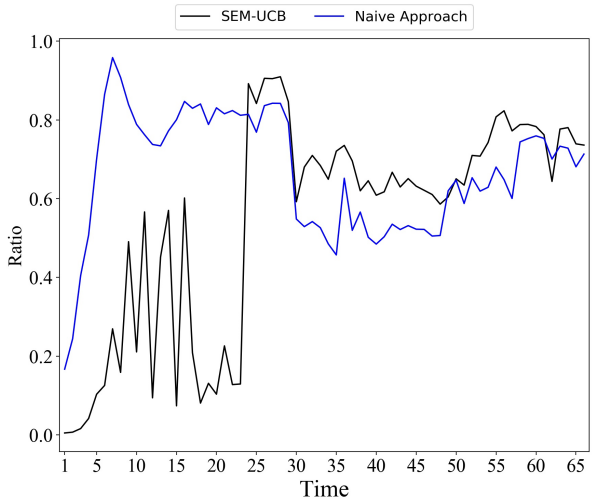


Figure A.4: The ratio of the amount of contributions of the selected regions by SEM-UCB and the naive approach over the total number of daily new infections in the country for each day.

A.2.3 Signals of overall daily new cases of Covid-19 infection

Italy has been severely affected by the Covid-19 pandemic. In April 2020, the country had the highest death toll in Europe. From the beginning of the pandemic, with the goal of containing the outbreak, the Italian government has put in place an increasing number of restrictions. Figure A.3 depicts the overall daily new cases of covid-19 of the 21 regions in Italy for the considered time interval in our numerical experiments. Due to space limitation, we use abbreviations for region names. Table A.1 lists the abbreviations together with the original names of the regions.

Appendix B

Additional Material for Chapter 4

B.1 Theoretical proofs

B.1.1 Proof of theorem 2

The theoretical analysis relies on the provided regret upper bound for the CSB problem with causally related rewards in the stationary environment [116]. In addition, we used theoretical analysis of regret [173] which yields the results of the combinatorial semi-bandit without a graph structure for the rewards. For the non-stationary setting we require the following assumption on the delay d_i of the change point indexed by i and maximum delay d of the GLR change-point detection:

Assumption 9. *Let* $\Delta_{\min}^{\text{change}} = \min_i \max_{k \in \mathcal{K}} |\mu_{k,i} - \mu_{k,i-1}|$.

$$\nu_i - \nu_{i-1} \geq 2d_i \qquad \forall i \in \{1, ..., N\},$$

where $d \leq \dfrac{K \log T}{p\left(\Delta_{\min}^{\text{change}}\right)^2}$

We explicitly assume, that the maximum delay is bounded and that the delayed detection will always occur in the respective consecutive stationary segment. The upper bound on the maximum delay is taken from the proof of Corollary 4.3. in [173]. For the proof of Theorem 1 we explicitly need the false alarm probability in the stationary scenario;

Lemma 3 (Lemma 4.5 in [173])**.** *Under the stationary scenario, with confidence level* $\delta > 0$ *we have that.*

$$\mathbb{P}(\tau_1 \leq T) \leq K\delta$$

We also need the probability that the GLR detects change reasonably well within a delay d.

Lemma 4 (Lemma 12 in [23])**.** *Define the event $\mathcal{C}_i$ that up to change-point i all changes have been detected successfully within a small delay d:*

$$\mathcal{C}_i = \{\forall j \leq i, \tau_j \in \{\nu_j + 1, ..., \nu_j + d\}\}$$

then: $\mathbb{P}(\tau_i \leq \nu_i | \mathcal{C}_{i-1}) \leq K\delta$ and $\mathbb{P}(\tau_i \geq \nu_i + d) \leq \delta$, with τ_i as the detection time of the ith change-point.

Lemma 5. ([12]) *Let $z_1, z_2, \ldots, z_m$ be random variables and $z_i \in [0,1]$, $\forall i$. Moreover, $\mathbb{E}[z_t | z_1, \ldots, z_{t-1}] = \alpha$, for all $t = 1, \ldots, m$. Then, for all $D \geq 0$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^{m} z_i - m\alpha\right| \geq D\right] \leq e^{-\frac{2D^2}{m}}. \tag{B.1}$$

And finally we require the upper regret bound of the stationary case in [116]:

Lemma 6 (Improved version of theorem 1)**.** *Let $\omega_t^T = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_{t+1})$ and $\omega_{\max} = \max_t \max_k \omega_t[k]$, $k \in \mathcal{K}$. In the stationary case of the PS-SEM-UCB algorithm, the upper regret bound is given as:*

$$\mathcal{R}(T) \leq \left[\frac{4\omega_{\max}^2 m^2 (m+1) K \log(T)}{\Delta_{\min}^2} + \frac{\pi^2}{3} mK + K\right] \Delta_{\max},$$

with $\Delta_{\max}$ as the largest suboptimality gap and $\Delta_{\min}$ smallest suboptimality gap.

Proof. The proof follows mostly the work of [116] and also uses mostly the same notation.

The *index set* of a decision vector is defined as $\mathbf{x} \in \mathcal{X}$ by $\mathcal{I}(\mathbf{x}) = \left\{k \mid \mathbf{x}[k] \neq 0, \forall k \in [K]\right\}$ and the confidence bound of base arm k at time t is defined as $\mathbf{C}_t[k] = \sqrt{\frac{(m+1)\ln t}{n_{k,t}}}$. At each time t, we store the empirical average of instantaneous rewards $\hat{\mu}_{k,t}$ and the calculated confidence bounds $\mathbf{C}_t[k]$ of all base arms $k \in \{1, ..., K\}$ in vectors $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{C}_t$, respectively. We have $\mathbf{U}_t = \hat{\boldsymbol{\mu}}_t + \mathbf{C}_t$. In order to make the proof more readable, we use the equivalent formulation: $\mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{U}_{t-1}) \mathbf{x}_t = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}_t) \mathbf{U}_{t-1}$. At each time t, we define the *selection index* for a decision vector $\mathbf{x} \in \mathcal{X}$ as $I_t(\mathbf{x}) = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \operatorname{diag}(\mathbf{x}) \mathbf{U}_{t-1}$. We further simplify the notation, by excluding the time index t in $n_{k,t}$ and use n_k to denote the number of times that the base arm k has been observed up to the current time instance.

For any $\mathbf{x} \in \mathcal{X}$, the counter $\mathcal{T}_{\mathbf{x}}(t)$ is used to represent the total number of times the decision vector $\mathbf{x}$ is selected up to time t. Finally, for each base arm $k \in [K]$, we define a counter $\mathcal{T}_k(t)$ which is updated as follows. At each time t after the initialization phase that a suboptimal decision vector $\mathbf{x}_t$ is selected, there is at least one base arm $k \in [K]$ such that $k = \underset{k \in \mathcal{I}(\mathbf{x}_t)}{\operatorname{argmin}} n_{k,t}$. In this case, if the base arm k is unique, we increment $\mathcal{T}_k(t)$

by 1. If there are more than one such base arm, we break the tie and select one of them arbitrarily to increment its corresponding counter.

We start by rewriting the expected regret as

$$\mathcal{R}(T) = T\,\psi(\mathbf{x}^*) - \sum_{t=1}^{T} \psi(\mathbf{x}_t) = \sum_{\mathbf{x}:\psi(\mathbf{x})<\psi(\mathbf{x}^*)} \Delta(\mathbf{x})\mathbb{E}[\mathcal{T}_{\mathbf{x}}(T)], \tag{B.2}$$

with $\psi(\mathbf{x})$ as reward for superarm $\mathbf{x}$. Since we are in a stationary environment with constant base arm distributions and no graph change, we can leave out the subscripts $\mathcal{A}_t$, $\boldsymbol{\mu}_t$. Based on the definition of the counters $\mathcal{T}_k(t)$ for the base arms $k \in [K]$, at each time t that a suboptimal decision vector is selected, only one of such counters is incremented by 1. Thus, we have [56]

$$\mathbb{E}\left[\sum_{\mathbf{x}:\psi(\mathbf{x})<\psi(\mathbf{x}^*)} \mathcal{T}_{\mathbf{x}}(t)\right] = \mathbb{E}\left[\sum_{k=1}^{K} \mathcal{T}_k(t)\right], \tag{B.3}$$

which implies that

$$\sum_{\mathbf{x}:\psi(\mathbf{x})<\psi(\mathbf{x}^*)} \mathbb{E}\left[\mathcal{T}_{\mathbf{x}}(t)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\mathcal{T}_k(t)\right]. \tag{B.4}$$

Therefore, we observe that

$$\begin{aligned}\mathcal{R}(T) &= \sum_{\mathbf{x}:\psi(\mathbf{x})<\psi(\mathbf{x}^*)} \Delta(\mathbf{x})\mathbb{E}[\mathcal{T}_{\mathbf{x}}(T)] \\ &\overset{(*)}{\leq} \Delta_{\max} \sum_{k=1}^{K} \mathbb{E}[\mathcal{T}_k(T)],\end{aligned} \tag{B.5}$$

where $(*)$ follows from the definition of $\Delta_{\max}$.

Let $\mathbb{I}_k(t)$ denote the indicator function which is equal to 1 if $\mathcal{T}_k(t)$ is increased by 1 at time t, and is 0 otherwise. Therefore,

$$\mathcal{T}_k(T) = \sum_{t=K+1}^{T} \mathbb{1}\left\{\mathbb{I}_k(t) = 1\right\}. \tag{B.6}$$

If $\mathbb{I}_i(t) = 1$, it means that a suboptimal decision vector $\mathbf{x}_t$ is selected at time t. In this

case, $n_{k,t} = \min\left\{n_{j,t} | j \in \mathcal{I}(\mathbf{x}_t)\right\}$. Let $l = \left\lceil \frac{4(m+1)\ln T}{(\frac{\Delta_{\min}}{m w_{\max}})^2} \right\rceil$. Then,

$$\begin{aligned}
\mathscr{T}_k(T) &= \sum_{t=K+1}^{T} \mathbb{1}\left\{\mathbb{I}_k(t) = 1\right\} \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left\{\mathbb{I}_k(t) = 1 \ \&\ \mathscr{T}_k(t-1) \geq l\right\} \\
&\leq l + \sum_{t=K+1}^{T} \mathbb{1}\left\{I_t(\mathbf{x}^*) \leq I_t(\mathbf{x}_t) \ \&\ \mathscr{T}_k(t-1) \geq l\right\} \\
&= l + \sum_{t=K+1}^{T} \mathbf{1}\{\mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \mathrm{diag}(\mathbf{x}^*) \mathbf{U}_{t-1} \\
&\qquad \leq \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t-1})^{-1} \mathrm{diag}(\mathbf{x}_t) \mathbf{U}_{t-1} \ \&\ \mathscr{T}_k(t-1) \geq l\} \\
&= l + \sum_{t=K}^{T} \mathbf{1}\{\mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t})^{-1} \mathrm{diag}(\mathbf{x}^*) \mathbf{U}_{t} \\
&\qquad \leq \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_{t})^{-1} \mathrm{diag}(\mathbf{x}_{t+1}) \mathbf{U}_{t} \ \&\ \mathscr{T}_k(t) \geq l\}. \qquad \text{(B.7)}
\end{aligned}$$

Based on the definition of $\mathscr{T}_k(t)$, we have $\mathscr{T}_k(t) \leq n_{k,t}$, $\forall k \in [K]$. Therefore, when $\mathscr{T}_k(t) \geq l$, the following holds [56].

$$l \leq \mathscr{T}_k(t) \leq n_{j,t}, \quad \forall j \in \mathcal{I}(\mathbf{x}_{t+1}). \qquad \text{(B.8)}$$

Let $\mathbf{v}_{t+1}^\top = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_t)^{-1} \mathrm{diag}(\mathbf{x}^*)$ and $\mathbf{u}_{t+1}^\top = \mathbf{1}^\top (\mathbf{I} - \hat{\mathbf{W}}_t)^{-1} \mathrm{diag}(\mathbf{x}_{t+1})$. We order the elements in sets $\mathcal{I}(\mathbf{x}^*)$ and $\mathcal{I}(\mathbf{x}_{t+1})$ arbitrarily. In the following, our results are independent of the way we order these sets. Let v_k, $k = 1, \ldots, |\mathcal{I}(\mathbf{x}^*)| \leq m$, represent the kth element in $\mathcal{I}(\mathbf{x}^*)$ and u_k, $k = 1, \ldots, |\mathcal{I}(\mathbf{x}_{t+1})| \leq m$, represent the kth element in $\mathcal{I}(\mathbf{x}_{t+1})$.

According to the aforementioned, we have

$$\begin{aligned}
&\mathscr{T}_k(T) \leq l + \sum_{t=K}^{T} \mathbf{1}\left\{ \min_{0 < n_{v_1}, \ldots, n_{v_{|\mathcal{I}(\mathbf{x}^*)|}} \leq t} \sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|} \mathbf{v}_{t+1}^\top [v_j] (\hat{\boldsymbol{\mu}}_{v_j,t} + \mathbf{C}_t[v_j]) \leq \right. \\
&\left. \max_{l \leq n_{u_1}, \ldots, n_{u_{|\mathcal{I}(\mathbf{x}_{t+1})|}} \leq t} \sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|} \mathbf{u}_{t+1}^\top [u_j] (\hat{\boldsymbol{\mu}}_{u_j,t} + \mathbf{C}_t[u_j]) \right\} \\
&\leq l + \sum_{t=1}^{\infty} \sum_{n_{v_1}=1}^{t} \cdots \sum_{n_{v_{|\mathcal{I}(\mathbf{x}^*)|}}=1}^{t} \sum_{n_{u_1}=l}^{t} \cdots \sum_{n_{u_{|\mathcal{I}(\mathbf{x}_{t+1})|}}=l}^{t}
\end{aligned}$$

$$\mathbf{1}\left\{\sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|}\mathbf{v}_{t+1}^{\top}[v_j](\hat{\mu}_{v_j,t}+\mathbf{C}_t[v_j]) \leq \sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|}\mathbf{u}_{t+1}^{\top}[u_j](\hat{\mu}_{u_j,t}+\mathbf{C}_t[u_j])\right\}. \tag{B.9}$$

We define the Event $\mathcal{P}$ as

$$\sum_{j=1}^{|\mathcal{I}(\mathbf{x}^*)|}\mathbf{v}_{t+1}^{\top}[v_j](\hat{\mu}_{v_j,t}+\mathbf{C}_t[v_j]) \leq \sum_{j=1}^{|\mathcal{I}(\mathbf{x}_{t+1})|}\mathbf{u}_{t+1}^{\top}[u_j](\hat{\mu}_{u_j,t}+\mathbf{C}_t[u_j]). \tag{B.10}$$

If the Event $\mathcal{P}$ in Equation (B.10) is true, it implies that at least one of the following events must be true.

$$\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\mu}}_t+\mathbf{C}_t) \leq \mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t, \tag{B.11}$$

$$\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})(\hat{\boldsymbol{\mu}}_t-\mathbf{C}_t) \geq \mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t, \tag{B.12}$$

$$\mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t < \mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t + 2\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})\mathbf{C}_t. \tag{B.13}$$

First, we consider Equation (B.11) by finding the upper bound for

$$\mathbb{P}\left[\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\mu}}_t+\mathbf{C}_t) \leq \mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t\right] \tag{B.14}$$

We consider the following Event $\mathcal{E}$.

$$\begin{aligned}&\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\mu}}_t+\mathbf{C}_t)+\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t\\ &\leq \mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t+\mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t.\end{aligned} \tag{B.15}$$

If $\mathcal{E}$ is true, then at least one of the following must hold.

$$\underbrace{\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\mu}}_t+\mathbf{C}_t) \leq \mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t}_{\mathcal{I}}, \tag{B.16}$$

$$\underbrace{\mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t \leq \mathbf{1}^{\top}(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t}_{\mathcal{II}}. \tag{B.17}$$

Therefore, we have

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{I}]+\mathbb{P}[\mathcal{II}]. \tag{B.18}$$

Let $\mathbf{y}_t^{\top} = \mathbf{1}^{\top}(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)$. If Event $\mathcal{I}$ is true, then at least one of the following must hold.

$$\mathbf{y}_t^{\top}[v_1](\hat{\mu}_{v_1,t}+\mathbf{C}_t[v_1]) \leq \mathbf{y}_t^{\top}[v_1]\mu_{v_1,t}, \tag{B.19}$$

$$\mathbf{y}_t^{\top}[v_2](\hat{\mu}_{v_2,t}+\mathbf{C}_t[v_2]) \leq \mathbf{y}_t^{\top}[v_2]\mu_{v_2,t}, \tag{B.20}$$

$$\vdots$$
$$\mathbf{y}_t^\top[v_{|\mathcal{I}(\mathbf{x}^*)|}](\hat{\mu}_{v_{|\mathcal{I}(\mathbf{x}^*)|},t}+\mathbf{C}_t[v_{|\mathcal{I}(\mathbf{x}^*)|}]) \leq \mathbf{y}_t^\top[v_{|\mathcal{I}(\mathbf{x}^*)|}]\mu_{v_{|\mathcal{I}(\mathbf{x}^*)|},t}. \tag{B.21}$$

we conclude for any arm that:

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{y}_t^\top[v_k](\hat{\mu}_{v_k,t}+\mathbf{C}_t[v_k]) \leq \mathbf{y}_t^\top[v_k]\mu_{v_k,t}\right] \\
&\quad\overset{(a)}{=} \mathbb{P}\left[n_{v_k,t}(\hat{\mu}_{v_k,t}+\mathbf{C}_t[v_k]) \leq n_{v_k,t}\mu_{v_k,t}\right] \\
&\quad\overset{(b)}{\leq} e^{-(2/n_{v_k,t})n_{v_k,t}^2\mathbf{C}_t[v_k]^2} \\
&\quad\overset{(c)}{=} e^{-2(m+1)\ln t} \\
&\quad= t^{-2(m+1)},
\end{aligned} \tag{B.22}$$

where (a) holds since $\mathbf{y}_t^\top[v_k] \geq 0$, $\forall k$, (b) follows from Lemma 5, and (c) results from the definition of $\mathbf{C}_t$. Hence, for Event $\mathcal{I}$, we conclude that

$$\mathbb{P}[\mathcal{I}] \leq |\mathcal{I}(\mathbf{x}^*)|t^{-2(m+1)} \leq mt^{-2(m+1)}. \tag{B.23}$$

Now, we consider Event $\mathcal{II}$. Based on Theorem 1 in [21], we know that we can identify the adjacency matrix $\mathbf{W}$ uniquely by K samples gathered during the initialization period of our proposed algorithm. This means that with probability 1, after the time point $t_{\text{init}} = K < \infty$, $\hat{\mathbf{W}}_t = \mathbf{W}$ holds for all $t > t_{\text{init}}$. Therefore, for $t > K$, Event $\mathcal{II}$ holds with probability 1.

Combining the aforementioned results with Equation (B.18), we find the upper bound for Equation (B.14) as

$$\mathbb{P}\left[\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}^*)(\hat{\boldsymbol{\mu}}_t+\mathbf{C}_t) \leq \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t\right] \leq mt^{-2(m+1)} \tag{B.24}$$

For Equation (B.12), we have similar results as follows.

$$\mathbb{P}\left[\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})(\hat{\boldsymbol{\mu}}_t-\mathbf{C}_t) \geq \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t\right] \leq mt^{-2(m+1)}. \tag{B.25}$$

Finally, for Equation (B.13) we have

$$\begin{aligned}
&\mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - 2\mathbf{1}^\top(\mathbf{I}-\hat{\mathbf{W}}_t)^{-1}\text{diag}(\mathbf{x}_{t+1})\mathbf{C}_t \\
&\overset{(a)}{=} \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\text{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - 2\sum_{j\in\mathcal{I}(\mathbf{x}_{t+1})}\omega_t^\top[j]\mathbf{C}_t[j]
\end{aligned}$$

$$\overset{(b)}{=} \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - 2\sum_{j\in\mathcal{I}(\mathbf{x}_{t+1})}\omega_t^\top[j]\sqrt{\frac{(m+1)\ln t}{n_{j,t}}}$$
$$\overset{(c)}{\geq} \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - 2mw_{\max}\sqrt{\frac{(m+1)\ln T}{l}}$$
$$\overset{(d)}{\geq} \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - \Delta_{\min}$$
$$\overset{(e)}{\geq} \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\mathbf{I}-\mathbf{W})^{-1}\mathrm{diag}(\mathbf{x}_{t+1})\boldsymbol{\mu}_t - \Delta(\mathbf{x}_{t+1}) = 0, \tag{B.26}$$

where in (a) and (c) we used the definition of $\omega_t^\top$ and $w_{\max}$, respectively. Moreover, in (b) and (d), we substituted the value for $\mathbf{C}_t[j]$ and l, respectively. (e) follows from the definition of $\Delta_{\min}$. Hence, we conclude that Equation (B.13) never happens.

By using Equations (B.24), (B.25), and (B.26), we achieve the following.

$$\begin{aligned}\mathbb{E}[\mathscr{T}_k(T)] &\leq \left\lceil \frac{4(m+1)\ln T}{(\frac{\Delta_{\min}}{mw_{\max}})^2} \right\rceil + \sum_{t=1}^{\infty}\left[\sum_{n_{v_1}=1}^{t}\cdots\sum_{n_{v_m}=1}^{t}\sum_{n_{u_1}=l}^{t}\cdots\sum_{n_{u_m}=l}^{t} 2mt^{-2(m+1)}\right] \\ &\leq \frac{4w_{\max}^2m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + m\sum_{t=1}^{\infty}2t^{-2} \\ &\leq \frac{4w_{\max}^2m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3}m. \end{aligned} \tag{B.27}$$

Therefore, the expected regret is upper bounded as

$$\begin{aligned}\mathcal{R}(T) &\leq \Delta_{\max}\sum_{k=1}^{K}\mathbb{E}[\mathscr{T}_k(T)] \\ &\leq \sum_{k=1}^{K}\left[\frac{4w_{\max}^2m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3}m\right]\Delta_{\max} \\ &\leq \left[\frac{4w_{\max}^2m^2(m+1)K\ln T}{\Delta_{\min}^2} + K + \frac{\pi^2}{3}mK\right]\Delta_{\max}. \end{aligned} \tag{B.28}$$

□

Proof of theorem 2. We assume there are $N-1$ points in time $\{\nu_1,...,\nu_{N-1}\}$ which mark the changes of base arm distributions inside any group and $\sigma(i)$ as the segment between ν_i and ν_{i-1}. We assume that multiple arm changes within a group occur at the same time, they are counted as one change in total. We define the events $\mathcal{F}_i = \{\tau_i > \nu_i\}$ and $D_i = \{\tau_i \leq \nu_i + d\}$ with d as expected delay of the GLR change-point detector and τ_i as time step in which the GLR is triggered. Additionally, we define the event $\mathcal{C}_i = \mathcal{F}_1 \cap D_1 \cap ... \cap \mathcal{F}_i \cap D_i$ that all change-points up to time i have been detected successfully. We also define $\nu_0 = 0$, $\nu_N = T$ and $\mathcal{C}_0 = \{\}$ as empty placeholder.

For the regret we have:

$$\begin{aligned}\mathcal{R}(T) &\leq \sum_{g\in G} \mathbb{E}\left[\mathcal{R}_g(T-\nu_1^g)\right] + \mathbb{E}\left[\mathcal{R}_g(\nu_1^g)\right] \\ &= \sum_{g\in G} \mathbb{E}\left[\mathcal{R}_g(T-\nu_1^g)\right] + \mathbb{E}\left[\mathcal{R}_g(\nu_1^g)\mathbb{1}(\mathcal{F}_1)\right] + \mathbb{E}\left[\mathcal{R}_g(\nu_1^g)\mathbb{1}(\bar{\mathcal{F}}_1)\right]\end{aligned}$$

where $\mathcal{R}_g$ denotes the regret per group g which comes down to a different number of arms assigned to group g. Next we have to determine $\mathbb{E}\left[\mathcal{R}_g(T-\nu_1^g)\right]$, for the case of readability we will estimate the regret terms per arm while leaving out the subscript g as the estimation holds for all groups:

$$\mathbb{E}\left[\mathcal{R}(T-\nu_1)\right] \leq \mathbb{E}\left[\mathcal{R}(T-\nu_1)|\mathcal{C}_1\right] + \Delta_{\max} T(1-\mathbb{P}(\mathcal{C}_1)) \tag{B.29}$$

We can further decompose $\mathbb{E}\left[\mathcal{R}(T-\nu_1)|\mathcal{C}_1\right]$:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(T-\nu_1)|\mathcal{C}_1\right] &\leq \mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_1\right] + \mathbb{E}\left[\mathcal{R}(\nu_2-\nu_1)|\mathcal{C}_1\right] \\ &\leq \mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_1\right] + \mathbb{E}\left[\mathcal{R}(\nu_2-\nu_1)\mathbb{1}(\mathcal{F}_2)|\mathcal{C}_1\right] + \mathbb{E}\left[\mathcal{R}(\nu_2-\nu_1)\mathbb{1}(\bar{\mathcal{F}}_2)|\mathcal{C}_1\right]\end{aligned}$$

inserting the result into Equation (B.29) we receive:

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(T-\nu_1)\right] \leq\ & \mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_1\right] \\ &+ \mathbb{E}\left[\mathcal{R}(\nu_2-\nu_1)\mathbb{1}(\mathcal{F}_2)|\mathcal{C}_1\right] \\ &+ \mathbb{E}\left[\mathcal{R}(\nu_2-\nu_1)\mathbb{1}(\bar{\mathcal{F}}_2)|\mathcal{C}_1\right] \\ &+ \Delta_{\max} T(1-\mathbb{P}(\mathcal{C}_1))\end{aligned}$$

as for the estimation of $\mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_1\right]$ we essentially repeat the previous two steps:

$$\mathbb{E}\left[\mathcal{R}(T-\nu_2|\mathcal{C}_1)\right] \leq \mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_2\right] + \Delta_{\max} T(1-\mathbb{P}(\mathcal{F}_2\cap D_2|\mathcal{C}_1)) \tag{B.30}$$

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_2\right] &\leq \mathbb{E}\left[\mathcal{R}(T-\nu_3)|\mathcal{C}_2\right] + \mathbb{E}\left[\mathcal{R}(\nu_3-\nu_2)|\mathcal{C}_2\right] \\ &= \mathbb{E}\left[\mathcal{R}(T-\nu_3)|\mathcal{C}_2\right] + \mathbb{E}\left[\mathcal{R}(\nu_3-\nu_2)\mathbb{1}(\mathcal{F}_3)|\mathcal{C}_2\right] \\ &+ \mathbb{E}\left[\mathcal{R}(\nu_3-\nu_2)\mathbb{1}(\bar{\mathcal{F}}_3)|\mathcal{C}_2\right]\end{aligned}$$

By recursively repeating the steps we can finally estimate the upper bound on the regret as:

$$\mathcal{R}(T) \leq \sum_{g\in G}\sum_{i=1}^{N_g} \mathbb{E}\left[\mathcal{R}_g(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\mathcal{F}_i^g)|\mathcal{C}_{i-1}^g\right] \tag{B.31}$$

$$+\mathbb{E}\left[\mathcal{R}_g(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\bar{\mathcal{F}}_i^g)|\mathcal{C}_{i-1}^g\right] \tag{B.32}$$

$$+\Delta_{\max}T(1-\mathbb{P}(\mathcal{F}_i^g\cap D_i^g|\mathcal{C}_{i-1}^g)). \tag{B.33}$$

This is the upper regret bound rewritten to showcase the regret contribution per stationary segment. Regarding the graph changes in our setting, the upper bound for a general asynchronous case where distribution and graph changes occur independently from each other can simply be constructed by including the the effect of each individual graph change separately. Since the UCB algorithm is independent from the state of the graph, each graph change would simply contribute a constant term $K\Delta_{\max}$, stemming from the graph learning phase, to the upper regret bound. As for the regret we evaluate each terms in Equations (B.31), (B.32) and (B.33) separately. We start with the last term inside the sum $\Delta_{\max}T(1-\mathbb{P}(\mathcal{F}_i^g\cap D_i^g|\mathcal{C}_{i-1}^g))$:

$$\begin{aligned}\Delta_{\max}T(1-\mathbb{P}(\mathcal{F}_i^g\cap D_i^g|\mathcal{C}_{i-1}^g)) &= \Delta_{\max}T\mathbb{P}(\bar{\mathcal{F}}_i^g\cup\bar{D}_i^g|\mathcal{C}_{i-1}^g)\\ &= T\Delta_{\max}\delta(K_g+1),\end{aligned}$$

for which results we used Lemma 4. For the second term we have due to Lemma 3:

$$\mathbb{E}\left[\mathcal{R}(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\bar{\mathcal{F}}_i^g)|\mathcal{C}_{i-1}^g\right] = (\nu_i^g-\nu_{i-1}^g)\Delta_{\max}K_g\delta.$$

For the first term the results of the stationary case are used, while the delay in the detection and the graph changes are considered as well:

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\mathcal{F}_i^g)|\mathcal{C}_{i-1}^g\right]\\ &\quad\leq K_gR_0(\nu_i^g-\nu_{i-1}^g)+\left[(\nu_i^g-\nu_{i-1}^g)\frac{p+N_{\mathbf{W}}K/T}{\zeta}+d+K_g+\frac{\pi^2}{3}mK_g\right]\Delta_{\max},\end{aligned}$$

where we have the result from Lemma 6
$R_0(T)=\frac{4\omega_{\max}^2m^2(m+1)\log(T)}{\Delta_{\min}^2}\Delta_{\max}$ as the base regret per arm of the non-stationary case, excluding the additional terms coming from the base-arm exploration phase and the base arm initialization. In this step, the effect of the delay is included, due to the last segment being a good event and $\tau_{i-1}>\nu_{i-1}$ as indicated by the conditional expectation. Finally we combine the previously estimated expressions summarize the final regret expression as:

$$\mathcal{R}(T)\leq$$

$$
\begin{aligned}
&\sum_{g\in G}\left[\sum_{i=1}^{N_g} K_g R_0(v_i^g - v_{i-1}^g) + \left[(v_i^g - v_{i-1}^g)p/\zeta + d\right]\Delta_{\max}\right.\\
&\left.+ [(v_i^g - v_{i-1}^g)K_g\delta + (K_g+1)T\delta + K_g + \frac{\pi^2 m K_g}{3}]\Delta_{\max}\right]\\
&+ N_{\mathbf{W}} K \Delta_{\max}\\
&\leq \sum_{g\in G}\left[N_g K_g R_0(T) + \Delta_{\max}\delta T(K_g + N_g + N_g K_g)\right.\\
&\left.+ \left(Tp/\zeta + dN_g + K_g N_g + \frac{\pi^2 m K_g N_g}{3}\right)\Delta_{\max}\right]\\
&+ N_{\mathbf{W}} K \Delta_{\max}\\
&= \sum_{g\in G}\left[N_g K_g R_0(T) + (\delta T + 1 + \frac{\pi^2 m}{3})N_g K_g \Delta_{\max}\right]\\
&+ \left(Tp + dN_G + \delta T(K + N_G) + N_{\mathbf{W}} K\right)\Delta_{\max}.
\end{aligned}
$$

□

B.1.2 Proof of corollary 1

For the proof of the corollary we make use of Assumption 9.

Proof of corollary 1. We insert $d \leq \frac{K\log T}{p\left(\Delta_{\min}^{\text{change}}\right)^2}$ into our expression of theorem 2:

$$
\begin{aligned}
\mathcal{R}(T) \leq &\sum_{g\in G}\left[N_g \frac{4\omega_{\max}^2 m^2(m+1)K_g \log(T)}{\Delta_{\min}^2} + (\delta T + 1 + \frac{\pi^2}{3}m)N_g K_g\right]\Delta_{\max}\\
&+ \Delta_{\max}\delta T(K + N_G)\\
&+ \left[Tp + \frac{K\log T}{p\left(\Delta_{\min}^{\text{change}}\right)^2}N_G\right]\Delta_{\max} + K N_{\mathbf{W}}\Delta_{\max}
\end{aligned}
$$

By choosing $\delta = \frac{1}{T}$ and $p = \sqrt{\frac{N_G K \log T}{T}}$ we finally get:

$$
\begin{aligned}
&\mathcal{R}(T) \leq \\
&\sum_{g \in G} N_g K_g \left[\frac{4\omega_{\max}^2 m^2 (m+1) \log(T)}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} m \right] \Delta_{\max} \\
&+ \left[K + N_G + \sqrt{N_G K T \log T} + \frac{N_G \sqrt{KT \log T}}{\sqrt{N_G} \left(\Delta_{\min}^{\text{change}} \right)^2} \right] \Delta_{\max} + K N_{\mathbf{W}} \Delta_{\max} \qquad \text{(B.34)} \\
&= \mathcal{O} \left(\frac{\sum_{g \in G} N_g K_g \log T}{\Delta_{\min}} + \frac{\sqrt{N_G K T \log T}}{\left(\Delta_{\min}^{\text{change}} \right)^2} \right) \Delta_{\max} + K N_{\mathbf{W}} \Delta_{\max}
\end{aligned}
$$

□

B.2 More on the experiments

B.2.1 Synthetic datast

Figure B.1 provides the expected values of the base arms' instantaneous reward distributions for each distribution stationary segment in our synthetic data experiment. Figure B.2 is the visualization of the base arms inside optimal super arms across time. Dark rectangles represent the four selected arms in each round. Graph changes happen at times $t = 4000, 8000, 12000, 16000$ and distribution changes happen at times $t = 5000, 10000, 15000, 20000$.

B.2.2 Real-world application

For the real data experiment, we grouped the provinces as the following; Group 1 includes Abruzzo, Basilicata, Campania, Lazio, Molise, Puglia, and Calabria. Group 2 includes Emilia-Romagna, Marche, Bolzano, Trento, Toscana, Umbria, Veneto, and Friuli Venezia Giulia. Group 3 has Liguria, Lombardia, Piemonte, and Valle d'Aosta. Group 4 includes Sardegna, and Sicilia. Region of Lazio is detected at $t = 57$ to have its distribution changed. Region of Emilia-Romagna is detected at $t = 63$ to have its distribution changed. The change point detector of the region of Liguria sends its signal at $t = 70$, and the region of Sardegna is detected at $t = 75$ to have its distribution changed. Consequently, all the 4 groups restart their UCB developments. Table A.1 lists the abbreviations together with the original names of the 21 regions in Italy that we study in our numerical experiments.

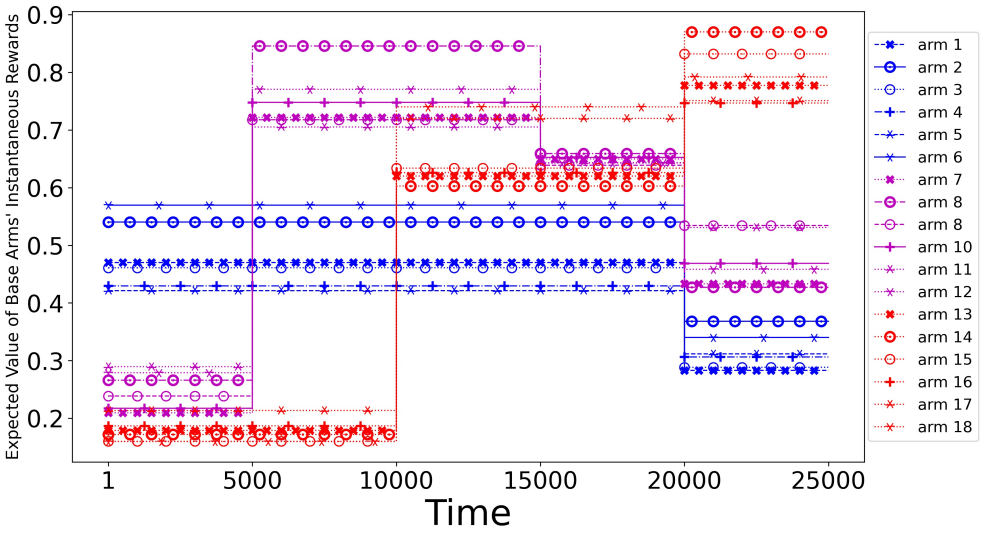


Figure B.1: Expected values of base arm's instantaneous rewards

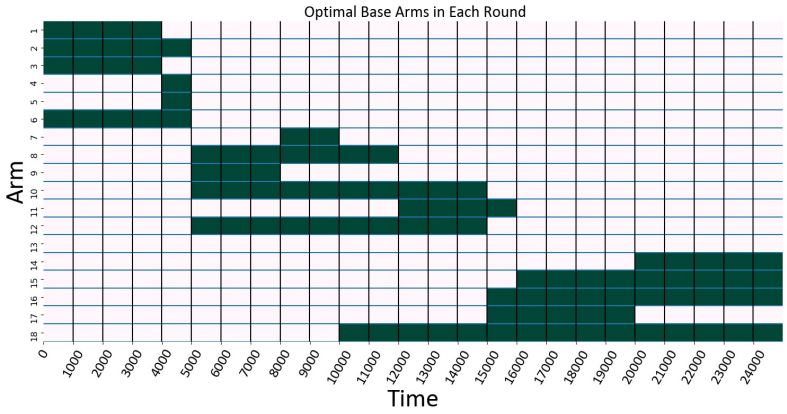


Figure B.2: Optimal super arms in synthetic dataset

Appendix C

Additional Material for Chapter 5

C.1 Some helper results

Proposition 1 (Bounds on norms of matrix products)**.** *Let* $\mathbf{M} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{N} \in \mathbb{R}^{n\times p}$. *Then*

$$\begin{aligned}
\|\mathbf{MN}\|_{q,1} &\leq \|\mathbf{M}\|_{\infty,1}\|\mathbf{N}\|_{q,1} \quad \forall q \in [1,\infty] \\
\|\mathbf{MN}\|_F &\leq \|\mathbf{M}\|\|\mathbf{N}\|_F \\
\|\mathbf{MN}\|_F &\leq \sqrt{\left\|\mathbf{M}^\top\mathbf{M}\right\|_{\infty,\infty}}\|\mathbf{N}\|_{2,1} \\
\|\mathbf{MN}\|_{2,1} &\leq \|\mathbf{M}\|_{2,1}\|\mathbf{N}\|
\end{aligned}$$

Proof.

First inequality For any $q \in [1,\infty]$, we have:

$$\left\|\mathbf{e}_i^\top\mathbf{MN}\right\|_q = \left\|\mathbf{e}_i^\top\mathbf{M}\sum_{j=1}^n \mathbf{e}_j\mathbf{e}_j^\top\mathbf{N}\right\|_q \leq \max_{1\leq j\leq n}\left|\mathbf{e}_i^\top\mathbf{Me}_j\right| \sum_{j=1}^n \left\|\mathbf{e}_j^\top\mathbf{N}\right\|_q = \max_{1\leq j\leq n}\left|(\mathbf{M})_{ij}\right|\|\mathbf{N}\|_{q,1}$$

Second inequality We have

$$\|\mathbf{MN}\|_F^2 = \sum_{j=1}^p \left\|\mathbf{MNe}_j\right\|^2 \leq \sum_{j=1}^p \|\mathbf{M}\|\left\|\mathbf{Ne}_j\right\|^2 = \|\mathbf{M}\|\|\mathbf{N}\|_F^2$$

Third inequality We have

$$\|\mathbf{MN}\|_F^2 = \mathrm{Tr}(\mathbf{MNN}^\top\mathbf{M}^\top) \leq \left\|\mathbf{M}^\top\mathbf{M}\right\|_{\infty,\infty}\left\|\mathbf{NN}^\top\right\|_{1,1}$$

Elements of (i,j) entry of matrix $\mathbf{NN}^\top$ is the inner product $\left\langle \mathbf{e}_i^\top\mathbf{N}, \mathbf{e}_j^\top\mathbf{N}\right\rangle$. Hence, we have

$$\left\|\mathbf{NN}^\top\right\|_{1,1} = \sum_{i,j}\left|\left\langle \mathbf{e}_i^\top\mathbf{N}, \mathbf{e}_j^\top\mathbf{N}\right\rangle\right| \leq \sum_{i,j}\left\|\mathbf{e}_i^\top\mathbf{N}\right\|\left\|\mathbf{e}_j^\top\mathbf{N}\right\| = \|\mathbf{N}\|_{2,1}^2\,.$$

Fourth inequality We have

$$\|\mathbf{MN}\|_{2,1} = \sum_{i=1}^{m}\|\mathbf{e}_i\mathbf{MN}\| \leq \sum_{i=1}^{m}\|\mathbf{e}_i\mathbf{M}\|\|\mathbf{N}\| = \|\mathbf{M}\|_{2,1}\|\mathbf{N}\|$$

□

Proposition 2 (Decomposition of a signal over a graph)**.** *For any $\mathcal{C} \in \mathcal{P}$*

- *Let $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}$ be a graph signal. Let us denote by $\mathbf{Z}_\mathcal{C}$ the signal obtained from $\mathbf{Z}$ by setting rows of vertices outside of $\mathcal{C}$ to zeros, and let $\mathbf{Z}_{|\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|\times d}$ be the signal obtained from $\mathbf{Z}_\mathcal{C}$ by removing the rows of vertices outside of $\mathcal{C}$. Also, let $\mathbf{B}_{|\mathcal{C}} \in \mathbb{R}^{|\mathcal{E}_\mathcal{C}|\times|\mathcal{C}|}$ be the matrix obtained by taking $\mathbf{B}_\mathcal{C}$, and removing rows of edges that link $\mathcal{C}$ to its outside, and the resulting null columns. It is clear that*

$$\mathbf{B}_\mathcal{C}\mathbf{Z} = \mathbf{B}_\mathcal{C}\mathbf{Z}_\mathcal{C} = \mathbf{B}_{|\mathcal{C}}\mathbf{Z}_{|\mathcal{C}} \tag{C.1}$$

- *Let $\mathbf{Q}_\mathcal{C} \coloneqq \mathbf{B}_\mathcal{C}^\dagger\mathbf{B}_\mathcal{C}$. Then*

$$\mathbf{I}_{|\mathcal{V}|} = \sum_{\mathcal{C}\in\mathcal{P}}\mathbf{J}_\mathcal{C} + \mathbf{Q}_\mathcal{C} \tag{C.2}$$

$$\mathbf{Q}_{\partial\mathcal{P}^c} \coloneqq= \mathbf{B}_{\partial\mathcal{P}^c}^\dagger\mathbf{B}_{\partial\mathcal{P}^c} = \sum_{\mathcal{C}\in\mathcal{P}}\mathbf{Q}_\mathcal{C} \tag{C.3}$$

where $\mathbf{J}_\mathcal{C} = \frac{\mathbf{1}_\mathcal{C}\mathbf{1}_\mathcal{C}^\top}{|\mathcal{C}|}$, $\mathbf{Q}_\mathcal{C} = \mathbf{B}_\mathcal{C}^\dagger\mathbf{B}_\mathcal{C}$ $\quad\forall\mathcal{C}\in\mathcal{P}$ and $\mathbf{Q}_{\partial\mathcal{P}^c} \coloneqq \mathbf{B}_{\partial\mathcal{P}^c}^\dagger\mathbf{B}_{\partial\mathcal{P}^c}$.

While $\sum_{\mathcal{C}\in\mathcal{P}}\mathbf{J}_\mathcal{C}$ projects each entry of a graph signal onto the mean vector value of its respective cluster, its residual $\mathbf{Q}_{\partial\mathcal{P}^c}$ can be interpreted as the projection onto the respective entries deviation from its cluster mean value.

Proof. Since the proof of the first point is trivial, we directly treat the second point. Denoting $\mathbf{B}_{|\mathcal{C}}^\dagger$ the pseudo-inverse of $\mathbf{B}_{|\mathcal{C}}$ it is a well-known linear algebra result that the matrix $Q_{|\mathcal{C}} \coloneqq \mathbf{B}_{|\mathcal{C}}^\dagger\mathbf{B}_{|\mathcal{C}}$ is the projector onto the null space of $\mathbf{B}_{|\mathcal{C}}$. Since $\mathcal{C}$ is connected,

the null space of $\mathbf{B}_{|\mathcal{C}}$ is unidimensional, and is generated by vector $\mathbf{1}_{|\mathcal{C}|} \in \mathbb{R}^{|\mathcal{C}|}$ having only ones as coordinates. Since the projector into that nullspace is $\mathbf{J}_{|\mathcal{C}|} \coloneqq \frac{\mathbf{1}_{|\mathcal{C}|}\mathbf{1}_{|\mathcal{C}|}}{|\mathcal{C}|}$, we deduce that

$$\begin{aligned} \mathbf{Z}_{|\mathcal{C}} &= \mathbf{J}_{|\mathcal{C}|}\mathbf{Z}_{|\mathcal{C}} + \mathbf{Q}_{|\mathcal{C}}\mathbf{Z}_{|\mathcal{C}} \\ \implies \mathbf{Z}_{\mathcal{C}} &= \mathbf{J}_{\mathcal{C}}\mathbf{Z}_{\mathcal{C}} + \mathbf{Q}_{\mathcal{C}}\mathbf{Z}_{\mathcal{C}} \\ &= \mathbf{J}_{\mathcal{C}}\mathbf{Z} + \mathbf{Q}_{\mathcal{C}}\mathbf{Z} \end{aligned}$$

where in the last line, $Q_{\mathcal{C}} \coloneqq \mathbf{B}_{\mathcal{C}}^{\dagger}\mathbf{B}_{\mathcal{C}}$. Consequently, we have

$$\begin{aligned} \mathbf{Z} &= \sum_{\mathcal{C}\in\mathcal{P}} \mathbf{Z}_{\mathcal{C}} \\ &= \sum_{\mathcal{C}\in\mathcal{P}} \mathbf{J}_{\mathcal{C}}\mathbf{Z} + \mathbf{Q}_{\mathcal{C}}\mathbf{Z} \end{aligned}$$

To prove the second point, we recall that $\mathbf{B}_{\partial\mathcal{P}^c}$ is the incidence matrix obtained by setting rows corresponding to edges in $\partial\mathcal{P}$ to zero. In other words, $\mathbf{B}_{\partial\mathcal{P}^c}$ is the incidence matrix of the graph after removing the boundary edges, and having exactly $|\mathcal{P}|$ connected components. Hence, $\mathbf{B}_{\partial\mathcal{P}^c}$ has a null space spanned by the set $\{\mathbf{1}_{\mathcal{C}}\}_{\mathcal{C}\in\mathcal{P}}$, and the orthogonal projector onto this null space is $\sum_{\mathcal{C}\in\mathcal{P}}\mathbf{J}_{\mathcal{C}}$. Combining this fact with the fact that $\mathbf{Q}_{\partial\mathcal{P}^c}$ is the projector onto the orthogonal of the null space of $\mathbf{B}_{\partial\mathcal{P}^c}$, we arrive at the second point. □

Proposition 3 (On the minimum topological centrality index of a graph vertex)**.** *Let $\mathcal{G}$ be a connected graph with incidence matrix $\mathbf{B}$ and vertex set size N, and let $\mathbf{L} \coloneqq \mathbf{B}^{\top}\mathbf{B}$. Let $c(\mathcal{G})$ denote the minimum value of inverses of diagonal element of $\mathbf{L}^{\dagger}$, called its minimum topological centrality index. Also let $a(\mathcal{G})$ be its algebraic connectivity, defined as the minimum non null eigenvalue of $\mathbf{L}$. Then*

- $c(\mathcal{G}) = \|\mathbf{L}\|_{\infty,\infty}^{-1}$.
- $c(\mathcal{G}) \geq a(\mathcal{G})$.
- *If $\mathcal{G}$ is weightless, then $c(\mathcal{G}) \leq \frac{N^2}{N-1}$.*

Proof. Since $\mathbf{L}$ is PSD, $\mathbf{L}^{\dagger}$ is PSD and hence $\left\|\mathbf{L}^{\dagger}\right\|_{\infty,\infty}$ is equal to the maximum diagonal entry of $\mathbf{L}^{\dagger}$. Taking the inverse proves the first point. Also, this implies that

$$c(\mathcal{G}) = \left\|\mathbf{L}^{\dagger}\right\|_{\infty,\infty}^{-1} \geq \left\|\mathbf{L}^{\dagger}\right\|^{-1} = a(\mathcal{G}), \tag{C.4}$$

where we used the fact that $\|\cdot\|_{\infty,\infty} \leq \|\cdot\|$ for matrices. This proves the second point of the proposition.

For the last point, assume $\mathcal{G}$ is weightless, let $\mathbf{L}_{comp}$ be the Laplacian of complete graph built on the vertices of $\mathcal{G}$. Then we have $\mathbf{L}_{comp} = N(\mathbf{I}_N - \mathbf{J}_N)$, where J is the square matrix of dimension N having $1/N$ as entries. From Fontan and Altafini [55, Lemma 4], we have

$$\mathbf{L}_{comp}^{\dagger} = (\mathbf{L}_{comp} + N\mathbf{J}_N)^{-1} - \frac{1}{N}\mathbf{J}_N = \frac{\mathbf{I}_N}{N} - \frac{1}{N}\mathbf{J}_N \tag{C.5}$$

which has diagonal elements $\frac{1}{N} - \frac{1}{N^2}$.

On the other hand, $\mathbf{L} \preccurlyeq \mathbf{L}_{comp}$ Hence, by Fontan and Altafini [55, lemma 4] we have for any $u \neq 0$

$$\mathbf{L}^{\dagger} = (\mathbf{L} + a\mathbf{J}_N)^{-1} - \mathbf{J}_N/a \succcurlyeq (\mathbf{L}_{comp} + a\mathbf{J}_N)^{-1} - \mathbf{J}_N/a = \mathbf{L}_{comp}^{\dagger}$$

This implies that the maximum diagonal entry of $\mathbf{L}^{\dagger}$ is at least equal to that of $\mathbf{L}_{comp}^{\dagger}$, *i.e.*to $\frac{1}{N} - \frac{1}{N^2}$. Taking the inverse of that entry finishes the proof.

□

C.2 Proofs of the different claims

C.2.1 Additional notation

The regularization term can be written more compactly using the incidence matrix of the graph $\mathbf{B} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ corresponding to an arbitrary orientation under the following form

$$\sum_{1\leq m<n\leq|\mathcal{V}|} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\| = \|\mathbf{B}\boldsymbol{\Theta}\|_{2,1} = \|\boldsymbol{\Theta}\|_{\mathcal{E}} \tag{C.6}$$

where the $\|\cdot\|_{2,1}$ norm denotes the sum of the L_2 norms o the rows of a matrix.[1] We provide notations that we use in the proofs of the different statements, in order to reduce the clutter. We define $\mathbf{E} \coloneqq \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ as the error signal, and its rows by $\{\boldsymbol{\varepsilon}_m\}_{m=1}^{|\mathcal{V}|}$.

While $\sum_{k=1}^{C} \mathbf{J}_{\mathcal{C}}$ projects each entry of a graph signal onto the mean vector value of its respective cluster, its residual $\mathbf{Q}_{\partial\mathcal{P}^c}$ can be interpreted as the projection onto the respective entries deviation from its cluster mean value.

Let $\boldsymbol{\eta}_m$ be a vector, vertically concatenated by noise terms of rewards received by node m, then we define $\mathbf{K} \in \mathbb{R}^{|\mathcal{V}|\times d}$ as the matrix of vertically concatenated row vectors $\boldsymbol{\eta}_m^{\top}\mathbf{X}_m$.

[1] It is possible that the notation $\|\cdot\|_{2,1}$ denotes the sum of $2-$norms of columns in the literature.

C.2.2 Oracle inequality

In this section, we present all intermediary theoretical results leading to Theorem 3 stating the oracle inequality. To reduce the clutter, we omit the dependence on t of several quantities. For instance, we write α and $\hat{\mathbf{\Theta}}$ instead of $\alpha(t)$ and $\hat{\mathbf{\Theta}}(t)$.

Lemma 7 (A first deterministic inequality)**.** *Let t be a time step. We have*

$$\frac{1}{2t\alpha}\sum_{m\in\mathcal{V}}\|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2+\|\mathbf{E}\|_{\partial\mathcal{P}^c}\leq\frac{1}{t\alpha}\langle\mathbf{K},\mathbf{E}\rangle+\|\mathbf{E}\|_{\partial\mathcal{P}} \tag{C.7}$$

Proof. By optimality of $\hat{\mathbf{\Theta}}$, we have

$$\frac{1}{2t}\sum_{m\in\mathcal{V}}\left\|\mathbf{X}_m\hat{\boldsymbol{\theta}}_m-\mathbf{y}_m\right\|^2+\alpha\left\|\hat{\mathbf{\Theta}}\right\|_{\mathcal{E}}\leq\frac{1}{2t}\sum_{m\in\mathcal{V}}\|\mathbf{X}_m\boldsymbol{\theta}_m-\mathbf{y}_m\|^2+\alpha\|\mathbf{\Theta}\|_{\mathcal{E}} \tag{C.8}$$

where the second line holds by definition of the observed rewards.

On the one hand, given a user index $m\in\mathcal{V}$, and since by definition of the observed rewards we have we have for the least squared terms

$$\begin{aligned}\left\|\mathbf{X}_m\hat{\boldsymbol{\theta}}_m-\mathbf{y}_m\right\|^2&=\left\|\mathbf{X}_m\hat{\boldsymbol{\theta}}_m-\mathbf{X}_m\boldsymbol{\theta}_m-\boldsymbol{\eta}_m\right\|^2\\&=\left\|\mathbf{X}_m\boldsymbol{\varepsilon}_m-\boldsymbol{\eta}_m\right\|^2\\&=\|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2+\|\mathbf{X}_m\boldsymbol{\theta}_m-\mathbf{y}_m\|^2-\boldsymbol{\eta}_m^\top\mathbf{X}_m\boldsymbol{\varepsilon}_m\end{aligned}$$

where we used the fact that $\mathbf{y}_m=\mathbf{X}_m\boldsymbol{\theta}_m+\boldsymbol{\eta}_m$, which holds by definition of the observed rewards. Summing over the users, and using the definition of $\mathbf{K}$, we have

$$\frac{1}{2t}\sum_{m\in\mathcal{V}}\left\|\mathbf{X}_m\hat{\boldsymbol{\theta}}_m-\mathbf{y}_m\right\|^2-\frac{1}{2t}\sum_{m\in\mathcal{V}}\|\mathbf{X}_m\boldsymbol{\theta}_m-\mathbf{y}_m\|^2=\frac{1}{2t}\sum_{m\in\mathcal{V}}\|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2-\frac{1}{t}\langle\mathbf{K},\mathbf{E}\rangle \tag{C.9}$$

On the other hand, we have for the estimated preference vectors

$$\begin{aligned}\left\|\hat{\mathbf{\Theta}}\right\|_{\mathcal{E}}&=\sum_{(m,n)\in\mathcal{E}}w_{mn}\left\|\hat{\boldsymbol{\theta}}_m-\hat{\boldsymbol{\theta}}_n\right\|\\&=\sum_{(m,n)\in\partial\mathcal{P}}w_{mn}\left\|\hat{\boldsymbol{\theta}}_m-\hat{\boldsymbol{\theta}}_n\right\|+\sum_{(m,n)\in\partial\mathcal{P}^c}w_{mn}\left\|\hat{\boldsymbol{\theta}}_m-\hat{\boldsymbol{\theta}}_n\right\|\\&=\left\|\hat{\mathbf{\Theta}}\right\|_{\partial\mathcal{P}}+\left\|\hat{\mathbf{\Theta}}\right\|_{\partial\mathcal{P}^c},\end{aligned}$$

For the true ones, and for any $\mathcal{C}\in\mathcal{P}$, let $\mathcal{E}_{\mathcal{C}}$ denote the edges linking the nodes of set of

nodes $\mathcal{C}$. It is clear that $\partial\mathcal{P}^c = \bigcup_{\mathcal{C}\in\mathcal{P}} \mathcal{E}_\mathcal{C}$ as a disjoint union, hence

$$\begin{aligned}
\|\boldsymbol{\Theta}\|_\mathcal{E} &= \sum_{(m,n)\in\mathcal{E}} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\| \\
&= \sum_{(m,n)\in\partial\mathcal{P}} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\| + \sum_{(m,n)\in\partial\mathcal{P}^c} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\| \\
&= \|\boldsymbol{\Theta}\|_{\partial\mathcal{P}} + \sum_{\mathcal{C}\in\mathcal{P}} \sum_{(m,n)\in\mathcal{E}_\mathcal{C}} w_{mn}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_n\| \\
&= \|\boldsymbol{\Theta}\|_{\partial\mathcal{P}}
\end{aligned}$$

where the last equality holds due to the cluster assumption.

Hence, we have

$$\begin{aligned}
\|\boldsymbol{\Theta}\|_\mathcal{E} - \left\|\hat{\boldsymbol{\Theta}}\right\|_\mathcal{E} &= \|\boldsymbol{\Theta}\|_{\partial\mathcal{P}} - \left\|\hat{\boldsymbol{\Theta}}\right\|_{\partial\mathcal{P}} - \left\|\hat{\boldsymbol{\Theta}}\right\|_{\partial\mathcal{P}^c} \\
&\leq \|\mathbf{E}\|_{\partial\mathcal{P}} - \left\|\hat{\boldsymbol{\Theta}}\right\|_{\partial\mathcal{P}^c},
\end{aligned} \tag{C.10}$$

where the first inequality holds due to the triangle inequality, and the last one since $\|\boldsymbol{\Theta}\|_{\partial\mathcal{P}^c} = 0$. Combining Equations (C.8), (C.9), and (C.10), we obtain the result of the statement. □

In the proof for the oracle inequality, we utilize projection operators on the graph signal, that we define as followed:

While $\sum_{k=1}^{C} \mathbf{J}_\mathcal{C}$ projects each entry of a graph signal onto the mean vector value of its respective cluster, its residual $\mathbf{Q}_{\partial\mathcal{P}^c}$ can be interpreted as the projection onto the respective entries deviation from its cluster mean value.

Lemma 8 (Bounding the error restricted to the boundary)**.** *The total variation of* $\mathbf{E}$ *restricted to the boundary verifies*

$$\|\mathbf{E}\|_{\partial\mathcal{P}} \leq w(\partial\mathcal{P}) \left(\sqrt{2} \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_\mathcal{G}(\mathcal{C})} \left\|\overline{\mathbf{E}}_\mathcal{P}\right\|_F + 2 \frac{\|\mathbf{E}\|_{\partial\mathcal{P}^c}}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_\mathcal{G}(\mathcal{C})}} \right) \tag{C.11}$$

Proof. The proof relies on a decomposition of the $\|\mathbf{E}\|_{\partial\mathcal{P}}$ term from Proposition 2. We have

$$\begin{aligned}
\|\mathbf{E}\|_{\partial\mathcal{P}} &= \left\| \sum_{\mathcal{C}\in\mathcal{P}} \mathbf{J}_\mathcal{C}\mathbf{E} + \mathbf{Q}_\mathcal{C}\mathbf{E} \right\|_{\partial\mathcal{P}} \\
&= \left\| \overline{\mathbf{E}}_\mathcal{P} + \mathbf{B}^\dagger_{\partial\mathcal{P}^c} \mathbf{B}_{\partial\mathcal{P}^c}\mathbf{E} \right\|_{\partial\mathcal{P}} \\
&\leq \left\|\overline{\mathbf{E}}_\mathcal{P}\right\|_{\partial\mathcal{P}} + \left\|\mathbf{B}^\dagger_{\partial\mathcal{P}^c} \mathbf{B}_{\partial\mathcal{P}^c}\mathbf{E}\right\|_{\partial\mathcal{P}}
\end{aligned} \tag{C.12}$$

where $\overline{\mathbf{E}}_{\mathcal{P}}$ is obtained by setting the error signal on every cluster to its mean. For the first term on the right-hand side, let us denote by $\boldsymbol{\varepsilon}_{\mathcal{C}}$ the value of any row of $\overline{\mathbf{E}}_{\mathcal{P}}$ belonging to cluster $\mathcal{C}$, which is equal to the mean of errors $\mathbf{E}$ over that cluster. Also, we denote by $(\overline{\mathbf{E}}_{\mathcal{P}})_{\partial\mathcal{P}}$ the signal obtained from $\overline{\mathbf{E}}_{\mathcal{P}}$ by setting its rows corresponding to nodes that are not adjacent to any edge in the boundary $\partial\mathcal{P}$ to zeros. Also, let $\partial_v\mathcal{C}$ denote the inner boundary of set of nodes $\mathcal{C}$,i.e. nodes of $\mathcal{C}$ that connect it to its complementary. Then it holds that:

$$\begin{aligned}
\left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_{\partial\mathcal{P}} &= \left\|\mathbf{B}_{\partial\mathcal{P}}\overline{\mathbf{E}}_{\mathcal{P}}\right\|_{2,1} \\
&= \left\|\mathbf{B}_{\partial\mathcal{P}}(\overline{\mathbf{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\|_{2,1} \\
&\leq \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\left\|(\overline{\mathbf{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\| \quad \text{(by Proposition 1)} \\
&\leq \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\left\|(\overline{\mathbf{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\|_{F} \\
&= \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}|\partial_v\mathcal{C}|\|\boldsymbol{\varepsilon}_{\mathcal{C}}\|^2} \\
&= \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}\frac{|\partial_v\mathcal{C}|}{|\mathcal{C}|}|\mathcal{C}|\|\boldsymbol{\varepsilon}_{\mathcal{C}}\|^2} \\
&\leq \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|\|\boldsymbol{\varepsilon}_{\mathcal{C}}\|^2} \\
&= \sqrt{2}w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}\mathcal{C}}\left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_{F} \qquad \text{(C.13)}
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\left\|\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c}\mathbf{B}_{\partial\mathcal{P}^c}\mathbf{E}\right\|_{\partial\mathcal{P}} &= \left\|\mathbf{B}_{\partial\mathcal{P}}\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c}\mathbf{B}_{\partial\mathcal{P}^c}\mathbf{E}\right\|_{2,1} \\
&\leq \left\|\mathbf{B}_{\partial\mathcal{P}}\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c}\right\|_{\infty,1}\|\mathbf{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|\mathbf{B}_{\partial\mathcal{P}}\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c}\right\|_{F}\|\mathbf{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|(\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c})^{\top}\mathbf{B}^{\top}_{\partial\mathcal{P}}\right\|_{F}\|\mathbf{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|\mathbf{B}^{\top}_{\partial\mathcal{P}}\right\|_{2,1}\sqrt{\left\|\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c}(\mathbf{B}^{\dagger}_{\partial\mathcal{P}^c})^{\top}\right\|_{\infty,\infty}}\|\mathbf{E}\|_{\partial\mathcal{P}^c} \quad \text{(by Proposition 1)} \\
&= \frac{\left\|\mathbf{B}^{\top}_{\partial\mathcal{P}}\right\|_{1,1}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\|\mathbf{E}\|_{\partial\mathcal{P}^c}\,.
\end{aligned}$$

$$= 2\frac{w(\partial\mathcal{P})}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\|\mathbf{E}\|_{\partial\mathcal{P}^c}\,. \tag{C.14}$$

The result is obtained by combining Equations (C.12), (C.13), and (C.14). □

Theorem 7 (Theorem 2.1 of Hsu *et al.* [70])**.** *At time step t, let $\mathbf{A}\in\mathbb{R}^{b\times t}$ where $b\in\mathbb{N}^*$, and let $\mathbf{v}\in\mathbb{R}^t$ be a random vector such that for some $\sigma\geq 0$, we have*

$$\mathbb{E}\left[\exp(\langle\mathbf{u},\mathbf{v}\rangle)\right]\leq\exp(\|\mathbf{u}\|^2\frac{\sigma^2}{2})\quad\forall\mathbf{u}\in\mathbb{R}^t.$$

Then for any $\delta\in(0,1)$, we have with a probability at least $1-\delta$:

$$\|\mathbf{A}\mathbf{v}\|^2\leq\sigma^2\left(\|\mathbf{A}\|_F^2+2\left\|\mathbf{A}^\top\mathbf{A}\right\|_F\sqrt{\log\frac{1}{\delta}}+2\|\mathbf{A}\|^2\log\frac{1}{\delta}\right).$$

Lemma 9 (Empirical process bound)**.** *Let $\mathbf{X}_m\in\mathbb{R}^{|\mathcal{T}_m|\times d}$ denotes the matrix of collected context vectors for task $m\in\mathcal{V}$, then, given collected context matrices $\{\mathbf{X}_m\}_{m\in\mathcal{V}}$, for any $\delta\in(0,1)$ we have with probability of at least $1-\delta$:*

$$\|\mathbf{K}\|_F\leq\frac{\alpha_\delta(t)}{\alpha_0}t,$$

where

$$\alpha_\delta(t):=\frac{\alpha_0\sigma}{t}\sqrt{t+2\sqrt{\sum_{m\in\mathcal{V}}\left|\mathcal{T}_m(t)\right|^2\log\frac{1}{\delta}}+2\max_{m\in\mathcal{V}}\left|\mathcal{T}_m(t)\right|\log\frac{1}{\delta}},\tag{C.15}$$

Proof. We recall that $\mathbf{K}\in\mathbb{R}^{t\times d}$ is the matrix obtained by stacking the row vectors $\boldsymbol{\eta}_m^\top\mathbf{X}_m$ vertically. On the one hand, we have

$$\|\mathbf{K}\|_F^2=\sum_{m\in\mathcal{V}}\left\|\mathbf{X}_m^\top\boldsymbol{\eta}_m\right\|^2=\left\|\mathbf{X}_\mathcal{V}^\top\boldsymbol{\eta}\right\|^2,\tag{C.16}$$

where $\mathbf{X}_\mathcal{V}:=\mathrm{diag}(\mathbf{X}_1,\cdots,\mathbf{X}_{|\mathcal{V}|})\in\mathbb{R}^{t\times d|\mathcal{V}|}$.

On the other one, for any $\mathbf{u}=(u_1,\cdots,u_t)\in\mathbb{R}^t$, denoting $P(t):=\exp(\sum_{\tau=1}^t u_\tau\eta_\tau)$, we have

$$\begin{aligned}
\mathbb{E}\left[P(t)\right]&=\mathbb{E}\left[\mathbb{E}\left[\exp\{u_t\eta_t\}P(t-1)|\mathcal{F}_{t-1}\right]\right]\quad\text{(by the law of total expectation)}\\
&=\mathbb{E}\left[P(t-1)\mathbb{E}\left[\exp(u_t\eta_t)|\mathcal{F}_{t-1}\right]\right]\quad\text{(because }\{\eta_s\}_{s=1}^{t-1}\text{ are }\mathcal{F}_{t-1}\text{ measurable.)}\\
&\leq\exp(\frac{1}{2}\sigma^2u_t^2)\mathbb{E}\left[P(t-1)\right]\quad\text{(by the conditional subgaussianity assumption)}
\end{aligned}$$

$$\leq \prod_{s=1}^{t} \exp(\frac{1}{2}\sigma^2 u_s^2) \quad \text{(by induction)}$$
$$= \exp(\frac{1}{2}\sigma^2 \|\mathbf{u}\|^2). \tag{C.17}$$

From Equations (C.16) and (C.17), we can apply Theorem 7 to matrix $\mathbf{X}_{\mathcal{V}}$ and random vector $\boldsymbol{\eta}$, which implies that with a probability at least $1-\delta$, we have

$$\|\mathbf{X}_{\mathcal{V}}\boldsymbol{\eta}\| \leq \sigma \sqrt{\mathrm{Tr}(\sum_{m\in\mathcal{V}} \mathbf{A}_m) + 2\sqrt{\sum_{m\in\mathcal{V}} \|\mathbf{A}_m\|_F^2 \log\frac{1}{\delta}} + 2\max_{m\in\mathcal{V}} \|\mathbf{A}_m\| \log\frac{1}{\delta}},$$

where we used the following equalities

$$\|\mathbf{X}_{\mathcal{V}}\|_F = \sum_{m\in\mathcal{V}} \mathrm{Tr}(\mathbf{A}_m),$$
$$\|\mathbf{X}_{\mathcal{V}}\|^2 = \max_{m\in\mathcal{V}} \|\mathbf{A}_m\|,$$
$$\left\|\mathbf{X}_{\mathcal{V}}\mathbf{X}_{\mathcal{V}}^\top\right\|_F^2 = \left\|\mathbf{X}_{\mathcal{V}}^\top\mathbf{X}_{\mathcal{V}}\right\|_F^2 = \sum_{m\in\mathcal{V}} \|\mathbf{A}_m\|_F^2.$$

To arrive to the the statement of the theorem, we use the fact that the context vectors have Euclidean norms of at most 1.

□

Proposition 4 (Probabilistic inequality)**.** *With a probability at least* $1-\delta$*, we have*

$$\frac{1}{2t\alpha}\sum_{m\in\mathcal{V}} \|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2 + a_1(\mathcal{G},\boldsymbol{\Theta})\|\mathbf{E}\|_{\partial\mathcal{P}^c} \leq a_2(\mathcal{G},\boldsymbol{\Theta})\left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_F + (1-\kappa)\|\mathbf{E}\|_{\partial\mathcal{P}}, \tag{C.18}$$

where $0 \leq \kappa < \frac{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}}{2w(\partial\mathcal{P})}$*,* $\frac{1}{\alpha_0} < \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\kappa w(\partial\mathcal{P})$ *and*

$$a_1(\mathcal{G},\boldsymbol{\Theta}) = 1 - \frac{\frac{1}{\alpha_0} + 2\kappa w(\partial\mathcal{P})}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \tag{C.19}$$
$$a_2(\mathcal{G},\boldsymbol{\Theta}) = \frac{1}{\alpha_0} + \sqrt{2}\kappa w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}. \tag{C.20}$$

Proof. The proof is a combination of the results of Lemmas 7 to 9. We have

$$\frac{1}{2t\alpha_\delta}\sum_{m\in\mathcal{V}} \|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2 + \|\mathbf{E}\|_{\partial\mathcal{P}^c} \leq \frac{1}{t\alpha_\delta}\langle\mathbf{K},\mathbf{E}\rangle + \|\mathbf{E}\|_{\partial\mathcal{P}} \quad \text{(by Lemma 7)}$$

$$
\begin{aligned}
&\leq \frac{1}{\alpha_0}\|\mathbf{E}\|_F + \kappa\|\mathbf{E}\|_{\partial\mathcal{P}} + (1-\kappa)\|\mathbf{E}\|_{\partial\mathcal{P}} \quad \text{(by Lemma 9)} \\
&\leq \frac{\left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_F}{\alpha_0} + \frac{\|\mathbf{E}\|_{\partial\mathcal{P}^c}}{\alpha_0 \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} + \kappa w(\partial\mathcal{P}) \left(\sqrt{2} \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})} \left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_F + 2\frac{\|\mathbf{E}\|_{\partial\mathcal{P}^c}}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \right) + \\
&+ (1-\kappa)\|\mathbf{E}\|_{\partial\mathcal{P}},
\end{aligned}
$$

where the last line is an application of Lemma 8. Grouping the terms by the type of norm applied to $\mathbf{E}$ finishes the proof. □

Theorem 8 (Oracle inequality, generalization of Theorem 5)**.** *Assume that the RE assumption holds for the empirical multi-task Gram matrix with constants $\phi > 0$ and*

$$
\kappa \in \left[0, \frac{1}{2w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}\right).
$$

Suppose that $\max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt$ for some $b > 0$. Then, with a probability at least $1-\delta(t)$, we have

$$
\left\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(t)\right\|_F \leq 2\frac{\sigma}{\phi^2\sqrt{t}} f(\mathcal{G},\boldsymbol{\Theta}) \sqrt{1 + 2b\sqrt{|\mathcal{V}| \log\frac{1}{\delta(t)}} + 2b\log\frac{1}{\delta(t)}},
$$

where

$$
f(\mathcal{G},\boldsymbol{\Theta}) \coloneqq \alpha_0 \left(a_2(\mathcal{G},\boldsymbol{\Theta}) + \sqrt{2}\mathbb{1}_{\leq 1}(\kappa) w(\partial\mathcal{P})\right) \left(\frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + \sqrt{2}\mathbb{1}_{\leq 1}(\kappa) w(\partial\mathcal{P})}{a_1(\mathcal{G},\boldsymbol{\Theta}) \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} + 1 \right).
$$

Proof. Using the previously established results, we obtain

$$
\begin{aligned}
&\frac{1}{2t}\sum_{m\in\mathcal{V}} \|\mathbf{X}_m\boldsymbol{\varepsilon}_m\|^2 + \alpha\|\mathbf{E}\|_{\partial\mathcal{P}^c} \\
&\leq \alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\mathbf{E}_{\mathcal{P}}\|_F + \alpha_\delta(1-\kappa)^+\|\mathbf{E}\|_{\partial\mathcal{P}} \quad \text{(by Proposition 4)} \\
&= \alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\mathbf{E}_{\mathcal{P}}\|_F + \alpha_\delta(1-\kappa)^+ \left\|\mathbf{B}_{\partial\mathcal{P}}\mathbf{B}_{\partial\mathcal{P}}^\dagger\mathbf{B}_{\partial\mathcal{P}}\mathbf{E}\right\|_{2,1} \text{(properties of pseudo-inverse)} \\
&\leq \alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\mathbf{E}_{\mathcal{P}}\|_F + \alpha_\delta\|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}\,\mathbb{1}_{\leq 1}(\kappa)(1-\kappa)^+ \left\|\mathbf{B}_{\partial\mathcal{P}}^\dagger\mathbf{B}_{\partial\mathcal{P}}\mathbf{E}\right\| \quad \text{(by Proposition 1)} \\
&\leq \alpha_\delta(a_2(\boldsymbol{\Theta},\mathcal{G}) + \mathbb{1}_{\leq 1}(\kappa)\sqrt{2}w(\partial\mathcal{P}))\|\mathbf{E}\|_{RE} \quad \text{(by definition of the } \|\|_{RE} \text{ norm)} \\
&\leq \alpha\frac{a_2(\boldsymbol{\Theta},\mathcal{G}) + \mathbf{1}_{\leq 1}(\kappa)\sqrt{2}w(\partial\mathcal{P})}{\phi\sqrt{t}} \sqrt{\sum_{m\in\mathcal{V}} \|\boldsymbol{\varepsilon}_m\|^2_{\mathbf{A}_m}} \quad \text{(using the RE assumption)}
\end{aligned}
$$

$$\leq \frac{\beta \alpha_\delta^2 (a_2(\boldsymbol{\Theta}, \mathcal{G}) + \mathbb{1}_{\leq 1}(\kappa) \|\mathbf{B}_{\partial \mathcal{P}}\|_{2,1})^2}{2\phi^2} + \frac{1}{2\beta t} \sum_{m \in \mathcal{V}} \|\mathbf{X}_m \boldsymbol{\varepsilon}_m\|^2, \tag{C.21}$$

where the last inequality holds for any $\beta > 0$, and is a consequence of the property that $uv \leq \frac{u^2 + v^2}{2}$ for any $u, v \in \mathbb{R}$.

As a result, we can bound the norm of $\mathbf{Q}_{\partial \mathcal{P}^c} \mathbf{E}$ as follows:

$$\begin{aligned}
\left\|\mathbf{Q}_{\partial \mathcal{P}^c} \mathbf{E}\right\|_F &= \left\|\mathbf{B}^\dagger_{\partial \mathcal{P}^c} \mathbf{B}_{\partial \mathcal{P}^c} \mathbf{E}\right\|_F \\
&\leq \sqrt{\left\|\mathbf{L}^\dagger_{\partial \mathcal{P}^c}\right\|_{\infty,\infty}} \|\mathbf{E}\|_{\partial \mathcal{P}^c} \\
&\leq \frac{2\alpha_\delta (a_2(\boldsymbol{\Theta}, \mathcal{G}) + \mathbb{1}_{\leq 1}(\kappa) \|\mathbf{B}_{\partial \mathcal{P}}\|_{2,1})^2}{\phi^2 a_1(\boldsymbol{\Theta}, \mathcal{G}) \min_{\mathcal{C} \in \mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \quad \text{(Equation (C.21) with } \beta = 1\text{).}
\end{aligned} \tag{C.22}$$

We can also bound the norm of $\overline{\mathbf{E}}_{\mathcal{P}}$ as follows:

$$\begin{aligned}
\left\|\overline{\mathbf{E}}_{\mathcal{P}}\right\|_F^2 &\leq \frac{1}{t\phi^2} \sum_{m \in \mathcal{V}} \|\mathbf{X}_m \boldsymbol{\varepsilon}_m\|^2 \quad \text{(by RE assumption on empirical multi-task Gram matrix)} \\
&\leq \frac{4\alpha_\delta^2 (a_2(\boldsymbol{\Theta}, \mathcal{G}) + \mathbb{1}_{\leq 1}(\kappa) \|\mathbf{B}_{\partial \mathcal{P}}\|_{2,1})^2}{\phi^4} \quad \text{(by Equation (C.21) with } \beta = 2\text{).}
\end{aligned} \tag{C.23}$$

The result is then obtained by combining Equations (C.22) and (C.23) along with using the fact that $\mathbf{E} = \overline{\mathbf{E}}_{\mathcal{P}} + \mathbf{Q}_{\partial \mathcal{P}^c} \mathbf{E}$ and the expressions of $a_1(\boldsymbol{\Theta}, \mathcal{G})$ and $a_2(\boldsymbol{\Theta}, \mathcal{G})$, and bounding $\alpha_\delta(t)$ as follows:

$$\begin{aligned}
\frac{\alpha_\delta(t)^2}{\alpha_0^2} &= \frac{\sigma^2}{t^2} \left(\sum_{m \in \mathcal{V}} \|\mathbf{X}_m\|_F^2 + 2\sqrt{\sum_{m \in \mathcal{V}} \left\|\mathbf{X}_m \mathbf{X}_m^\top\right\|_F^2 \log \frac{1}{\delta}} + 2 \max_{m \in \mathcal{V}} \|\mathbf{X}_m\|^2 \log \frac{1}{\delta} \right) \\
&\leq \frac{\sigma^2}{t^2} \left(t + 2\sqrt{\sum_{m \in \mathcal{V}} \left|\mathcal{T}_m(t)\right|^2 \log \frac{1}{\delta}} + 2 \max_{m \in \mathcal{V}} \left|\mathcal{T}_m(t)\right| \log \frac{1}{\delta} \right) \\
&\leq \frac{\sigma^2}{t^2} \left(t + 2t\sqrt{\log \frac{1}{\delta}} + 2t \log \frac{1}{\delta} \right) \\
&\leq 2\frac{\sigma^2}{t} \left(1 + \sqrt{\log \frac{1}{\delta}} \right)^2
\end{aligned}$$

□

Proposition 5 (Optimal value of α_0)**.** *Assume* $\kappa \geq 1$. *Let*

$$h_1 := \sqrt{2}\kappa w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})},$$

$$h_2 := \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\kappa w(\partial\mathcal{P}),$$

and $r := \frac{h_1}{h_2}$. *Choosing*

$$\alpha_0 := \frac{1}{h_2}\left(1+\sqrt{1+\frac{h_2}{h_1}}\right) \tag{C.24}$$

minimizes $f(\boldsymbol{\Theta},\mathcal{G})$. *In this case, we have* $f(\boldsymbol{\Theta},\mathcal{G}) = \mathcal{O}(1+2\sqrt{r})$ *as r vanishes.*

Proof. Using the definition of $f(\boldsymbol{\Theta},\mathcal{G})$ from Definition 2, it is easy to see that $f(\boldsymbol{\Theta},\mathcal{G}) = \alpha_0(h_1+h_2)\frac{\alpha_0 h_1+1}{\alpha_0 h_2-1}$. Deriving with respect to α_0 and solving yields the stated optimal value. Injecting this value into $f(\boldsymbol{\Theta},\mathcal{G})$, we get

$$\min_{\alpha_0>0} f(\boldsymbol{\Theta},\mathcal{G}) = \frac{h_1+h_2}{h_2}\left(1+\sqrt{1+\frac{h_2}{h_1}}\right)\frac{1+\frac{h_1}{h_2}\left(1+\sqrt{1+\frac{h_2}{h_1}}\right)}{\sqrt{1+\frac{h_2}{h_1}}} \tag{C.25}$$

$$= (1+r)\left(1+\sqrt{1+\frac{1}{r}}\right)\frac{1+r\left(1+\sqrt{1+\frac{1}{r}}\right)}{\sqrt{1+\frac{1}{r}}} \tag{C.26}$$

$$\leq \frac{1+r}{\sqrt{r}}(\sqrt{r}+1+\frac{r}{2})\sqrt{r}(1+\sqrt{r}(\sqrt{r}+\frac{r}{2}+1)) \tag{C.27}$$

$$= (1+r)(1+\sqrt{r}+\frac{r}{2})(1+\sqrt{r}+r+\frac{r\sqrt{r}}{2}) \tag{C.28}$$

□

C.2.3 Inheriting the RE condition from the true to the empirical data Gram matrix

C.2.3.1 From the adapted to the empirical multi-task Gram matrix

Lemma 10 (Bounding a quadratic form using projections)**.** *Let* $\mathbf{M}_1,\cdots,\mathbf{M}_p \in \mathbb{R}^{d\times d}$ *be symmetric matrices, and let* $\mathbf{J} := \frac{1}{p}\mathbf{1}\mathbf{1}^\top$, *and* $\mathbf{Q} = \mathbf{I}-\mathbf{J}$. *Then, for any* $\mathbf{Z} \in \mathbb{R}^{p\times d}$ *with*

rows $\{\mathbf{z}_i\}_{i=1}^p$*, we have:*

$$\left|\sum_{i=1}^p \mathbf{z}_i^\top \mathbf{M}_i \mathbf{z}_i\right| \leq \frac{1}{p}\left\|\sum_{i=1}^p \mathbf{M}_i\right\| \|\mathbf{Z}\|_{\mathbf{J}}^2 + 2\sqrt{\left\|\frac{1}{p}\sum_{i=1}^p \mathbf{M}_i^2\right\|}\|\mathbf{Z}\|_{\mathbf{Q}}\|\mathbf{Z}\|_{\mathbf{J}} + \max_{1\leq i\leq p}\|\mathbf{M}_i\|\|\mathbf{Z}\|_{\mathbf{Q}}^2$$

Proof. We have

$$\begin{aligned}\left|\sum_{i=1}^p \mathbf{z}_i^\top \mathbf{M}_i \mathbf{z}_i\right| &= \left|\sum_{i=1}^p \bar{\mathbf{z}}^\top \mathbf{M}_i \bar{\mathbf{z}} + 2\sum_{i=1}^p (\mathbf{z}_i - \bar{\mathbf{z}})^\top \mathbf{M}_i \bar{\mathbf{z}} + \sum_{i=1}^p (\mathbf{z}_i - \bar{\mathbf{z}})^\top \mathbf{M}_i (\mathbf{z}_i - \bar{\mathbf{z}})\right| \\ &\leq \left|\bar{\mathbf{z}}^\top \sum_{i=1}^p \mathbf{M}_i \bar{\mathbf{z}}\right| + 2\left|\sum_{i=1}^p \mathbf{e}_i^\top \mathbf{Q}\mathbf{Z}\mathbf{M}_i \bar{\mathbf{z}}\right| + \left|\sum_{i=1}^p \mathbf{e}_i^\top \mathbf{Q}\mathbf{Z}\mathbf{M}_i \mathbf{Z}^\top \mathbf{Q}\mathbf{e}_i\right| \end{aligned} \tag{C.29}$$

where we used the fact that $\mathbf{z}_i - \bar{\mathbf{z}} = \mathbf{Z}^\top \mathbf{e}_i - \mathbf{Z}^\top \mathbf{J}\mathbf{e}_i = \mathbf{Z}^\top \mathbf{Q}\mathbf{e}_i$.

Let us now examine every term on the right-hand side of Equation (C.29). For the first term, we have

$$\left|\bar{\mathbf{z}}^\top \sum_{i=1}^p \mathbf{M}_i \bar{\mathbf{z}}\right| \leq \left\|\sum_{i=1}^p \mathbf{M}_i\right\| \|\bar{\mathbf{z}}\|^2 = \left\|\frac{1}{p}\sum_{i=1}^p \mathbf{M}_i\right\| \|\mathbf{Z}\|_{\mathbf{J}}^2. \tag{C.30}$$

For the second term, we have

$$\begin{aligned}\left|\sum_{i=1}^p \mathbf{e}_i^\top \mathbf{Q}\mathbf{Z}\mathbf{M}_i \bar{\mathbf{z}}\right| &\leq \left\|\sum_{i=1}^p \mathbf{M}_i \mathbf{Z}^\top \mathbf{Q}\mathbf{e}_i\right\| \|\bar{\mathbf{z}}\| \\ &= \left\|\sum_{i=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i)\,\mathrm{vec}(\mathbf{Z}^\top \mathbf{Q})\right\| \|\bar{\mathbf{z}}\| \\ &\leq \left\|\sum_{i=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i)\right\| \left\|\mathrm{vec}(\mathbf{Z}^\top \mathbf{Q})\right\| \|\bar{\mathbf{z}}\| \\ &= \left\|\sum_{i=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i)\right\| \|\mathbf{Q}\mathbf{Z}\|_F \|\bar{\mathbf{z}}\| \\ &= \sqrt{\left\|(\sum_{i=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i))^\top \sum_{i=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i)\right\|} \|\mathbf{Q}\mathbf{Z}\|_F \|\bar{\mathbf{z}}\| \\ &= \sqrt{\left\|\sum_{i=1}^p \sum_{j=1}^p (\mathbf{e}_i^\top \otimes \mathbf{M}_i))(\mathbf{e}_j \otimes \mathbf{M}_j)\right\|} \|\mathbf{Q}\mathbf{Z}\|_F \|\bar{\mathbf{z}}\| \end{aligned}$$

$$= \sqrt{\left\| \sum_{i=1}^{p} \sum_{j=1}^{p} (\mathbf{e}_i^\top \mathbf{e}_j \otimes \mathbf{M}_i \mathbf{M}_j) \right\|} \|\mathbf{QZ}\|_F \|\bar{\mathbf{z}}\|$$

$$= \sqrt{\left\| \sum_{i=1}^{p} \mathbf{M}_i^2 \right\|} \|\mathbf{QZ}\|_F \|\bar{\mathbf{z}}\| . \tag{C.31}$$

Finally, for the last term, we have

$$\begin{aligned} \left| \sum_{i=1}^{p} \mathbf{e}_i^\top \mathbf{QZM}_i \mathbf{Z}^\top \mathbf{Qe}_i \right| &\leq \sum_{i=1}^{p} \|\mathbf{M}_i\| \left\| \mathbf{Z}^\top \mathbf{Qe}_i \right\|^2 \\ &\leq \max_{1 \leq i \leq p} \|\mathbf{M}_i\| \sum_{i=1}^{p} \left\| \mathbf{Z}^\top \mathbf{Qe}_i \right\|^2 \\ &= \max_{1 \leq i \leq p} \|\mathbf{M}_i\| \|\mathbf{QZ}\|_F^2 . \end{aligned} \tag{C.32}$$

Combining Equations (C.30), (C.31), and (C.32) yields the result. □

We also define an operator norm that is induced by the $\|\|_{\mathrm{RE}}$.

Definition 5 ((RE,$\mathcal{S}$)-induced operator norm)**.** *Let $\{\mathbf{M}_m\}_{m \in \mathcal{V}} \subseteq \mathbb{R}^{d \times d}$ be symmetric matrices associated to the graph nodes $\mathcal{V}$, and let $\mathbf{M}_{\mathcal{V}} \coloneqq \operatorname{diag}\left(\mathbf{M}_1, \cdots, \mathbf{M}_{|\mathcal{V}|}\right) \in \mathbb{R}^{d|\mathcal{V}| \times d|\mathcal{V}|}$. For any cluster $\mathcal{C} \in \mathcal{P}$, let the cluster mean and mean of squares associated to those matrices be given by*

$$\overline{\mathbf{M}}_{\mathcal{C}} \coloneqq \frac{1}{|\mathcal{C}|} \sum_{m \in \mathcal{C}} \mathbf{M}_m, \qquad \overline{\mathbf{M}^2}_{\mathcal{C}} \coloneqq \frac{1}{|\mathcal{C}|} \sum_{m \in \mathcal{C}} \mathbf{M}_m^2 .$$

The RE-induced operator norm of $\mathbf{M}_{\mathcal{V}}$ is defined as

$$\|\mathbf{M}\|_{\mathrm{RE},\mathcal{S}} \coloneqq \max_{\mathcal{C} \in \mathcal{P}} \left\| \overline{\mathbf{M}}_{\mathcal{C}} \right\| \vee \sqrt{\min_{\mathcal{C} \in \mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C} \in \mathcal{P}} \left\| \overline{\mathbf{M}^2}_{\mathcal{C}} \right\|} \vee \min_{\mathcal{C} \in \mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{m \in \mathcal{V}} \|\mathbf{M}_m\| . \tag{C.33}$$

C.2.3.2 Linking the adapted to the empirical Gram

We first start by establishing that given the closeness of two PSD matrices in a certain sense, the RE condition can be transferred between them.

Proposition 6 (Restricted spectral norm)**.** *Let $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ verifying*

$$a_1(\mathcal{G}, \boldsymbol{\Theta}) \|\mathbf{Z}\|_{\partial \mathcal{P}^c} \leq a_2(\mathcal{G}, \boldsymbol{\Theta}) \left\| \overline{\mathbf{Z}}_{\mathcal{P}} \right\|_F + (1 - \kappa)^+ \|\mathbf{Z}\|_{\partial \mathcal{P}}$$

Let $\{\mathbf{M}_m\}_{m\in\mathcal{V}} \subseteq \mathbb{R}^{d\times d}$ *be symmetric matrices associated to the graph nodes* $\mathcal{V}$*, and let* $\mathbf{M}_{\mathcal{V}} \coloneqq \operatorname{diag}(\mathbf{M}_1,\cdots,\mathbf{M}_{|\mathcal{V}|}) \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$*. Then we have:*

$$\left|\sum_{m\in\mathcal{V}} \mathbf{z}_m^\top \mathbf{M}_m \mathbf{z}_m\right| \leq \|\mathbf{M}\|^2_{\mathrm{RE},\mathcal{S}} \left(1 + \frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+ \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^2 \|\mathbf{Z}\|^2_{RE}. \tag{C.34}$$

Proof. For any cluster $\mathcal{C}$, we denote by $\mathbf{B}_{\mathcal{C}}$ the incidence matrix obtained by setting the rows of $\mathbf{B}$ outside the edges linking nodes in $\mathcal{C}$ to null vectors. The latter's nullspace is the span of the vector $\mathbf{1}_{\mathcal{C}}$ having coordinates 1 at nodes in $\mathcal{C}$ and zeros elsewhere. Hence, the projector onto the orthogonal of $\mathbf{1}_{\mathcal{C}}$ is $\mathbf{Q}_{\mathcal{C}} \coloneqq \mathbf{B}_{\mathcal{C}}^\dagger \mathbf{B}_{\mathcal{C}}$.

On the one hand, for any signal $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}$ we have

$$\|\mathbf{Z}\|_{\partial\mathcal{P}^c} = \sum_{\mathcal{C}\in\mathcal{P}} \|\mathbf{B}_{\mathcal{C}}\mathbf{Z}\|_{2,1} \geq \sum_{\mathcal{C}\in\mathcal{P}} \frac{\left\|\mathbf{B}_{\mathcal{C}}^\dagger \mathbf{B}_{\mathcal{C}}\mathbf{Z}\right\|_F}{\sqrt{\left\|\mathbf{L}_{\mathcal{C}}^\dagger\right\|_{\infty,\infty}}} \geq \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} \sum_{\mathcal{C}\in\mathcal{P}} \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}}$$

Hence, by the proposition's assumptions, $\mathbf{Z}$ verifies

$$\begin{aligned}
\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} a_1(\mathcal{G},\boldsymbol{\Theta}) \sum_{\mathcal{C}\in\mathcal{P}} \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}} &\leq (a_2(\mathcal{G},\boldsymbol{\Theta}) \left\|\overline{\mathbf{Z}}_{\mathcal{P}}\right\|_F + (1-\kappa)\|\mathbf{Z}\|_{\partial\mathcal{P}}) \\
&\leq a_2(\mathcal{G},\boldsymbol{\Theta}) \left\|\overline{\mathbf{Z}}_{\mathcal{P}}\right\|_F + (1-\kappa)^+ \|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1} \left\|\mathbf{B}_{\partial\mathcal{P}}^\dagger \mathbf{B}_{\partial\mathcal{P}}\mathbf{Z}\right\| \\
&\leq (a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+ \|\mathbf{B}\|_{2,1}) \|\mathbf{Z}\|_{RE}
\end{aligned}$$

From Lemma 10, we have

$$\begin{aligned}
&\left|\sum_{m\in\mathcal{V}} \mathbf{z}_m^\top \mathbf{M}_m \mathbf{z}_m\right| \leq \sum_{\mathcal{C}\in\mathcal{P}} \left|\sum_{m\in\mathcal{C}} \mathbf{z}_m^\top \mathbf{M}_m \mathbf{z}_m\right| \\
&\leq \sum_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\mathbf{M}}_{\mathcal{C}}\right\| \|\mathbf{Z}\|^2_{\mathbf{J}_{\mathcal{C}}} + 2\sum_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}} \|\mathbf{Z}\|_{\mathbf{J}_{\mathcal{C}}} + \sum_{\mathcal{C}\in\mathcal{P}} \max_{m\in\mathcal{C}} \|\mathbf{M}_m\| \|\mathbf{Z}\|^2_{\mathbf{Q}_{\mathcal{C}}},
\end{aligned} \tag{C.35}$$

where we used Equation (C.1).

This allows us to bound every term in Equation (C.35). For the second term on the right-hand side, we have

$$\begin{aligned}
&\sum_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}} \|\mathbf{Z}\|_{\mathbf{J}_{\mathcal{C}}} \\
\leq &\max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} \left\|\overline{\mathbf{Z}}_{\mathcal{P}}\right\|_F \sqrt{\sum_{\mathcal{C}\in\mathcal{P}} \|\mathbf{Z}\|^2_{\mathbf{Q}_{\mathcal{C}}}}
\end{aligned}$$

$$\leq \frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-\frac{1}{2}}}{a_1(\mathcal{G},\boldsymbol{\Theta})} \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} (a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+ \|\mathbf{B}\|_{2,1}) \|\mathbf{Z}\|_{RE}^2 \tag{C.36}$$

As for the third term, we have

$$\begin{aligned}\sum_{\mathcal{C}\in\mathcal{P}} \max_{m\in\mathcal{C}} \|\mathbf{M}_m\| \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}}^2 &\leq \max_{m\in\mathcal{V}} \|\mathbf{M}_m\| \left(\sum_{\mathcal{C}\in\mathcal{P}} \|\mathbf{Z}\|_{\mathbf{Q}_{\mathcal{C}}}\right)^2 \\ &\leq \max_{m\in\mathcal{V}} \|\mathbf{M}_m\| \frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1}}{a_1(\mathcal{G},\boldsymbol{\Theta})^2} (a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+ \|\mathbf{B}\|_{2,1})^2 \|\mathbf{Z}\|_{RE}^2\end{aligned} \tag{C.37}$$

Consequently, denoting

$$v = \frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+ \|\mathbf{B}\|_{2,1}}{a_1(\mathcal{G},\boldsymbol{\Theta})},$$

and combining Equations (C.35), (C.36), and (C.37), we obtain

$$\begin{aligned}&\left|\sum_{m\in\mathcal{V}} \mathbf{z}_m^\top \mathbf{M}_m \mathbf{z}_m\right| \\ &\left(\max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\mathbf{M}}_{\mathcal{C}}\right\| + 2v \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} + v^2 \max_{i\in\mathcal{V}} \|\mathbf{M}_i\|\right) \|\mathbf{Z}\|_{RE}^2 \\ \leq &\left(\max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\mathbf{M}}_{\mathcal{C}}\right\| \vee \sqrt{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\mathbf{M}^2}_{\mathcal{C}}\right\|} \vee \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{i\in\mathcal{V}} \|\mathbf{M}_i\|\right) (1+v)^2 \|\mathbf{Z}\|_{RE}^2,\end{aligned}$$

which finishes the proof. □

Proposition 7 (Inheritance of a RE condition from a close matrix)**.** *Assume that the matrix* $\mathbf{V}_{\mathcal{V}}$ *verifies the RE condition with constant* $\phi > 0$*, and that* $\left\|\frac{\mathbf{A}_{\mathcal{V}}}{t} - \mathbf{V}_{\mathcal{V}}\right\|_{\text{op,RE}} \leq \gamma\phi^2$ *for some* $\gamma \in \left(0, \left(1 + \frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^{-2}\right)$*. Then* $\frac{\mathbf{A}_{\mathcal{V}}}{t}$ *verifies the RE condition with constant*

$$\hat{\phi} = \phi \sqrt{1 - \gamma\left(1 + \frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^2} \tag{C.38}$$

Proof. From Proposition 4, we know that

$$\begin{aligned}
\frac{1}{t}\varepsilon_{\mathcal{V}}^{\top}\mathbf{A}_{\mathcal{V}}\varepsilon_{\mathcal{V}} &= \frac{1}{|\mathcal{V}|}\varepsilon_{\mathcal{V}}^{\top}\mathbf{V}_{\mathcal{V}}\varepsilon_{\mathcal{V}} + \varepsilon_{\mathcal{V}}^{\top}\mathbf{\Delta}_{\mathcal{V}}\varepsilon_{\mathcal{V}} \\
&\geq \frac{1}{|\mathcal{V}|}\varepsilon_{\mathcal{V}}^{\top}\mathbf{V}_{\mathcal{V}}\varepsilon_{\mathcal{V}} - \left|\varepsilon_{\mathcal{V}}^{\top}\mathbf{\Delta}_{\mathcal{V}}\varepsilon_{\mathcal{V}}\right| \\
&\geq \left(\phi^2 - \max_{m\in\mathcal{V}}\|\mathbf{\Delta}_{\mathcal{V}}\|_{\text{op,RE}}\left(1+\frac{a_2(\mathcal{G},\mathbf{\Theta})+(1-\kappa)^+\|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}}{a_1(\mathcal{G},\mathbf{\Theta})}\right)^2\right)\|\mathbf{E}\|_{\text{RE}}^2 \\
&\geq \left(\phi^2 - \gamma\phi^2\left(1+\frac{a_2(\mathcal{G},\mathbf{\Theta})+(1-\kappa)^+\|\mathbf{B}_{\partial\mathcal{P}}\|_{2,1}}{a_1(\mathcal{G},\mathbf{\Theta})}\right)^2\right)\|\mathbf{E}\|_{\text{RE}}^2
\end{aligned}$$

where the third inequality is an applicaiton of Proposition 6. □

Theorem 9 (Matrix Freedman Inequality, Tropp [149])**.** *Consider a matrix martingale* $\{\mathbf{M}(t)\}_{t\geq 1}$ *with dimension* $d_1 \times d_2$*. Let* $\{\mathbf{N}(t)\}_{t\geq 1}$ *be the associated difference sequence. Assume that for some* $A > 0$*, we have* $\left\|\mathbf{N}(t)\right\| \leq A \quad \forall t \geq 1$ *almost surely. Define for any* $t \geq 1$*:*

$$\begin{aligned}
\mathbf{W}_{col}(t) &:= \sum_{\tau=1}^{t}\mathbb{E}\left[\mathbf{N}(\tau)\mathbf{N}(\tau)^{\top}|\mathcal{F}_{\tau-1}\right] \\
\mathbf{W}_{row}(t) &:= \sum_{\tau=1}^{t}\mathbb{E}\left[\mathbf{N}(\tau)^{\top}\mathbf{N}(\tau)|\mathcal{F}_{\tau-1}\right].
\end{aligned}$$

Then, for any $u, v > 0$*,*

$$\mathbb{P}\left[\exists t \geq 1; \left\|\mathbf{M}(t)\right\| \geq u \text{ and } \|\mathbf{W}_{col}\|(t) \vee \left\|\mathbf{W}_{row}(t)\right\| \leq v\right] \leq (d_1+d_2)\exp\left(-\frac{3u^2}{6v+2Au}\right)$$

Corollary 3. *Let* $\{\mathbf{N}(\tau)\}_{\tau=1}^{t}$ *by a sequence of matrices of dimension* $d_1 \times d_2$*, adapted to filtration* $\{\mathcal{F}_\tau\}_{\tau=1}^{t}$*. Let* $\{t_i\}_{i=1}^{N}$ *an increasing sequence with elements in* $[t]$ *for some* $N \leq t$*. Consider the sequence* $\{\mathbf{M}(n)\}_{\tau=1}^{N}$ *of random matrices defined by*

$$\mathbf{M}(n) = \sum_{i=1}^{n}\mathbf{N}(t_i) - \mathbb{E}\left[\mathbf{N}(t_i)|\mathcal{F}_{t_i-1}\right] \tag{C.39}$$

Then $\{\mathbf{M}(n)\}_{n=1}^{N}$ *is a martingale adapted to the filtration* $\{\mathcal{F}_{t_n}\}_{n=1}^{N}$*.*
Moreover, if $\left\|\mathbf{N}(\tau)\right\| \leq b \quad \forall \tau \in [t]$ *for some* $b > 0$*, then we have*

$$\mathbb{P}\left[\left\|\mathbf{M}(N)\right\| \geq u\right] \leq (d_1+d_2)\exp\left(-\frac{3u^2}{6Nb^2+2\sqrt{2}bu}\right). \tag{C.40}$$

Proof. We denote $\mathbb{E}\left[\cdot|\mathcal{F}_s\right]$ as $\mathbb{E}_s[\cdot]$ for any $s \in \mathbb{N}$. Also, let $\mathbf{C}(s) \coloneqq \mathbb{E}_{s-1}\left[\mathbf{N}(s)\right]$, which is $\mathcal{F}_{s-1}$-measurable by construction. We have for any $n \in [N]$,

$$\mathbb{E}_{t_{n-1}}\left[\mathbf{C}(t_n)\right] = \mathbb{E}_{t_{n-1}}\left[\mathbb{E}_{t_n-1}\left[\mathbf{N}(t_n)\right]\right] = \mathbb{E}_{t_{n-1}}\left[\mathbf{N}(t_n)\right] \tag{C.41}$$

$$\implies \mathbb{E}_{t_{n-1}}\left[\mathbf{N}(t_n) - \mathbf{C}(t_n)\right] = 0 \tag{C.42}$$

where the first equality is due to the tower rule since $\mathcal{F}_{t_{n-1}} \subset \mathcal{F}_{t_n-1}$. Also, we have for any $\tau \geq 1$

$$\left\|\mathbf{N}(\tau) - \mathbf{C}(\tau)\right\|^2 = \left\|(\mathbf{N}(\tau) - \mathbf{C}(\tau))^2\right\| \tag{C.43}$$

$$\leq \mathrm{Tr}((\mathbf{N}(\tau) - \mathbf{C}(\tau))^2) \tag{C.44}$$

$$= \mathrm{Tr}((\mathbf{N}(\tau) - \mathbf{C}(\tau))^2) \tag{C.45}$$

$$= \left\|\mathbf{N}(\tau)\right\|_F^2 - 2\mathrm{Tr}(\mathbf{C}(\tau)\mathbf{N}(\tau)) + \mathrm{Tr}(\mathbf{C}(\tau)^2) \tag{C.46}$$

$$\leq \left\|\mathbf{N}(\tau)\right\|_F^2 + \mathrm{Tr}(\mathbf{C}(\tau)^2) \leq 2b^2 \tag{C.47}$$

Hence $\mathbf{N}(\tau) - \mathbf{C}(\tau)$ is integrable for any $\tau \geq 1$. This shows that $\mathbf{M}(n)$ is a sequence of partial sums of matrix martingale differences, hence it is a matrix martingale.

The second part of the corollary statement is a consequence of Theorem 9. The boundedness of the sequence of martingale differences has already been established above. To verify the second requirement of the theorem, let us compute bounds on the norms of $\mathbf{W}_{\mathrm{col}}$ and $\mathbf{W}_{\mathrm{row}}$ from Theorem 9. Notice that the two matrices are equal since the difference sequence matrices $\mathbf{N}(t_s)$ are symmetric. Hence, for any $n \in [N]$, we have

$$\left\|\mathbf{W}_{\mathrm{col}}(N)\right\| \vee \left\|\mathbf{W}_{\mathrm{row}}(N)\right\| \leq \mathrm{Tr}(\mathbf{W}_{\mathrm{col}}(N)) \vee \mathrm{Tr}(\mathbf{W}_{\mathrm{row}}(N)) \tag{C.48}$$

$$= \mathrm{Tr}\left(\sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[(\mathbf{N}(t_n) - \mathbf{C}(t_n))^2\right]\right) \tag{C.49}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\left\|\mathbf{N}(t_n)\right\|_F^2\right] - \mathbb{E}_{t_n-1}\left[2\mathrm{Tr}(\mathbf{C}(t_n)\mathbf{N}(t_n))\right] + \mathrm{Tr}(\mathbf{C}(t_n)^2) \tag{C.50}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\left\|\mathbf{N}(t_n)\right\|_F^2\right] - \mathrm{Tr}(\mathbf{C}(t_n)^2) \tag{C.51}$$

$$\leq \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\left\|\mathbf{N}(t_n)\right\|_F^2\right] \leq Nb^2. \tag{C.52}$$

By Theorem 9, we have for any $u > 0$

$$2d\exp\left(-\frac{3u^2}{6Nb^2+2\sqrt{2}bu}\right) \geq \mathbb{P}\left[\exists n\geq 1; \left\|\mathbf{M}(n)\right\| \geq u \text{ and } \left\|\mathbf{W}_{\text{col}}(n)\right\| \leq Nb^2\right] \quad \text{(C.53)}$$

$$\geq \mathbb{P}\left[\left\|\mathbf{M}(N)\right\| \geq u \text{ and } \left\|\mathbf{W}_{\text{col}}(N)\right\| \leq Nb^2\right] \quad \text{(C.54)}$$

$$= \mathbb{P}\left[\left\|\mathbf{M}(N)\right\| \geq u\right] \quad \text{(C.55)}$$

where the last line holds because we showed that the inequality $\left\|\mathbf{W}_{\text{col}}(N)\right\| \leq Nb^2$ holds almost surely. □

Proposition 8 (Concentration of the empirical multi-task Gram matrix around the adapted one)**.** *Let $t \geq 1$, $b > 0$. Then we have:*

$$\mathbb{P}\left[\left\|\frac{\mathbf{A}_{\mathcal{V}}(t)}{t} - \mathbf{V}_{\mathcal{V}}\right\|_{\text{op,RE}} > \gamma \middle| \max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt\right] \leq$$
$$d(2|\mathcal{P}|\,e^{-A_1 t} + (|\mathcal{V}| + |\mathcal{P}|)e^{-A_2 t} + 2|\mathcal{V}|\,e^{-A_3 t}),$$

where

$$A_1 := \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} |\mathcal{C}|\, t}{6b + 2\sqrt{2}\gamma}$$

$$A_2 := \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t}{6b + 2\sqrt{2}\gamma\sqrt{\dfrac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}} |\mathcal{C}|}}}$$

$$A_3 := \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2 t}{6b + 2\sqrt{2}\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}$$

Proof. For $\gamma > 0$, let us define

$$\boldsymbol{\Delta}_m := \frac{\mathbf{A}_{\mathcal{V}}}{t} - \mathbf{V}_{\mathcal{V}} \quad \text{and } G_{\text{Gram},\gamma} := \left\{\frac{1}{t}\|\boldsymbol{\Delta}_{\mathcal{V}}\|_{\text{RE},\mathcal{S}} \leq \gamma\right\},$$

where $\boldsymbol{\Delta}_{\mathcal{V}}$ is block diagonal matrix formed by $\{\boldsymbol{\Delta}_m\}_{m\in\mathcal{V}}$. We also define $\overline{\boldsymbol{\Delta}}_{\mathcal{C}}$ and $\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}$ in the same pattern of Definition 5. We can express the complementary of this event as the disjunction of a finite number of events as follows:

$$G^c_{\text{Gram},\gamma} \quad \text{(C.56)}$$

$$= \left\{ \max_{\mathcal{C}\in\mathcal{P}} \left\| \overline{\boldsymbol{\Delta}}_{\mathcal{C}} \right\| \vee \sqrt{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C}\in\mathcal{P}} \left\| \overline{\boldsymbol{\Delta}^2}_{\mathcal{C}} \right\|} \vee \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{m\in\mathcal{V}} \|\boldsymbol{\Delta}_m\| > t\gamma \right\} \tag{C.57}$$

$$= \bigcup_{\mathcal{C}\in\mathcal{P}} \left\{ \left\| \overline{\boldsymbol{\Delta}}_{\mathcal{C}} \right\| > t\gamma \right\} \cup \bigcup_{\mathcal{C}\in\mathcal{P}} \left\{ \left\| \overline{\boldsymbol{\Delta}^2}_{\mathcal{C}} \right\| > t^2\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) \right\} \cup \bigcup_{m\in\mathcal{V}} \left\{ \|\boldsymbol{\Delta}_m\| > t\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) \right\} \tag{C.58}$$

The first and third event can be bounded by considering the sequence $\mathbf{x}\mathbf{x}^\top(\tau)$ adapted to the filtration $\{\mathcal{F}_\tau\}$, verifying $\left\| \mathbf{x}\mathbf{x}^\top(\tau) \right\| \leq 1$.

Bounding the probability of the first event Let $\mathcal{C} \in \mathcal{P}$ be a cluster. By definition, we have

$$\begin{aligned} |\mathcal{C}|\,\overline{\boldsymbol{\Delta}}_{\mathcal{C}}(t) &= \sum_{m\in\mathcal{C}} \sum_{\tau\in\mathcal{T}_m(t)} \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right] \\ &= \sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right] \end{aligned}$$

We will apply Corollary 3 for the sequence of time indices in $\mathcal{C}$, *i.e.* $\bigcup_{m\in\mathcal{V}} \mathcal{T}_m(t)$. Hence $|\mathcal{C}|\,\overline{\boldsymbol{\Delta}_{\mathcal{C}}}$ is a martingale sequence, and we have

$$\begin{aligned} \mathbb{P}\left[\left\| \overline{\boldsymbol{\Delta}}_{\mathcal{C}}(t) \right\| > \gamma t \middle| \max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt \right] &\leq 2d \exp\left(\frac{-3\gamma^2|\mathcal{C}|^2t^2}{6\sum_{m\in\mathcal{C}}\left|\mathcal{T}_m(t)\right| + 2\sqrt{2}\gamma|\mathcal{C}|t} \right) \\ &\leq 2d \exp\left(\frac{-3\gamma^2|\mathcal{C}|^2t^2}{6|\mathcal{C}|bt + 2\sqrt{2}\gamma|\mathcal{C}|t} \right) \\ &= 2d \exp\left(\frac{-3\gamma^2|\mathcal{C}|t}{6b + 2\sqrt{2}\gamma} \right) \\ &\leq 2d \exp\left(\frac{-3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|t}{6b + 2\sqrt{2}\gamma} \right) \end{aligned} \tag{C.59}$$

Bounding the probability of the third event Let $m \in \mathcal{V}$ be a task index. We apply Corollary 3 for the sequence of time steps in $\mathcal{T}_m(t)$. We have

$$\boldsymbol{\Delta}_m(t) = \sum_{\tau\in\mathcal{T}_m(t)} \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right]$$

is a martingale sequence, hence

$$\begin{aligned}\mathbb{P}\left[\left\|\boldsymbol{\Delta}_m(t)\right\| > \gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t \,\middle|\, \max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt\right] &\leq 2d \exp\left(\frac{-3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2 t^2}{6\left|\mathcal{T}_m(t)\right| + 2\sqrt{2}\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t}\right) \\ &\leq 2d \exp\left(\frac{-3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2 t^2}{6bt + 2\sqrt{2}\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t}\right) \\ &= 2d \exp\left(\frac{-3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2 t}{6b + 2\sqrt{2}\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}\right).\end{aligned} \tag{C.60}$$

Bounding the probability of the second event Let $\mathcal{C} \in \mathcal{P}$ be a cluster, and let us denote $\mathbf{e}_m$ the m^{th} canonical vector of $\mathbb{R}^{|\mathcal{C}|}$. We have

$$\begin{aligned}\left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}(t)\right\| &= \frac{1}{|\mathcal{C}|}\left\|\sum_{m\in\mathcal{C}}\left(\sum_{\tau\in\mathcal{T}_m(t)} \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)^2\right\| \\ &= \frac{1}{|\mathcal{C}|}\left\|\sum_{m\in\mathcal{C}} \mathbf{e}_m^\top \otimes \left(\sum_{\tau\in\mathcal{T}_m(t)} \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)\right\|^2 \\ &= \frac{1}{|\mathcal{C}|}\left\|\sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \mathbf{e}_{m(\tau)}^\top \otimes \left(\mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)\right\|^2 \\ &= \frac{1}{|\mathcal{C}|}\left\|\sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \mathbf{e}_{m(\tau)}^\top \otimes \mathbf{x}\mathbf{x}(\tau) - \mathbb{E}\left[\mathbf{e}_{m(\tau)} \otimes \mathbf{x}\mathbf{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right\|^2,\end{aligned}$$

where the last equality holds since $m(\tau)$ is measurable w.r.t. $\mathcal{F}_{\tau-1}$. We will apply the Corollary 3 to the set of time steps $\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)$ and the adapted sequence $\mathbf{e}_{m(\tau)}^\top \otimes \mathbf{x}\mathbf{x}(\tau)$ of matrices in $\mathbb{R}^{d\times d|\mathcal{C}|}$. Hence we have

$$\begin{aligned}&\mathbb{P}\left[\sqrt{\left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}(t)\right\|} > \gamma t \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} \,\middle|\, \max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt\right] \\ &\leq d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2|\mathcal{C}| \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t^2}{6\sum_{m\in\mathcal{C}}\left|\mathcal{T}_m(t)\right| + 2\sqrt{2}\gamma\sqrt{|\mathcal{C}| \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t}}\right)\end{aligned}$$

$$
\begin{aligned}
&\leq d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6|\mathcal{C}|\,b+2\sqrt{2}\gamma\sqrt{|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}}\right)\\
&= d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6b+2\sqrt{2}\gamma\sqrt{\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{|\mathcal{C}|}}}\right)\\
&\leq d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6b+2\sqrt{2}\gamma\sqrt{\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|}}}\right) \qquad \text{(C.61)}
\end{aligned}
$$

Union bound We conclude the result of the statement via a union bound using Equation (C.58). □

Proposition 9 (Concentration of the empirical multi-task Gram matrix around the adapted one, simplified)**.** *Let $t \geq 1$, $b > 0$. Assume that $\max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt$. Then we have:*

$$
\mathbb{P}\left[\left\|\frac{\mathbf{A}_{\mathcal{V}}}{t}-\mathbf{V}_{\mathcal{V}}\right\|_{\text{op,RE}} > \gamma\right] \leq 6d|\mathcal{V}|\exp\left(\frac{-3\gamma^2(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C})\wedge\tilde{c}_{\mathcal{G}}(\mathcal{C})^2)t}{6b+2\sqrt{2}\gamma}\right),
$$

where $\tilde{c}_{\mathcal{G}}(\mathcal{C}) \coloneqq c_{\mathcal{G}}(\mathcal{C})\wedge|\mathcal{C}| \quad \forall\mathcal{C}\in\mathcal{P}$.

Proof. The proof will rely on simple calculus inequalities. Hence, let $u = \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})$, $v = \min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|$, $f = 3\gamma^2$, $g = 6b$, $h = 2\sqrt{2}\gamma$, which are all positive. Then, we have

$$
\begin{aligned}
A_1 &= \frac{fu}{f+g} \geq \frac{(u\wedge v)f}{f+g} \geq (u\wedge v)\frac{(1\wedge u\wedge v)f}{f+g(1\wedge u\wedge v)}\\
A_2 &= \frac{fv}{f+g\frac{v}{u}} \geq \frac{(v\wedge u)f}{f+g\frac{v\wedge u}{u}} \geq \frac{(v\wedge u)f}{f+g} \geq (u\wedge v)\frac{(1\wedge u\wedge v)f}{f+(1\wedge u\wedge v)g}\\
A_3 &= \frac{fv^2}{f+gv} \geq \frac{(v\wedge u)^2}{f+(v\wedge u)g} \geq (u\wedge v)\frac{(1\wedge u\wedge v)f}{f+(1\wedge u\wedge v)g}
\end{aligned}
$$

where we used the fact that functions of the form $x \mapsto \frac{x}{\beta_1 x+\beta_2}$ for positive β_1,β_2 are increasing on $\mathbb{R}_+$.

As a final step, we use the inequality $\dfrac{(1\wedge x)f}{f+(1\wedge x)g} \geq \dfrac{x\wedge 1}{f+g}$ taken for $x = u\wedge v$, we apply

the $\exp(-\cdot t)$ function and we use the result of Proposition 8, we deduce the result. □

C.2.3.3 From the true to the adapted Gram matrix

For all of the proofs in this subsection, we follow an approach similar to that of Oh *et al.* [119]. In particular, we use their Lemma 10.

Theorem 10 (Lemma 10 of Oh *et al.* [119])**.** *Under Assumption 2 on the context generating distribution, let $t \geq 1$. We have for any $\boldsymbol{\theta} \in \mathbb{R}^d$:*

$$\sum_{\mathbf{x} \in \mathcal{A}(t)} \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top \mathbb{1}\left\{\mathbf{x} \in \underset{\tilde{\mathbf{x}} \in \mathcal{A}(t)}{\arg\max} \langle \boldsymbol{\theta}, \tilde{\mathbf{x}} \rangle\right\}\right] \succcurlyeq \frac{1}{2\nu\omega}\overline{\boldsymbol{\Sigma}} \tag{C.62}$$

Proposition 10 (RE condition from the true to the adapted Gram matrix)**.** *Under Assumption 2, for any $t \geq 1$, the adapted Gram matrix $\mathbf{V}_{\mathcal{V}}(t)$ verifies the compatibility condition with constants κ and $\dfrac{\phi}{\sqrt{2\nu\omega}}$.*

Proof. For $t \geq 1$, we have

$$\mathbb{E}\left[\mathbf{x}(t)\mathbf{x}(t)^\top | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{\mathbf{x} \in \mathcal{A}(t)} \mathbf{x}(t)\mathbf{x}(t)^\top | \mathcal{F}_{t-1}\right] \tag{C.63}$$

Let $m \in \mathcal{V}$. We have

$$\begin{aligned}
\mathbf{V}_m(t) &= \frac{1}{t}\sum_{\tau \in \mathcal{T}_m(t)} \mathbb{E}\left[\mathbf{x}(\tau)\mathbf{x}(\tau)^\top | \mathcal{F}_{\tau-1}\right] \\
&= \frac{1}{t}\sum_{\tau \in \mathcal{T}_m(t)} \mathbb{E}\left[\mathbb{E}\left[\mathbf{x}(\tau)\mathbf{x}(\tau)^\top | \boldsymbol{\theta}_m(\tau-1), \mathcal{F}_{\tau-1}\right] | \mathcal{F}_{\tau-1}\right] \quad \text{(total expectation law)} \\
&= \frac{1}{t}\sum_{\tau \in \mathcal{T}_m(t)} \mathbb{E}\left[\mathbf{x}(\tau)\mathbf{x}(\tau)^\top | \boldsymbol{\theta}_m(\tau-1)\right] \quad (\mathbf{x}(\tau) \text{ is fully determined by } \boldsymbol{\theta}_m(\tau-1)) \\
&= \frac{1}{t}\sum_{\tau \in \mathcal{T}_m(t)} \mathbb{E}\left[\sum_{\mathbf{x} \in \mathcal{A}(\tau)} \mathbf{x}\mathbf{x}^\top \mathbb{1}\left\{\mathbf{x} \in \underset{\tilde{\mathbf{x}} \in \mathcal{A}(t)}{\arg\max} \langle \boldsymbol{\theta}, \tilde{\mathbf{x}} \rangle\right\} | \boldsymbol{\theta}_m(\tau-1)\right] \\
&\succcurlyeq \frac{1}{2\nu\omega}\overline{\boldsymbol{\Sigma}} \quad \text{(by Theorem 10).}
\end{aligned} \tag{C.64}$$

Now, let $\mathbf{Z} \in \mathcal{S}$, where $\mathcal{S}$ is defined with constant κ of Assumption 4. Then

$$\sum_{m \in \mathcal{V}} \|\mathbf{z}\|_{\mathbf{V}_m(t)} \geq \frac{1}{2\nu\omega}\sum_{m \in \mathcal{V}} \|\mathbf{z}_m\|_{\overline{\boldsymbol{\Sigma}}} \quad \text{by Equation (C.64)}$$

$$\geq \frac{\phi^2}{2\nu\omega}\|\mathbf{Z}\|_{RE}^2 \quad \text{(by Assumption 4)},$$

which finishes the proof. □

Theorem 11 (RE condition holding for the empirical multi-task Gram matrix, generalization of Theorem 4)**.** *Under Assumptions 2 and 4, let $t \geq 1$, and let κ, ϕ be the constants from Assumption 4. Assume that $\max_{m\in\mathcal{V}} \left|\mathcal{T}_m(t)\right| \leq bt$. Then, for any*

$$\gamma \in \left(0, \left(1 + \frac{a_2(\mathcal{G},\mathbf{\Theta}) + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1(\mathcal{G},\mathbf{\Theta})}\right)^{-2}\right),$$

the empirical multi-task Gram matrix verifies the RE condition with constants κ and $\hat{\phi}$, with

$$\hat{\phi} = \tilde{\phi}\sqrt{1 - \gamma\left(1 + \frac{a_2(\mathcal{G},\mathbf{\Theta}) + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1(\mathcal{G},\mathbf{\Theta})}\right)^2}, \tag{C.65}$$

with a probability at least equal to $1 - 6d|\mathcal{V}|\exp(\frac{-3\gamma^2\tilde{\phi}^4(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_\mathcal{G}(\mathcal{C}) \wedge \tilde{c}_\mathcal{G}(\mathcal{C})^2)t}{6b + 2\sqrt{2}\gamma\tilde{\phi}^2})$,

where $\tilde{\phi} \coloneqq \frac{\phi}{\sqrt{2\nu\omega}}$ *and* $\tilde{c}_\mathcal{G}(\mathcal{C}) \coloneqq c_\mathcal{G}(\mathcal{C}) \wedge |\mathcal{C}| \quad \forall \mathcal{C} \in \mathcal{P}$.

Proof. For the sake of readability, let $\tilde{\phi} = \frac{\phi}{\sqrt{2\nu\omega}}$ the compatibility constant of the adapted Gram matrix, according to Proposition 10. Then:

$$1 - 6d|\mathcal{V}|\exp(\frac{-3\gamma^2\tilde{\phi}^4(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_\mathcal{G}(\mathcal{C}) \wedge \tilde{c}_\mathcal{G}(\mathcal{C})^2)t}{6b + 2\sqrt{2}\gamma\tilde{\phi}^2}) \tag{C.66}$$

$$\leq \mathbb{P}\left[\left\|\frac{\mathbf{A}_\mathcal{V}}{t} - \mathbf{V}_\mathcal{V}\right\|_{\text{op,RE}} \leq \gamma\tilde{\phi}^2\right] \quad \text{(by Proposition 9)} \tag{C.67}$$

$$\leq \mathbb{P}\left[\frac{\mathbf{A}_\mathcal{V}}{t} \text{ satisfies the RE condition with constant } \kappa \text{ and } \hat{\phi}\right] \quad \text{(by Proposition 7)}, \tag{C.68}$$

where $\hat{\phi} = \tilde{\phi}\sqrt{1 - \gamma\left(1 + \frac{a_2(\mathcal{G},\mathbf{\Theta}) + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1(\mathcal{G},\mathbf{\Theta})}\right)^2}$. □

C.2.4 Regret bound

Lemma 11 (Concentration of the fraction of observations per task)**.** *Assume that* $|\mathcal{V}| \geq 2$*. Then for* $\delta \in (0,1)$*, we have with a probability at least* $1-\delta$*:*

$$\max_{m\in\mathcal{V}} \frac{\left|\mathcal{T}_m(t)\right|}{t} \leq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{1}{t|\mathcal{V}|}\log\frac{|\mathcal{V}|}{\delta}} + \frac{4}{3t}\log\frac{|\mathcal{V}|}{\delta}. \tag{C.69}$$

Proof. We have $\left|\mathcal{T}_m(t)\right| \coloneqq \sum_{\tau=1}^{t}[m(\tau)=m]$, where $\forall t, \forall m \in \mathcal{V}, \mathbb{P}\left[m(t)=m\right] = \frac{1}{|\mathcal{V}|}$, meaning that the binary variable $[m(t)=m]$ follows a Bernoulli distribution $\mathcal{B}(\frac{1}{\mathcal{V}})$. Then, the random variable $X_t \coloneqq [m(t)=m] - \frac{1}{|\mathcal{V}|}$ has mean 0, variance $\frac{1}{|\mathcal{V}|}(1-\frac{1}{|\mathcal{V}|})$, and verifies $|X_t| \leq 1 - \frac{1}{|\mathcal{V}|}$ since $|\mathcal{V}| \geq 2$. As a result, via the Bernstein inequality, we have for any $m \in \mathcal{V}$, and for any $w \geq 0$,

$$\mathbb{P}\left[\frac{\left|\mathcal{T}_m(t)\right|}{t} \geq \frac{1}{|\mathcal{V}|} + w\right] \leq \exp\left(-\frac{tw^2}{2(1-\frac{1}{|\mathcal{V}|})(\frac{1}{|\mathcal{V}|}+\frac{w}{3})}\right) \leq \exp\left(-\frac{tw^2}{2(\frac{1}{|\mathcal{V}|}+\frac{w}{3})}\right)$$

For the right-hand side to hold with a probability at most $\delta \in (0,1)$, it is sufficient to have

$$\begin{aligned}
& t\frac{w^2}{2(\frac{1}{|\mathcal{V}|}+\frac{w}{3})} \geq \log\frac{1}{\delta} \\
\Longleftarrow\ & \frac{w^2}{2} \geq \frac{2\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}}{t} \text{ and } \frac{w^2}{2} \geq \frac{2w\log\frac{1}{\delta}}{3t} \\
\Longleftarrow\ & w = 2\sqrt{\frac{\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}}{t}} + \frac{4\log\frac{1}{\delta}}{3t}
\end{aligned}$$

Hence, and via a union bound, we get

$$\begin{aligned}
& \mathbb{P}\left[\frac{\left|\mathcal{T}_m(t)\right|}{t} \geq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}} + \frac{4}{3t}\log\frac{1}{\delta}\right] \leq \delta \\
\Longrightarrow\ & \mathbb{P}\left[\max_{m\in\mathcal{V}}\frac{\left|\mathcal{T}_m(t)\right|}{t} \geq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}}{t}} + \frac{4\log\frac{1}{\delta}}{3t}\right] \leq |\mathcal{V}|\,\delta
\end{aligned}$$

The result is obtained by adjusting the value of δ. □

Theorem 12 (Regret bound, generalization of Theorem 5)**.** *Let the mean horizon per node be* $\overline{T} = \frac{T}{|\mathcal{V}|}$*. Under Assumptions 1, 2 and 4 and* $\kappa > 0$*, the expected regret of the*

Network Lasso Bandit algorithm is upper bounded as follows:

$$\mathcal{R}(\overline{T}) \leq \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\phi^2}\left(\sqrt{|\mathcal{V}|}+\sqrt{\log(\overline{T}|\mathcal{V}|)}+\sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)+\right.$$
$$\left.+\frac{1}{A}\log(d|\mathcal{V}|)+\sqrt{|\mathcal{V}|}\right).$$

$$\textit{with } A=\frac{3\gamma^2\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C})\wedge\tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}}+2\sqrt{2}\gamma}.$$

Proof. For any time step t, we will define a list of good events under which the Oracle inequality and the RE condition for the empirical multi-task Gram matrix both hold with high probability. Then, we will use those bounds to sum up over time steps until horizon T.

Good events We formalize these requirements as three families of time-depending "good" events.

- $G_{\text{pro}}(t)$ is the event that the mean of the empirical process bounded by $\alpha(t)$ up to a constant c, which is equivalent to saying that it converges:

$$G_{\text{pro}}(t) := \left\{\frac{1}{t}\|\mathbf{K}\|_F \leq \frac{\alpha(t)}{\alpha_0}\right\} \tag{C.70}$$

- $G_{sel}(t)$ is the event that the number of selections of all tasks is bounded by its expected value up to a small constant $\rho(t)$

$$G_{sel}(t) := \left\{\max_{m\in\mathcal{V}}\frac{\left|\mathcal{T}_m(t)\right|}{t} \leq \frac{1}{|\mathcal{V}|}+\frac{\rho(t)}{t}\right\} \tag{C.71}$$

- $G_{\text{RE}}(t)$ is the event that the empirical multi-task Gram matrix $\frac{1}{t}\mathbf{A}_{\mathcal{V}}(t)$ satisfies the RE condition.

$$G_{\text{RE}}(t) := \left\{\frac{1}{t}\mathbf{A}_{\mathcal{V}}(t) \text{ verifies the RE condition with constants } \kappa,\hat{\phi}\right\} \tag{C.72}$$

Event $G_{\text{pro}}(t)$ is the most straightforward to cover since our bound on the empirical process given in Lemma 9 holds with a probability of at least $1-\delta(t)$, thus:

$$\mathbb{P}\left[G_{\text{pro}}(t)^c|G_{sel}(t)\right] \leq \delta(t), \tag{C.73}$$

where we included the time dependency on $\delta(t)$ in contrast to the previous section.

This way we emphasize to adjust $\delta(t)$ after each round, to guarantee a sub linear regret bound. The probability of event $G_{sel}(t)$ can be determined using Bernstein's inequality:

From Lemma 11 we can select $\rho(t) = 2\sqrt{\frac{t}{|\mathcal{V}|}\log\frac{|\mathcal{V}|}{\delta_{sel}(t)}} + \frac{4}{3}\log\frac{|\mathcal{V}|}{\delta_{sel}(t)}$ as well as

$$\mathbb{P}\left[G_{sel}(t)^c\right] \leq \delta_{sel}(t).$$

C.2.4.1 Instantaneous regret decomposition

Now, given the event probabilities, we condition the instantaneous regret $r(t)$ on the good events at a time $t > t_0$. We have for its expectation:

$$\begin{aligned}\mathbb{E}\left[r(t)\right] &\leq \mathbb{E}\left[r(t)|G_{sel}(t)\right] + 2\mathbb{P}\left[G_{sel}(t)^c\right] \\ &\leq \mathbb{E}\left[r(t)|G_{\text{pro}}(t) \cap G_{\text{RE}}(t) \cap G_{sel}(t)\right] \\ &+ 2\left(\mathbb{P}\left[G_{\text{pro}}(t)^c|G_{sel}(t)\right] + \mathbb{P}\left[G_{\text{RE}}(t)^c|G_{sel}(t)\right] + \mathbb{P}\left[G_{sel}(t)^c\right]\right), \end{aligned} \tag{C.74}$$

where we used the worst case bound $r(t) \leq 2$ if any one of the good events does not hold.

Bounding the regret Inserting our results of the event probabilities, the oracle inequality and the decomposition of the expected instantaneous regret in Equation (C.74) and bounding the sum over rounds, yields the final result. Thus, we start by bounding the sum over the first term i.e. the expected regret in case all good events hold:

$$\sum_{t=1}^{T}\mathbb{E}\left[r(t)|G_{\text{pro}}(t) \cap G_{\text{RE}}(t) \cap G_{sel}(t)\right] \leq \sum_{t=1}^{T}\left\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(t)\right\|_F$$

Taking the result of our oracle inequality in Theorem 3, we point out that only $\alpha(t)$ is time dependent such that the rest of the terms can be pulled outside the sum:

$$\begin{aligned}&\sum_{t=1}^{T}\left\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(t)\right\|_F \leq \sum_{t=1}^{T} 2\frac{\sigma}{\hat{\phi}^2\sqrt{t}} f(\mathcal{G},\boldsymbol{\Theta})\sqrt{1 + 2b\sqrt{|\mathcal{V}|\log\frac{1}{\delta(t)}} + 2b\log\frac{1}{\delta(t)}} \\ &= \frac{2\sigma}{\hat{\phi}^2} f(\mathcal{G},\boldsymbol{\Theta})\sum_{t=1}^{T}\sqrt{\frac{1}{t} + \frac{2b}{t}\sqrt{2|\mathcal{V}|\log(t)} + \frac{4b}{t}\log(t)} \\ &\leq \frac{2\sigma}{\hat{\phi}^2} f(\mathcal{G},\boldsymbol{\Theta})\int_0^T \frac{1}{\sqrt{t}} + \sqrt{\frac{2b}{t}\left(\sqrt{2|\mathcal{V}|\log(T)} + 2\log(T)\right)}\, dt\end{aligned}$$

$$\leq \frac{2\sigma}{\hat{\phi}^2} f(\mathcal{G},\boldsymbol{\Theta}) \left(2\sqrt{T} + \left(\frac{\sqrt{8T}}{|\mathcal{V}|} + 4\sqrt[4]{\frac{32\log(|\mathcal{V}|T)T}{|\mathcal{V}|}} + \sqrt{\frac{16}{3}\log(|\mathcal{V}|T)\log(T)} \right) \right.$$
$$\left. \left(\sqrt[4]{2|\mathcal{V}|\log(T)} + \sqrt{2\log(T)} \right) \right)$$
$$= \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\hat{\phi}^2} \left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|} \right) \right),$$

where

$$f(\mathcal{G},\boldsymbol{\Theta}) := \left(a_2(\mathcal{G},\boldsymbol{\Theta}) + \sqrt{2}\mathbb{1}_{\leq 1}(\kappa) w(\partial\mathcal{P}) \right) \left(\frac{a_2(\mathcal{G},\boldsymbol{\Theta}) + \sqrt{2}\mathbb{1}_{\leq 1}(\kappa) w(\partial\mathcal{P})}{a_1(\mathcal{G},\boldsymbol{\Theta}) \min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} + 1 \right).$$

We upper bounded the sum with an integral i.e. $\sum_{t=1}^{T} f(t) \leq \int_0^T f(t)dt$ for monotonically decreasing functions $f(t)$ in the last inequality. Also b is the bound on the concentration of the fraction of observation per task provided by Lemma 11. For $t_0 = \sqrt{|\mathcal{V}|}$ we find by inserting the result to Lemma 11 for all $t > t_0$:

$$\frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{1}{t|\mathcal{V}|}\log\frac{|\mathcal{V}|}{\delta}} + \frac{4}{3t}\log\frac{|\mathcal{V}|}{\delta} \leq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{2\log(|\mathcal{V}|\sqrt{|\mathcal{V}|})}{\sqrt{|\mathcal{V}|}|\mathcal{V}|}} + \frac{8\log(|\mathcal{V}|\sqrt{|\mathcal{V}|})}{3\sqrt{|\mathcal{V}|}}$$
$$= \frac{1}{|\mathcal{V}|} + \frac{2}{\sqrt{|\mathcal{V}|}} \left[\sqrt{\frac{3}{\sqrt{|\mathcal{V}|}}\log(|\mathcal{V}|)} + 2\log(|\mathcal{V}|) \right]$$
$$= \mathcal{O}\left(\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} \right) = b.$$

Finally we bound the sum over the instantaneous regret term for the bad events:

$$\sum_{t=1}^{T} 2\left(\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{sel}(t) \right] + \mathbb{P}\left[G_{\text{RE}}(t)^c | G_{sel}(t) \right] + \mathbb{P}\left[G_{sel}(t)^c \right] \right)$$

By construction, we have $\max(\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{sel}(t) \right], \mathbb{P}\left[G_{sel}(t)^c \right]) \leq \delta(t) = \frac{1}{t^2}$. Hence,

$$\sum_{t=1}^{T} \mathbb{P}\left[G_{\text{pro}}(t)^c | G_{sel}(t) \right] + \mathbb{P}\left[G_{sel}(t)^c \right] \leq 2\sum_{t=1}^{T} \frac{1}{t^2} \leq 2\left(1 + \int_1^T \frac{dt}{t^2} \right) \leq 4 \qquad \text{(C.75)}$$

As for the RE condition event, letting $A \coloneqq \dfrac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}{6b+2\sqrt{2}\gamma}$, we have for any $t_0 \geq 1$

$$\begin{aligned}
\sum_{t=t_0}^{T} \mathbb{P}\left[G_{\mathrm{RE}}(t)^c | G_{sel}(t)\right] &\leq 6d|\mathcal{V}| \sum_{t=t_0}^{T} \exp(-At) \quad \text{(by Theorem 4)} \\
&\leq 6d|\mathcal{V}| \frac{e^{-At_0}}{1-e^{-A}} \leq 6d|\mathcal{V}| e^{-At_0}\left(1+\frac{1}{A}\right) \\
&\leq 6d|\mathcal{V}| e^{-At_0}\left(1+\frac{1}{A}\right)
\end{aligned}$$

where in the last line, we used the inequality $\exp(A) \geq A+1$. Hence, for any $u>0$, choosing

$$t_0 = \left\lceil \sqrt{|\mathcal{V}|} \right\rceil \vee \left\lceil \frac{1}{A}\log\left(\frac{6d|\mathcal{V}|\,(1+\frac{1}{A})}{u}\right) \right\rceil$$

implies that $\sum_{t=t_0}^{T} \mathbb{P}\left[G_{\mathrm{RE}}(t)^c | G_{sel}(t)\right] \leq u$. Before we continue with the regret bound, we need to find an appropriate bound on $\frac{f(\mathcal{G},\boldsymbol{\Theta})}{\hat{\phi}^2}$. Given our result in Theorem 3 and assuming that $\kappa > 1$, we get:

$$\begin{aligned}
\frac{f(\mathcal{G},\boldsymbol{\Theta})}{\hat{\phi}^2} &= \frac{\alpha_0 a_2(\mathcal{G},\boldsymbol{\Theta})}{\hat{\phi}^2}\left(\frac{a_2(\mathcal{G},\boldsymbol{\Theta})}{a_1(\mathcal{G},\boldsymbol{\Theta})\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}+1\right) \\
&= \frac{\left(\sqrt{2}\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\alpha_0+1\right)\left(\frac{\sqrt{2}\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\alpha_0+1}{\alpha_0(\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}-2\kappa w(\partial\mathcal{P}))-1}+1\right)}{1-\gamma\left(1+\frac{\sqrt{2}\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\alpha_0+1}{\alpha\left(1-\frac{2\kappa w(\partial\mathcal{P})}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\right)-\frac{1}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}}\right)^2} \\
&= \mathcal{O}\left(\frac{\max_{\mathcal{C}\in\mathcal{P}}\iota_{\mathcal{G}}(\mathcal{C})+\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}+1}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}+\max_{\mathcal{C}\in\mathcal{P}}\iota_{\mathcal{G}}(\mathcal{C})+1\right)
\end{aligned}$$

$$= \mathcal{O}\left(\frac{1}{\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}}\right).$$

The first big $\mathcal{O}$ notation is obtained due to the fact that for for large $\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}$ and small $\max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}$ the denominator term i.e. $\hat{\phi}^2$ behaves like $1-\gamma$, which leaves the numerator dominating the rest of the term. Now, we simply have to insert all our results into the sum of instantaneous regrets:

$$\begin{aligned}
\mathcal{R}(\overline{T}) &\leq t_0 + 2u + 8 + \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|}\right)\right) \\
&\leq \left\lceil\sqrt{|\mathcal{V}|}\right\rceil + \left\lceil\frac{1}{A}\log(\frac{6d|\mathcal{V}|\,(1+\frac{1}{A})}{u})\right\rceil + 2u + 8 \\
&+ \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|}\right)\right) \\
&\leq \left\lceil\sqrt{|\mathcal{V}|}\right\rceil + \left\lceil\frac{1}{A}\log(12d|\mathcal{V}|\,(1+A))\right\rceil + \frac{1}{A} + 8 \\
&+ \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|}\right)\right) \\
&\leq \left\lceil\sqrt{|\mathcal{V}|}\right\rceil + \left\lceil\frac{1}{A}\log(12d|\mathcal{V}|\,(1+A))\right\rceil + \frac{1}{A} + 8 \\
&+ \mathcal{O}\left(\frac{f(\mathcal{G},\boldsymbol{\Theta})\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|}\right)\right) \\
&= \mathcal{O}\left(\frac{1}{A}\log(d|\mathcal{V}|) + \sqrt{\frac{\overline{T}}{\min\limits_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{\left|\mathcal{V}\log(\overline{T}|\mathcal{V}|)\right|}\right)\right),
\end{aligned}$$

where we set $u = \frac{1}{2A}$ in the third inequality. □

Proof of Corollary 2. For $\kappa > 1$ and $\gamma = \frac{1}{2}\left(1 + \frac{a_2(\mathcal{G},\boldsymbol{\Theta})}{a_1(\mathcal{G},\boldsymbol{\Theta})}\right)^{-2}$ we have $\hat{\phi} = \frac{\phi}{2\sqrt{\nu\omega}}$. We find an appropriate bound on $f(\mathcal{G},\boldsymbol{\Theta})$ by minimizing the expression w.r.t. α_0 as done in Proposition 5 such that:

$$\min_{\alpha_0>0} f(\boldsymbol{\Theta},\mathcal{G}) = \mathcal{O}\left(1+\sqrt{\frac{\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}-2\kappa w(\partial\mathcal{P})}}\right),$$

which holds if the expression $\sqrt{\frac{\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}}$ is negligible compared to 1, which is guaranteed given the assumptions $\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})} \gg \kappa w(\partial\mathcal{P})$ and that

$$\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})} \leq 1$$

for which the expression simplifies to:

$$\min_{\alpha_0>0} f(\boldsymbol{\Theta},\mathcal{G}) = \mathcal{O}\left(1+\sqrt{\frac{\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}}\right).$$

Inserting everything into our regret bound from Theorem 12 we get:

$$\mathcal{R}(\overline{T}) = \mathcal{O}\Bigg(\frac{\nu\omega\left(1+\sqrt{\frac{\kappa w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}}\right)}{\phi^2}$$
$$\sqrt{\overline{T}}\left(\sqrt{|\mathcal{V}|}+\sqrt{\log(\overline{T}|\mathcal{V}|)}+\sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)+\frac{1}{A}\log(d|\mathcal{V}|)+\sqrt{|\mathcal{V}|}\Bigg)$$

If we assume that $|\mathcal{V}| \gg \log T$ and $|\mathcal{V}| = \mathcal{O}(T)$, then then we have further

$$\sqrt{|\mathcal{V}|}+\sqrt{\log(\overline{T}|\mathcal{V}|)}+\sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)} = \mathcal{O}(\sqrt{T})$$

which yields the result. □

C.3 Additional related work

Homophily and modularity in social networks Given the large number of users on social networks, one may be able to learn their preferences more quickly by leveraging the similarities between them. This idea relies on the notion of *homophily* in social networks [109, 53]. In modelling social networks, users' preferences relationships are encoded in a graph, where neighboring nodes are users with similar preferences. This graph can be known *a priori* or it can be inferred from previously collected feedback [50]. Exploiting this information and integrating them into bandit algorithms can lead to

a significant increase in performance Yang *et al.* [169]. Indeed, the knowledge of user relations allows the algorithm to tackle the data sparsity issue that is inherent to bandit settings.

Another fundamental point that can be used for integration of information from social networks is that, social networks show large *modularity* measures [114] [26]. This implies that we have high density of edges within clusters and low density of edges between clusters. As a result, users can be clustered based on the graph topology and a preference vector can be learned for each cluster, substantially reducing the dimensionality of the problem. In other words, discovering the clustering structure of users can reduce the computational burden of large social networks. Consequently, there have been attempts in exploiting the clustered structures of social networks in bandit algorithms [59, 115, 168, 97, 117, 46].

Bandit meta-learning In contrast to the multi-task setting, meta learning deals with sequentially arriving tasks that have to be learnt and generalizing the gained information to improve performance for future tasks. Here, as in the multi-task setting, it is assumed that the tasks share some common structure that is ought to be learnt and exploited. In the work of [24] it is assumed that the tasks were sampled from a common distribution such that they are concentrated around an affine subspace, which is learnt through PCA algorithm. The resulting projection matrices could then be exploited to improve learning for new tasks in an adapted UCB and Thompson sampling approach.
Other lines of work are [35, 87, 19], which learns the mean of the distribution under the assumption that the covariance of the prior is known or [123] which generalizes this assumption and attempts to learn the covariance as well.

C.4 More on the experiments

C.4.1 About experiments of the main paper

The experiments have been conducted with an intel i7 CPU with 12 2.6 GHz cores and 32 GB of RAM. The two experiments with the highest number of tasks (200) and dimension (80) take about 8 hours, parallelized over the 12 cores.

To generate clusters, we generate $|\mathcal{P}|$ variables $v_{ii\in\mathcal{P}}$ from the uniform distribution, then we use them to construct a categorical distribution with probabilities proportional to e^{v_i}. These probabilities defines the cluster proportions.

C.4.2 Solving the Network Lasso problem

We implement the Primal-Dual algorithm proposed in Jung [74] to solve the Network Lasso problem but we do not vectorize the matrices (in the sense of stacking their columns into a vector), which speeds up computation.

C.4.3 Algebraic connectivity vs topological centrality index

Given two fully connected graphs weightless $\mathcal{G}_1$ and $\mathcal{G}_2$ with size 100 each, we progressively link them by edges, we construct the Laplacian $\mathbf{L}$ of the resulting graph $\mathcal{G}$. We measure the minimum topological centrality index $\min_{1\leq i\in 200}(L_{\mathcal{C}}^{\dagger})_{ii}^{-1}$, and the algebraic connectivity, i.e. the minimum non-null eigenvalue of L.

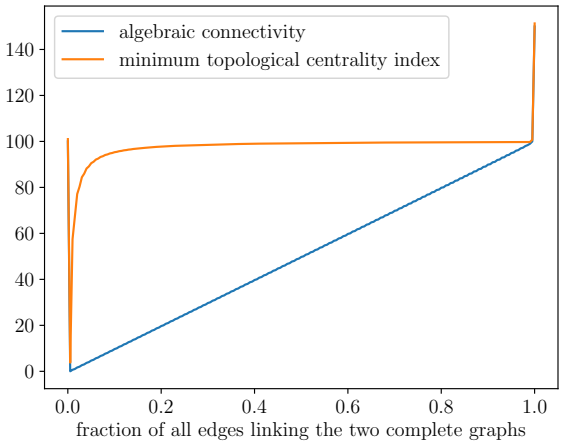


Figure C.1: Minimum Topological centrality index vs Algebraic Connectivity, for a graph formed by connecting two fully connected initial graphs $\mathcal{G}_1, \mathcal{G}_2$ with size 100 each.

Clearly, the minimum topological centrality index grows faster than the algebraic connectivity in this case, and seems to saturate at some level that is reached in a linear progress by the algebraic connectivity.

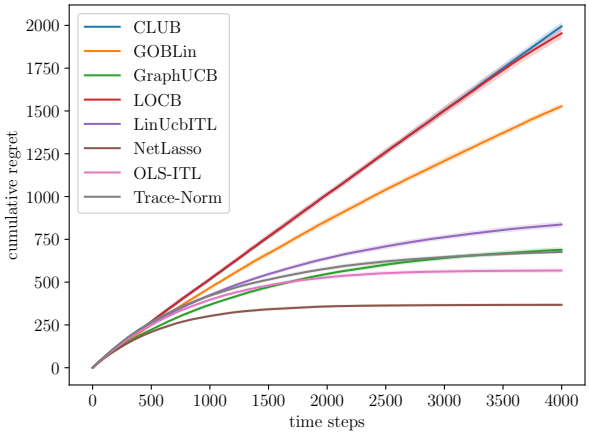
C.4.4 Limitations

The first limitation of the paper is the restriction to the setting of i.i.d generated action sets. This restriction is common to all papers relying on Lasso-type optimization objectives [18, 119, 34, 6, 36]. Also, we do not provide a lower bound for the regret, a challenge that we let for future work. Besides, our optimization problem is not strongly convex, which can be mitigated by adding a squared L^2 norm regularization. However, such an addition would probably drastically change the theoretical analysis.

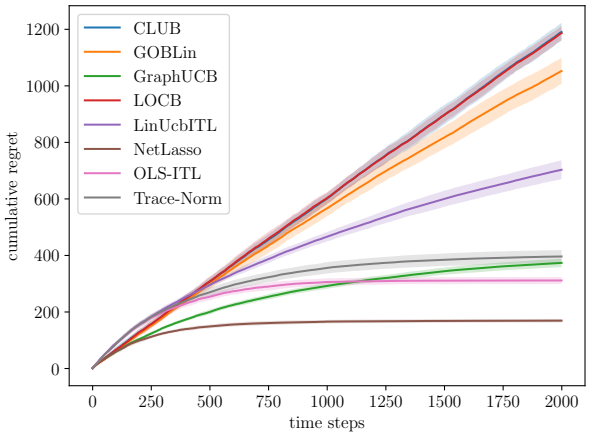
C.4.5 Broader impacts

As our method can be applied to transfer knowledge between users of a recommender system, it has the potential to improve their overall experience by learning their preferences quickly. However, one must be careful with the strength of the integrated prior knowledge as it can lead to an adverse effect of slowing down the learning process.

C.4.6 Experiments where the number of clusters is higher than the dimension



(a) $|\mathcal{V}| = 200, |\mathcal{P}| = 25, d = 10, p = 0.5, q = 0.05$



(b) $|\mathcal{V}| = 200, |\mathcal{P}| = 20, d = 2, p = 0.5, q = 0.1$

Figure C.2: Synthetic data experiments showing the cumulative regret of Network Lasso Policy as a function of time-steps compared to other baselines, for the case where the number of clusters exceeds the dimension

Notation	Meaning
$\mathcal{V}$	Set of graph vertices
$\mathcal{E}$	Set of graph edges
$\mathbf{B}_I \in \mathbb{R}^{\lvert\mathcal{E}\rvert\times\lvert\mathcal{V}\rvert}, I \subseteq \mathcal{E}$	Graph incidence matrix obtained by setting rows of edges outside I to zeros
$\mathbf{B}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{E}\rvert\times\lvert\mathcal{V}\rvert}$	cf. Definition 1
$\mathbf{L} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	$\mathbf{B}^\top\mathbf{B}$
$\boldsymbol{\theta}_m \in \mathbb{R}^d$	True preference vector of user/bandit m
$\boldsymbol{\Theta} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times d}$	Matrix of true vertically concatenated row preference vectors
$\partial\mathcal{P} \subseteq \mathcal{E}$	Boundary of $\mathcal{P}$: set of edges connecting nodes from different clusters
$c_{\mathcal{G}}(\mathcal{C})$	Minimum topological centrality index of a node of $\mathcal{C}$, restricted to the graph having nodes $\mathcal{C}$
$w(\partial\mathcal{P})$	Total weight of $\partial\mathcal{P}$, *i.e.*, sum of weights of edges in $\mathcal{P}$
$\lVert\cdot\rVert$	Euclidean norm for vectors, largest singular value for matrices
$\lVert\cdot\rVert_{\mathbf{A}}$	Semi-norm defined by PSD matrix $\mathbf{A}$: $\lVert\mathbf{x}\rVert_{\mathbf{A}}^2 \coloneqq \mathbf{x}^\top\mathbf{A}\mathbf{x}$
$\lVert\cdot\rVert_F$	Matrix Frobenius norm
$\lVert\cdot\rVert_{p,q}$	q-norm of the vector with coordinates equal to the p-norm of rows
$\lVert\cdot\rVert_I, I \subseteq \mathcal{E}$	Total variation norm of a signal over edges of I
$\mathbf{A}^\dagger$	Moore-Penrose pseudo-inverse of matrix $\mathbf{A}$
vec	Vectorization operator, concatenating columns vertically
$\otimes$	Kronecker product
$\mathbf{1}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	Vector with elements equal to 1 at coordinates corresponding to vertices in $\mathcal{C}$, and 0 elsewhere
$\mathbf{J}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	Equal to $\frac{\mathbf{1}_{\mathcal{C}}\mathbf{1}_{\mathcal{C}}^\top}{\lvert\mathcal{C}\rvert}$
$\mathbf{Q}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	Equal to $\mathbf{B}_{\mathcal{C}}^\dagger\mathbf{B}_{\mathcal{C}}$
$\mathbf{Q}_I \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}, I \subseteq \mathcal{E}$	Equal to $\mathbf{B}_I^\dagger\mathbf{B}_I$
$\mathbf{e}_k$	Elementary vector of dimension depending on the context
σ	Subgaussianity constant/variance proxy

Table C.1: Notation table (parameters independent of time t).

Notation	Meaning
$\mathcal{T}_m(t)$	Set of time steps user m has been encountered before time t
$\hat{\boldsymbol{\theta}}_m \in \mathbb{R}^d$	Estimated preference vector of user/bandit m
$\boldsymbol{\varepsilon}_m \in \mathbb{R}^d$	Estimation error for user/bandit m: $\hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m$
$\mathbf{E} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d}$	Vertical concatenation of row vectors $\boldsymbol{\varepsilon}_m$
$\boldsymbol{\eta}_m \in \mathbb{R}^{\lvert\mathcal{T}_m(t)\rvert}$	Vector of subgaussian noise for user m
$\mathbf{x}(t) \in \mathbb{R}^d$	Context vector received at time t
$m(t) \in \mathbb{N}$	User at time t
$\mathbf{X}_m \in \mathbb{R}^{\lvert\mathcal{T}_m(t)\rvert \times d}$	Data matrix of user m
$\mathbf{X} \in \mathbb{R}^{t \times d}$	Data matrix of context vectors of all users
$\mathbf{A}_m \in \mathbb{R}^{d \times d}$	$\mathbf{X}_m^\top \mathbf{X}_m$ (potentially associated to time t)
$\mathbf{A}_\mathcal{V} \in \mathbb{R}^{d\lvert\mathcal{V}\rvert \times d\lvert\mathcal{V}\rvert}$	$\operatorname{diag}(\mathbf{A}_1, \cdots, \mathbf{A}_m)$
$\mathbf{K} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d}$	Matrix of vertically concatenated row vectors $\boldsymbol{\eta}_m^\top \mathbf{X}_m$

Table C.2: Notation table (parameters dependent on time t).

Appendix D

Additional Material for Chapter 6

D.1 Proof of Lemma 2

Here we first present an auxiliary lemma in the following.

Lemma 12. *Let A, B be a Hermitian matrix in $\mathbb{R}^{d\times d}$ and suppose $A, B \succ 0$, then $A \succeq B$ if and only if $B^{-1} \succeq A^{-1}$ [69].*

The proof of Lemma 2 is as follows.

Proof. Let $\mathcal{H}_t$ denote the history by the end of time step t hence $\{\mathcal{H}_t\}_{t=0}^{\infty}$ is a filtration. Note that $\hat{p}_t$ is $\mathcal{H}_{t-1}$-adaptive and $\mathcal{S}_t$ is also $\mathcal{H}_{t-1}$-adaptive[157]. For $t \in \{1,2,\ldots,T\}$ and $e = (u,v) \in \bar{\mathcal{E}}_t$, define

$$\eta_t(e) = y_t(e) - \boldsymbol{\theta}^T \boldsymbol{x}_e - c_{u,t}.$$

Note that $\eta_t(e)$ is $\mathcal{H}_{t-1}$-adaptive. Besides, it satisfies $\left\|\eta_t(e)\right\| \leq 1$ and $\mathbb{E}[\eta_t(e)|\mathcal{H}_{t-1}] = 0$ [157], hence they are conditionally sub-Gaussian with constant $R = 1$ [164]. We further define

$$\begin{aligned}
\boldsymbol{V}_t &= \sigma^2 \boldsymbol{M}_t = \sigma^2 \boldsymbol{I} + \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,\tau} \boldsymbol{x}_e \boldsymbol{x}_e^T \\
\boldsymbol{S}_t &= \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,\tau} \boldsymbol{x}_e \eta_\tau(e) \\
&= \boldsymbol{b}_t - \sigma^2(\boldsymbol{M}_t - I)\boldsymbol{\theta} - \sum_{\tau=1}^{t} \sum_{e\in\tilde{\mathcal{E}}_\tau} \omega_{e,t} \boldsymbol{x}_e c_{u,t}.
\end{aligned}$$

Thus we have

$$\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta} = \boldsymbol{M}_t^{-1}(\sigma^{-2}\boldsymbol{S}_t + \sigma^{-2}\sum_{\tau=1}^{t}\sum_{e\in\tilde{\mathcal{E}}_\tau}\omega_{e,\tau}\boldsymbol{x}_e c_{u,\tau} - \boldsymbol{\theta}).$$

It implies

$$\begin{aligned}|\boldsymbol{x}_e^T(\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta})| &\le |\boldsymbol{x}_e\boldsymbol{M}_{t-1}^{-1}\boldsymbol{\theta}| + \sigma^{-2}|\boldsymbol{x}_e\boldsymbol{M}_{t-1}^{-1}\boldsymbol{S}_{t-1}| + \sigma^{-2}|\boldsymbol{x}_e\boldsymbol{M}_{t-1}^{-1}\sum_{\tau=1}^{t}\sum_{e\in\tilde{\mathcal{E}}_\tau}\omega_{e,\tau}\boldsymbol{x}_e c_{u,\tau}| \\ &\le \|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}(\|\boldsymbol{\theta}\|_2 + \underbrace{\sigma^{-2}\|\boldsymbol{S}_{t-1}\|_{\boldsymbol{M}_{t-1}^{-1}}}_{\text{stochastic error}} + \underbrace{\sigma^{-2}\left\|\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}\omega_{e,\tau}\boldsymbol{x}_e c_{u,\tau}\right\|_{\boldsymbol{M}_{t-1}^{-1}}}_{\text{corruption error}}),\end{aligned} \tag{D.1}$$

where Equation (D.1) is based on Cauchy-Schwarz inequality and matrix operator inequality.

The stochastic error can be bounded by the concentration Lemma J.2 in [2]: According to the Appendix I.1 in [67], we introduce the auxiliary vectors $\tilde{\boldsymbol{x}}_{e,t} = \sqrt{\omega_{e,t}}\boldsymbol{x}_e$ and $\tilde{\eta}_t(e) = \sqrt{\omega_{e,t}}\eta_t(e)$. Then, it holds

$$\left\|\tilde{\boldsymbol{x}}_{e,t}\right\|_2 \le 1, \quad \tilde{\eta}_t(e) \text{ is R-sub Gaussian}, R \le 1. \tag{D.2}$$

With this notation, with probability at least $1-\delta$, it holds

$$\begin{aligned}\sigma^{-2}\|\boldsymbol{S}_{t-1}\|_{\boldsymbol{M}_{t-1}^{-1}} &= \sigma^{-2}\left\|\sum_{\tau=1}^{t}\sum_{e\in\tilde{\mathcal{E}}_\tau}\omega_{e,\tau}\boldsymbol{x}_e\eta_\tau(e)\right\|_{\boldsymbol{M}_{t-1}^{-1}} \\ &= \sigma^{-2}\left\|\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}\tilde{\boldsymbol{x}}_{e,\tau}\tilde{\eta}_\tau(e)\right\|_{\boldsymbol{M}_{t-1}^{-1}} \\ &\le \sigma^{-2}\sqrt{2\log(\frac{\det(\boldsymbol{M}_{t-1})^{1/2}\det(\boldsymbol{M}_0)^{-1/2}}{\delta})},\end{aligned}$$

where the last inequality is based on the Lemma J.2 and Theorem 1 in [2]. It is satisfied with probability at least $1-\delta$ with $\delta \in (0,1)$. Notice that $\det(\boldsymbol{M}_0) = \det(I) = 1$. Moreover, from the trace-determinant inequality, we have

$$\det(\boldsymbol{M}_{t-1})^{1/d} \le \frac{\mathrm{Tr}(\boldsymbol{M}_{t-1})}{d} = 1 + \frac{1}{d}\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}\tilde{\boldsymbol{x}}_{e,\tau}\tilde{\boldsymbol{x}}_{e,\tau}^T \le 1 + \frac{(t-1)E^*}{d},$$

where the last inequality follows from the assumption that $\left\|\tilde{\boldsymbol{x}}_{e,t}\right\|_2 \leq 1$ and $|\tilde{\mathcal{E}}_t| \leq E^*$. Thus, with probability at least $1-\delta$, the stochastic error can be bounded by

$$\sigma^{-2}\|\boldsymbol{S}_{t-1}\|_{\boldsymbol{M}_{t-1}^{-1}} \leq \sigma^{-2}\sqrt{d\log(1+\frac{tE^*}{d})+2\log(\frac{1}{\delta})}.$$

The corruption error can be bounded by

$$\begin{aligned}
&\sigma^{-2}\left\|\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}\omega_{e,\tau}\boldsymbol{x}_e c_{u,\tau}\right\|_{\boldsymbol{M}_{t-1}} \\
&\leq \sigma^{-2}\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}\left\|\boldsymbol{M}_{t-1}^{-1/2}\omega_{e,\tau}\boldsymbol{x}_e c_{u,\tau}\right\|_2 \\
&= \sigma^{-2}\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}|c_{u,\tau}|\times\omega_{e,\tau}\left\|\boldsymbol{M}_{t-1}^{-1/2}\boldsymbol{x}_e\right\|_2 \\
&\leq \sigma^{-2}\sum_{\tau=1}^{t-1}\sum_{e\in\tilde{\mathcal{E}}_\tau}|c_{u,\tau}|\lambda \leq \sigma^{-2}\lambda E^c C,
\end{aligned}$$

where the first inequality holds due to the inequality that $\|a+b\|_2 \leq \|a\|+\|b\|$. The second one holds due to the definition of $\omega_{e,t}$. By the assumption $\|\boldsymbol{\theta}\| \leq \Theta$, and after substituting the results, we arrive at

$$\begin{aligned}
&|\boldsymbol{x}_e^T(\hat{\boldsymbol{\theta}}_{t-1}-\boldsymbol{\theta})| \\
&\leq \|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}(\sigma^{-2}\sqrt{d\log(1+\frac{E^*T}{d})+2\log(\frac{1}{\delta})}+\sigma^{-2}\lambda E^c C+\Theta)
\end{aligned} \tag{D.3}$$

□

D.2 Proof of Theorem 6

Proof. The scaled regret at time t is $R_t^{\alpha\gamma} = f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma}f_{D,\boldsymbol{P}}(\mathcal{S}_t)$. Using the naive bound $R_t^{\alpha\gamma} \leq n-K$, the scaled cumulative regret can be decomposed into

$$\begin{aligned}
R^{\alpha\gamma}(T) = &\sum_{t=1}^{T}\mathbb{P}(\xi_{t-1})\mathbb{E}\left\{[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma}f_{D,\boldsymbol{P}}(\mathcal{S}_t)]\middle|\xi_{t-1}\right\} \\
&+\sum_{t=1}^{T}\mathbb{P}(\bar{\xi}_{t-1})\mathbb{E}\left\{[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma}f_{D,\boldsymbol{P}}(\mathcal{S}_t)]\middle|\bar{\xi}_{t-1}\right\}
\end{aligned}$$

$$\leq \sum_{t=1}^{T} \mathbb{E}\left\{ [f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma} f_{D,\boldsymbol{P}}(\mathcal{S}_t)] \Big| \xi_{t-1} \right\} + \sum_{t=1}^{T} \mathbb{P}(\bar{\xi}_{t-1})(n-K).$$

Select $\delta = \frac{1}{nT}$, then with probability at least $1 - \frac{1}{nT}$, the good event

$$\xi_{t-1} = \left\{ |\boldsymbol{x}_e^T(\hat{\boldsymbol{\theta}}_{\tau-1} - \boldsymbol{\theta})| \leq \beta \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{\tau-1}^{-1} \boldsymbol{x}_e}, \forall e \in \mathcal{E}, \forall \tau \leq t \right\}$$

happens $\forall t \in \{1, 2, \ldots, T\}$ with

$$\beta = \sigma^{-2} \sqrt{d \log(1 + \frac{E^* T}{d}) + 2\log(nT)} + \sigma^{-2} \lambda E^c C + \Theta$$

with $\forall t \in [T]$.

Besides, the ORACLE indicates that $f_{D,\hat{\boldsymbol{P}}_t}(\mathcal{S}_t) \geq \gamma \max_{\mathcal{S} \in \mathcal{V}} f_{D,\hat{\boldsymbol{P}}_t}(\mathcal{S})$ with probability at least α. Thus, $\mathbb{E}[f_{D,\hat{\boldsymbol{P}}_t}(\mathcal{S}_t)] \geq \alpha\gamma \mathbb{E}[\max_{\mathcal{S} \in \mathcal{V}} f_{D,\hat{\boldsymbol{P}}_t}(\mathcal{S})]$. Under the good event ξ_{t-1}, we have $p(e) \leq \hat{p}_{\tau-1}(e)$, $\forall e \in \mathcal{E}$, $\forall \tau \leq t$. Thus, based on the monotonicity of f in the probability weight, we have $\mathbb{E}[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt})] \leq \mathbb{E}[f_{D,\hat{\boldsymbol{P}}_{t-1}}(\mathcal{S}^{opt})]$. Combining all the previous inequalities, we can obtain [164, 157]

$$\mathbb{E}[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt})] \leq \mathbb{E}[f_{D,\hat{\boldsymbol{P}}_{t-1}}(\mathcal{S}^{opt})] \leq \mathbb{E}[\max_{\mathcal{S}:|\mathcal{S}|=K} f_{D,\hat{\boldsymbol{P}}_{t-1}}(\mathcal{S})] \leq \frac{1}{\alpha\gamma} \mathbb{E}[f_{D,\hat{\boldsymbol{P}}_{t-1}}(\mathcal{S}_t)]. \quad \text{(D.4)}$$

Therefore, under the good event ξ_{t-1}, $\forall t \in \{1, \ldots, T\}$ and according to the 1-Norm bounded smoothness condition, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \left[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt}) - \frac{1}{\alpha\gamma} f_{D,\boldsymbol{P}}(\mathcal{S}_t) \right] \\
&\leq \sum_{t=1}^{T} \left[\frac{1}{\alpha\gamma} f_{D,\hat{\boldsymbol{P}}_{t-1}}(\mathcal{S}_t) - \frac{1}{\alpha\gamma} f_{D,\boldsymbol{P}}(\mathcal{S}_t) \right] \\
&\leq \frac{B}{\alpha\gamma} \sum_{t=1}^{T} \sum_{e \in \tilde{\mathcal{E}}_t} |\hat{p}_{e,t-1} - p_e| \\
&= \frac{B}{\alpha\gamma} \sum_{t=1}^{T} \sum_{e \in \tilde{\mathcal{E}}_t} |\boldsymbol{x}_e^T(\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}) + \beta \|\boldsymbol{x}_e\|_{\boldsymbol{M}_{t-1}^{-1}}| \\
&\leq \frac{2B\beta}{\alpha\gamma} \sum_{t=1}^{T} \sum_{e \in \tilde{\mathcal{E}}_t} \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{t-1}^{-1} \boldsymbol{x}_e}, \\
&= \frac{2B\beta}{\alpha\gamma} \Bigg[\underbrace{\sum_{t:\omega_{e,t}=1} \sum_{e \in \tilde{\mathcal{E}}_t} \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{t-1}^{-1} \boldsymbol{x}_e}}_{I_1} + \underbrace{\sum_{t:\omega_{e,t}<1} \sum_{e \in \tilde{\mathcal{E}}_t} \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{t-1}^{-1} \boldsymbol{x}_e}}_{I_2} \Bigg]
\end{aligned} \quad \text{(D.5)}$$

where $\tilde{\mathcal{E}}_t$ refers to the set of observed edges at time step t.

Definition 6. *For any time step t and any directed edge $e \in \mathcal{E}$, we define the event*

$$O_t(e) = \{\textit{edge } e \textit{ is observed at round } t\}. \tag{D.6}$$

Based on the Definition 6, we have

$$\sum_{e\in\tilde{\mathcal{E}}_t} \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{t-1}^{-1} \boldsymbol{x}_e} = \sum_{e\in\mathcal{E}} \mathbb{1}(O_t(e)) \sqrt{\boldsymbol{x}_e^T \boldsymbol{M}_{t-1}^{-1} \boldsymbol{x}_e}. \tag{D.7}$$

For the term I_1 defined in Equation (D.5), we consider for all rounds $t \in [T]$, there exists $e \in \tilde{\mathcal{E}}_t$ with $\omega_{e,t} = 1$ and we assume these rounds can be listed as $\{t_1, t_2, \ldots, t_q\}$ for simplicity. With this notation, for each $i \leq q$, we can construct the auxiliary covariance matrix $\boldsymbol{A}_i = \boldsymbol{I} + \frac{1}{\sigma^2}\sum_{j=1}^{i}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{t_j}(e))\omega_{e,t_j}\boldsymbol{x}_e\boldsymbol{x}_e^T$ [164]. According to the definition of matrix $\boldsymbol{M}_t$, we have

$$\boldsymbol{M}_{t_i} \succeq \boldsymbol{I} + \frac{1}{\sigma^2}\sum_{j=1}^{i}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{t_j}(e))\omega_{e,t_j}\boldsymbol{x}_e\boldsymbol{x}_e^T = \boldsymbol{A}_{u,i}. \tag{D.8}$$

According to Lemma 12, we have

$$\boldsymbol{x}_e^T \boldsymbol{M}_{t_i}^{-1} \boldsymbol{x}_e \leq \boldsymbol{x}_e^T \boldsymbol{A}_i^{-1} x_e \tag{D.9}$$

Therefore, the term I_1 defined in Equation (D.5) is bounded as shown in the following lemma.

Lemma 13. *For time step $t = 1, \ldots, T$, and I_1 as defined in Equation (D.5), we have*

$$I_1 \leq \sqrt{\frac{TdE^* \log(1 + \frac{TE^*}{d\sigma^2})}{\log(1 + \frac{1^2}{\sigma^2})}} \tag{D.10}$$

Proof. Define $z_{e,i} = \sqrt{\boldsymbol{x}_e^T \boldsymbol{A}_{i-1}^{-1} \boldsymbol{x}_e}$, and we have

$$\boldsymbol{A}_i = \boldsymbol{A}_{i-1} + \frac{1}{\sigma^2}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{t_i}(e))\boldsymbol{x}_e\boldsymbol{x}_e^T. \tag{D.11}$$

For all $i \in \{1, 2, \ldots, q\}$, $e \in \tilde{\mathcal{E}}_{t_i}$ (e is observed at time step t_i), we have that

$$\det(\boldsymbol{A}_i) \geq \det(\boldsymbol{A}_{i-1} + \frac{1}{\sigma^2}\boldsymbol{x}_e\boldsymbol{x}_e^T)$$

$$
\begin{aligned}
&= \det\left[\boldsymbol{A}_{i-1}^{\frac{1}{2}}\left(\boldsymbol{I}+\frac{1}{\sigma}\boldsymbol{A}_{i-1}^{-\frac{1}{2}}\boldsymbol{x}_e\boldsymbol{x}_e^T\frac{1}{\sigma}\boldsymbol{A}_{i-1}^{-\frac{1}{2}}\right)\boldsymbol{A}_{i-1}^{\frac{1}{2}}\right] \\
&= \det(\boldsymbol{A}_{i-1})\left(1+\frac{1}{\sigma^2}\boldsymbol{x}_e\boldsymbol{A}_{i-1}^{-1}\boldsymbol{x}_e^T\right) \\
&= \det(\boldsymbol{A}_{u,i-1})\left(1+\frac{z_{e,i}^2}{\sigma^2}\right). \qquad \text{(D.12)}
\end{aligned}
$$

Hence, we have

$$
\det(\boldsymbol{A}_i)^{|\tilde{\mathcal{E}}_{t_i}|} \geq \det(\boldsymbol{A}_{i-1})^{|\tilde{\mathcal{E}}_{t_i}|}\prod_{e\in\tilde{\mathcal{E}}_{t_i}}\left(1+\frac{z_{e,i}^2}{\sigma^2}\right).
$$

Denote E^* as defined in Definition 4, it is easy to obtain $|\tilde{\mathcal{E}}_{t_i}| \leq E^*$, $\forall i \in \{1,2,\ldots,q\}$ and we have

$$
\det(\boldsymbol{A}_{t_q})^{E^*} \geq \det(\boldsymbol{A}_0)^{E^*}\prod_{i=1}^{q}\prod_{e\in\tilde{\mathcal{E}}_{t_i}}\left(1+\frac{z_{e,i}^2}{\sigma^2}\right).
$$

$\boldsymbol{A}_{u,0} = \boldsymbol{I}$, $\forall u \in \mathcal{V}$. Besides, based on the algorithm, we have

$$
\mathrm{Tr}(\boldsymbol{A}_{t_q}) = \mathrm{Tr}(\boldsymbol{I}+\frac{1}{\sigma^2}\sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}}\boldsymbol{x}_e\boldsymbol{x}_e^T) = d+\frac{1}{\sigma^2}\sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}}\|\boldsymbol{x}_e\|^2 \leq d+\frac{TE^*}{\sigma^2}, \qquad \text{(D.13)}
$$

the last inequality is based on the bound of $\|x_e\| \leq 1$, $\forall e \in \mathcal{E}$ and $t_q \leq T$. According to the trace-determinant inequality, we have $\frac{1}{d}\mathrm{Tr}(A_{t_q}) \geq [\det(A_{t_q})]^{\frac{1}{d}}$, thus we have [164]

$$
\left(1+\frac{TE^*L^2}{d\sigma^2}\right)^{dE^*} \geq \left[\frac{1}{d}\mathrm{Tr}(\boldsymbol{A}_{t_q})\right]^{dE^*} \geq [\det(\boldsymbol{A}_{t_q})]^{E^*} \geq \prod_{i=1}^{q}\prod_{e\in\tilde{\mathcal{E}}_{t_{i-1}}}\left(1+\frac{z_{e,i}^2}{\sigma^2}\right). \qquad \text{(D.14)}
$$

Take the logarithm on the both sides, we have

$$
dE^*\log(1+\frac{TE^*L^2}{d\sigma^2}) \geq \sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}}\log\left(1+\frac{z_{e,i}^2}{\sigma^2}\right), \qquad \text{(D.15)}
$$

where $z_{e,i}^2 = \boldsymbol{x}_e^T\boldsymbol{A}_{i-1}^{-1}\boldsymbol{x}_e \leq \boldsymbol{x}_e^T\boldsymbol{A}_{u,0}^{-1}\boldsymbol{x}_e = \|x_e\|^2 \leq 1$. And it is easy to prove that for any $y \in [0,1]$, we have $y \leq \frac{\log(1+\frac{y}{\sigma^2})}{\log(1+\frac{1}{\sigma^2})}$. Thus, we have $z_{e,i}^2 \leq \frac{\log(1+\frac{z_{e,i}^2}{\sigma^2})}{\log(1+\frac{1}{\sigma^2})}$.

And then, it is satisfied that

$$\sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}} z_{e,i}^2 \le \frac{1}{\log(1+\frac{1}{\sigma^2})}\sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}} \log(1+\frac{z_{e,i}^2}{\sigma^2}) \le \frac{dE^*\log(1+\frac{TE^*L^2}{d\sigma^2})}{\log(1+\frac{1}{\sigma^2})}. \tag{D.16}$$

Finally, with Cauchy-Schwarz inequality, we can obtain

$$\begin{aligned}
I_1 &= \sum_{t:\omega_{e,t}=1}\sum_{e\in\mathcal{E}} \mathbb{1}(O_t(e))\sqrt{\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e} \\
&\le \sum_{i=1}^{q}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{t_i}(e))\sqrt{\boldsymbol{x}_e^T\boldsymbol{M}_{t_{i-1}}^{-1}\boldsymbol{x}_e} \\
&\le \sum_{i=1}^{q}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{t_i}(e))\sqrt{\boldsymbol{x}_e^T\boldsymbol{A}_{i-1}^{-1}\boldsymbol{x}_e} \\
&= \sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_{i-1}}} z_{e,i} \le \sqrt{qE^*\left[\sum_{i=1}^{q}\sum_{e\in\tilde{\mathcal{E}}_{t_i}} z_{e,i}^2\right]} \\
&\le E^*\sqrt{\frac{Td\log(1+\frac{TE^*L^2}{d\sigma^2})}{\log(1+\frac{1^2}{\sigma^2})}}.
\end{aligned} \tag{D.17}$$

That completes the proof of Lemma 13. □

For the second term I_2 defined in Equation (D.5), according to the definition for weight $\omega_{e,t} < 1$, we have $\omega_{e,t} = \frac{\lambda}{\sqrt{\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e}}$, which implies that

$$I_2 = \sum_{t:\omega_{e,t}<1}\sum_{e\in\tilde{\mathcal{E}}_t}\sqrt{\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e} = \sum_{t:\omega_{e,t}<1}\sum_{e\in\tilde{\mathcal{E}}_t}\omega_{e,t}\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e/\lambda,$$

where the second equation holds due to the definition of $\omega_{e,t}$. Now, we assume the rounds with weight $\omega_{e,t} < 1$ can be listed as $\{\tau_1,\ldots,\tau_k\}$. And we introduce the vector $\tilde{\boldsymbol{x}}_{e,i} = \sqrt{\omega_{e,\tau_i}}\boldsymbol{x}_e$ if at time step τ_i, edge e is observable and matrix $\tilde{\boldsymbol{M}}_i$ as

$$\tilde{\boldsymbol{M}}_i = \boldsymbol{I} + \frac{1}{\sigma^2}\sum_{j=1}^{i}\omega_{e,\tau_j}\boldsymbol{x}_e\boldsymbol{x}_e^T = I + \sum_{j=1}^{i}\tilde{\boldsymbol{x}}_{e,i}\tilde{\boldsymbol{x}}_{e,i}^T. \tag{D.18}$$

According to Lemma 12, we also have $(\tilde{\boldsymbol{M}}_i)^{-1} \succeq \boldsymbol{M}_{\tau_i}^{-1}$. Therefore, for each $i\in\{1,\ldots,k\}$, we have

$$\boldsymbol{x}_e^T(\tilde{\boldsymbol{M}}_i)^{-1}\boldsymbol{x}_e^T \ge \boldsymbol{x}_e^T\boldsymbol{M}_{\tau_i}^{-1}\boldsymbol{x}_e^T. \tag{D.19}$$

Follow the same process in the proof of Lemma 13, define $\tilde{z}_{e,i} = \sqrt{\tilde{\boldsymbol{x}}_{e,i}^T(\tilde{\boldsymbol{M}}_{i-1})^{-1}\tilde{\boldsymbol{x}}_{e,i}}$ and we have

$$\tilde{\boldsymbol{M}}_i = \tilde{\boldsymbol{M}}_{i-1} + \frac{1}{\sigma^2}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{\tau_i}(e))\tilde{\boldsymbol{x}}_{e,i}\tilde{\boldsymbol{x}}_{e,i}^T \tag{D.20}$$

Similarly, for all $e \in \tilde{\mathcal{E}}_{\tau_i}$ and $i \in \{1,2,\ldots,k\}$, we can easily obtain

$$\det(\tilde{\boldsymbol{M}}_i) \geq \det(\tilde{\boldsymbol{M}}_{i-1})(1+\frac{(\tilde{z}_{e,i})^2}{\sigma^2}), \tag{D.21}$$

and

$$\det(\tilde{\boldsymbol{M}}_i)^{|\tilde{\mathcal{E}}_{\tau_i}|} \geq \det(\tilde{\boldsymbol{M}}_{i-1})^{|\tilde{\mathcal{E}}_{\tau_i}|} \prod_{e\in\tilde{\mathcal{E}}_{\tau_{i-1}}} \left(1+\frac{(\tilde{z}_{e,i-1})^2}{\sigma^2}\right).$$

Follow the same process in Lemma 13, it is satisfied that

$$\sum_{i=1}^{k}\sum_{e\in\tilde{\mathcal{E}}_{\tau_i}} (\tilde{z}_{e,i})^2 \leq \frac{1}{\log(1+\frac{1}{\sigma^2})}\sum_{i=1}^{k}\sum_{e\in\tilde{\mathcal{E}}_{\tau_i}} \log(1+\frac{(\tilde{z}_{e,i})^2}{\sigma^2}) \leq \frac{dE^*\log(1+\frac{TE^*}{d\sigma^2})}{\log(1+\frac{1}{\sigma^2})}. \tag{D.22}$$

And then

$$\begin{aligned}
I_2 &= \sum_{t:\omega_{e,t}<1}\sum_{e\in\mathcal{E}} \mathbb{1}(O_t(e))\sqrt{\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e} \\
&= \sum_{t:\omega_{e,t}<1}\sum_{e\in\mathcal{E}} \mathbb{1}(O_t(e))\omega_{e,t}\boldsymbol{x}_e^T\boldsymbol{M}_{t-1}^{-1}\boldsymbol{x}_e/\lambda \\
&\leq \sum_{i=1}^{k}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{\tau_i}(e))\omega_{e,\tau_i}\boldsymbol{x}_e^T M_{\tau_{i-1}}^{-1}\boldsymbol{x}_e/\lambda \\
&\leq \sum_{i=1}^{k}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{\tau_i}(e))\omega_{e,\tau_i}\boldsymbol{x}_e^T(\tilde{\boldsymbol{M}}_{i-1})^{-1}\boldsymbol{x}_e/\lambda \\
&= \sum_{i=1}^{k}\sum_{e\in\mathcal{E}} \mathbb{1}(O_{\tau_i}(e))\tilde{\boldsymbol{x}}_{e,i}^T\tilde{\boldsymbol{M}}_{i-1}^{-1}\tilde{\boldsymbol{x}}_{e,i}/\lambda \\
&= \sum_{i=1}^{k}\sum_{e\in\tilde{\mathcal{E}}_{\tau_i}} \frac{\tilde{z}_{e,i}^2}{\lambda} \leq \frac{dE^*\log(1+\frac{TE^*}{d\sigma^2})}{\lambda\log(1+\frac{1}{\sigma^2})}
\end{aligned} \tag{D.23}$$

Therefore, combine with the bound of I_1 in Lemma 13, we can obtain

$$\sum_{t=1}^{T}\mathbb{E}\left\{[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt})-\frac{1}{\alpha\gamma}f_{D,\boldsymbol{P}}(\mathcal{S}_t)]\Big|\xi_{t-1}\right\}\leq\frac{2B\beta}{\alpha\gamma}\left(E^*\sqrt{\frac{Td\log(1+\frac{TE^*}{d\sigma^2})}{\log(1+\frac{1^2}{\sigma^2})}}+\frac{dE^*\log(1+\frac{TE^*}{d\sigma^2})}{\lambda\log(1+\frac{1}{\sigma^2})}\right).$$

Select $\lambda=\frac{\sqrt{d}}{E^cC}$, the $\alpha\gamma$-scaled regret is upper bounded

$$\begin{aligned}R^{\alpha\gamma}(T)&\leq\sum_{t=1}^{T}\mathbb{E}\left\{[f_{D,\boldsymbol{P}}(\mathcal{S}^{opt})-\frac{1}{\alpha\gamma}f_{D,\boldsymbol{P}}(\mathcal{S}_t)]\Big|\xi_{t-1}\right\}+\sum_{t=1}^{T}\mathbb{P}(\bar{\xi}_{t-1})(n-K)\\&\leq O(dBE^*\sqrt{T}\log(nT)+BE^*E^cCd\log(nT))\end{aligned}\tag{D.24}$$

□

Abbreviations

Symbol	Description
MAB	Multi-Armed Bandit
CSB	Combinatorial Semi-Bandit
UCB	Upper Confidence Bound
ML	Machine Learning
IM	Influence Maximization
OIM	Online Influence Maximization
MLB	Multi-task contextual Linear Bandit
GLR	Generalized Likelihood Ratio
CUSUM	Cumulative Sum
SEM	Structural Equation Models
CTS	Combinatorial Thompson Sampling
DAG	Directed Acyclic Graph
MSE	Mean Squared Error
RE	Restricted Eigenvalue
OSQP	Operator Splitting Quadratic Program
DTV	Directed Total Variation
CMAB	Combinatorial Multi-Armed Bandit
PSD	Positive Semi-Definite
LT	Linear Threshold
IC	Independent Cascade
ICSB	Independent Cascade Semi-Bandit

Bibliography

[1] Abbasi-Yadkori, Y., Pal, D., and Szepesvári, C. (2011a). Improved algorithms for linear stochastic bandits. *Proceedings of the 24th Annual Conference on Learning Theory*, pages 231–247.

[2] Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011b). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, **24**.

[3] Abeille, M. and Lazaric, A. (2017). Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR.

[4] Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on Machine Learning*, **28**, 127–135.

[5] Akoglu, L., Tong, H., and Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, **29**, 626–688.

[6] Ariu, K., Abe, K., and Proutiere, A. (2022). Thresholded Lasso Bandit. In *Proceedings of the 39th International Conference on Machine Learning*, pages 878–928. PMLR.

[7] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, **3**, 397–422.

[8] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, **47**, 235–256.

[9] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, **32**, 48–77.

[10] Awerbuch, B. and Kleinberg, R. (2008). Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, **74**(1), 97–114.

[11] Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021). On multi-armed bandit designs for dose-finding clinical trials. *Journal of Machine Learning Research*, **22**, 4.

[12] Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, **19**(3), 357 – 367.

[13] Backstrom, L. and Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644.

[14] Baingana, B. and Giannakis, G. B. (2015). Switched dynamic structural equation models for tracking social network topologies. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 682–686.

[15] Baingana, B., Mateos, G., and Giannakis, G. B. (2014). Proximal-gradient algorithms for tracking cascades over social networks. *IEEE Journal of Selected Topics in Signal Processing*, **8**(4), 563–575.

[16] Ban, Y. and He, J. (2021). Local clustering in contextual multi-armed bandits. In *Proceedings of the Web Conference 2021*, pages 2335–2346.

[17] Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, **28**.

[18] Bastani, H. and Bayati, M. (2019). Online Decision Making with High-Dimensional Covariates. *Operations Research*.

[19] Basu, S., Kveton, B., Zaheer, M., and Szepesvari, C. (2021). No Regrets for Learning the Prior in Bandits. In *Advances in Neural Information Processing Systems*.

[20] Bazerque, J. A., Baingana, B., and Giannakis, G. B. (2013a). Identifiability of sparse structural equation models for directed and cyclic networks. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 839–842. IEEE.

[21] Bazerque, J. A., Baingana, B., and Giannakis, G. B. (2013b). Identifiability of sparse structural equation models for directed and cyclic networks. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 839–842.

[22] Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, **27**.

[23] Besson, L. and Kaufmann, E. (2019). The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits. *Proceedings of Machine Learning Research vol XX*, **1**, 35.

[24] Bilaj, S., Dhouib, S., and Maghsudi, S. (2024). Meta learning in bandits within shared affine subspaces. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR.

[25] Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. (2021). Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR.

[26] Borge-Holthoefer, J., Rivero, A., García, I., Cauhé, E., Ferrer, A., Ferrer, D., Francos, D., Iniguez, D., Pérez, M. P., Ruiz, G., *et al.* (2011). Structural and dynamical patterns on online social networks: the spanish may 15th movement as a case study. *PloS one*, **6**(8).

[27] Bouneffouf, D. and Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

[28] Bridgwater, A. and Bóta, A. (2021). Identifying regions most likely to contribute to an epidemic outbreak in a human mobility network. In *2021 Swedish Artificial Intelligence Society Workshop (SAIS)*, pages 1–4. IEEE.

[29] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg.

[30] Bull, M. (2021). The italian government response to covid-19 and the making of a prime minister. *Contemporary Italian Politics*, pages 1–17.

[31] Burgoon, M., Alvaro, E., Grandpre, J., and Voulodakis, M. (2002). Revisiting the theory of psychological reactance. *The persuasion handbook*, pages 213–232.

[32] Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology*, **9**(5), e1003068.

[33] Cao, Y., Wen, Z., Kveton, B., and Xie, Y. (2019). Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 418–427. PMLR.

[34] Cella, L. and Pontil, M. (2021). Multi-task and meta-learning with sparse linear bandits. In *Uncertainty in Artificial Intelligence*. PMLR.

[35] Cella, L., Lazaric, A., and Pontil, M. (2020). Meta-learning with stochastic linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR.

[36] Cella, L., Lounici, K., Pacreau, G., and Pontil, M. (2023). Multi-task representation learning with stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*.

[37] Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, **78**(5), 1404–1422.

[38] Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A gang of bandits. *Advances in neural information processing systems*, **26**.

[39] Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*.

[40] Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 199–208.

[41] Chen, W., Wang, C., and Wang, Y. (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1029–1038.

[42] Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR.

[43] Chen, W., Wang, Y., Yuan, Y., and Wang, Q. (2016). Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, **17**, 1746–1778.

[44] Chen, W., Wang, L., Zhao, H., and Zheng, K. (2021). Combinatorial semi-bandit in the non-stationary environment. In *Uncertainty in Artificial Intelligence*, pages 865–875. PMLR.

[45] Cheng, X. and Maghsudi, S. (2023). Distributed consensus algorithm for decision-making in multi-agent multi-armed bandit. *arXiv preprint arXiv:2306.05998*.

[46] Cheng, X., Pan, C., and Maghsudi, S. (2023). Parallel online clustering of bandits via hedonic game. In *International Conference on Machine Learning*, pages 5485–5503. PMLR.

[47] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings.

[48] Di Lorenzo, P., Banelli, P., Isufi, E., Barbarossa, S., and Leus, G. (2018). Adaptive graph signal processing: Algorithms and optimal sampling strategies. *IEEE Transactions on Signal Processing*, **66**(13), 3584–3598.

[49] Dong, X., Thanou, D., Frossard, P., and Vandergheynst, P. (2016). Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, **64**(23), 6160–6173.

[50] Dong, X., Thanou, D., Rabbat, M., and Frossard, P. (2019). Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*.

[51] Dore, M. (2018). Controversial humor in advertising: Social and cultural implications. In *Not All Claps and Cheers*, pages 132–145. Routledge.

[52] Duncan, A. (2004). Powers of the adjacency matrix and the walk matrix.

[53] Easley, D., Kleinberg, J., *et al.* (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*, volume 1. Cambridge university press Cambridge.

[54] Fitzsimons, G. J. and Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, **23**(1), 82–94.

[55] Fontan, A. and Altafini, C. (2021). On the properties of laplacian pseudoinverses. In *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE.

[56] Gai, Y., Krishnamachari, B., and Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, **20**(5), 1466–1478.

[57] Gao, S., Zhang, Z., Su, S., Wen, J., and Sun, L. (2020). Fair-aware competitive event influence maximization in social networks. *IEEE Transactions on Network Science and Engineering*, **7**(4), 2528–2540.

[58] Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer.

[59] Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *International conference on machine learning*, pages 757–765. PMLR.

[60] Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. (2017). On context-dependent clustering of bandits. In *International Conference on machine learning*, pages 1253–1262. PMLR.

[61] Giannakis, G. B., Shen, Y., and Karanikolas, G. V. (2018). Topology identification and learning over graphs: Accounting for nonlinearities and dynamics. *Proceedings of the IEEE*, **106**(5), 787–807.

[62] Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.

[63] Guaitoli, G. and Pancrazi, R. (2021). Covid-19: Regional policies and local infection risk: Evidence from italy with a modelling study. *The Lancet Regional Health-Europe*, **8**, 100169.

[64] Gupta, A., Koren, T., and Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR.

[65] Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396.

[66] Hartland, C., Baskiotis, N., Gelly, S., Sebag, M., and Teytaud, O. (2007). Change point detection and meta-bandits for online learning in dynamic environments. In *CAp 2007: 9è Conférence francophone sur l'apprentissage automatique*, pages 237–250.

[67] He, J., Zhou, D., Zhang, T., and Gu, Q. (2022). Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in Neural Information Processing Systems*, **35**, 34614–34625.

[68] Herbster, M., Pasteris, S., Vitale, F., and Pontil, M. (2021). A gang of adversarial bandits. *Advances in Neural Information Processing Systems*, **34**.

[69] Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

[70] Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, **17**.

[71] Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. (2021). Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*. PMLR.

[72] Huo, X. and Fu, F. (2017). Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, **4**(11), 171377.

[73] Huyuk, A. and Tekin, C. (2019). Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1322–1330. PMLR.

[74] Jung, A. (2020). Networked Exponential Families for Big Data Over Networks. *IEEE Access*, **8**.

[75] Jung, A. and Vesselinova, N. (2019). Analysis of network lasso for semi-supervised regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 380–387. PMLR.

[76] Jung, A., Tran, N., and Mara, A. (2018). When Is Network Lasso Accurate? *Frontiers in Applied Mathematics and Statistics*, **3**.

[77] Kalofolias, V. (2016). How to learn a graph from smooth signals. In *Artificial intelligence and statistics*, pages 920–929. PMLR.

[78] Kaplan, D. (2008). *Structural equation modeling: Foundations and extensions*, volume 10. SAGE publications.

[79] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146.

[80] Khatri, I., Gupta, A., Choudhry, A., Tyagi, A., Vishwakarma, D. K., and Prasad, M. (2023). Cks: a community-based k-shell decomposition approach using community bridge nodes for influence maximization (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 16240–16241.

[81] Kim, G.-S. and Paik, M. C. (2019). Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, **32**.

[82] Kong, F., Xie, J., Wang, B., Yao, T., and Li, S. (2023a). Online influence maximization under decreasing cascade model. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2197–2204, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[83] Kong, F., Xie, J., Wang, B., Yao, T., and Li, S. (2023b). Online influence maximization under decreasing cascade model. *arXiv preprint arXiv:2305.15428*.

[84] Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015a). Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR.

[85] Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015b). Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, **28**.

[86] Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. (2015c). Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR.

[87] Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. (2021). Meta-thompson sampling. In *International Conference on Machine Learning*. PMLR.

[88] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, **6**(1), 4–22.

[89] Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, **20**(1), 96–1.

[90] Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, **29**.

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

[92] Lei, S., Maniu, S., Mo, L., Cheng, R., and Senellart, P. (2015). Online influence maximization. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654.

[93] Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

[94] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010a). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670.

[95] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010b). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

[96] Li, S., Karatzoglou, A., and Gentile, C. (2016). Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548.

[97] Li, S., Chen, W., and Leung, K.-S. (2019a). Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*.

[98] Li, S., Kong, F., Tang, K., Li, Q., and Chen, W. (2020). Online influence maximization under linear threshold model. *Advances in Neural Information Processing Systems*, **33**, 1192–1204.

[99] Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

[100] Li, Y., Lou, E. Y., and Shan, L. (2019b). Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*.

[101] Liu, F., Lee, J., and Shroff, N. (2018). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[102] Liu, Y. and Shrum, L. J. (2002). What is interactivity and is it always such a good thing? implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of advertising*, **31**(4), 53–64.

[103] Lorenzo, P., Barbarossa, S., and Banelli, P. (2018). Sampling and recovery of graph signals. In *Cooperative and Graph Signal Processing*, pages 261–282. Elsevier.

[104] Lykouris, T., Mirrokni, V., and Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122.

[105] Maghsudi, S. and Hossain, E. (2016). Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*, **23**, 64–73.

[106] Maillard, O.-A. (2019). *Mathematics of statistical sequential decision making*. Ph.D. thesis, Université de Lille, Sciences et Technologies.

[107] Mastakouri, A. and Schölkopf, B. (2020). Causal analysis of covid-19 spread in germany. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3153–3163. Curran Associates, Inc.

[108] Matz, G. and Dittrich, T. (2020). Learning signed graphs from data. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5570–5574. IEEE.

[109] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, **27**(1), 415–444.

[110] Mellor, J. and Shapiro, J. (2013). Thompson sampling in switching environments with bayesian online change detection. In *Artificial intelligence and statistics*, pages 442–450. PMLR.

[111] Misra, K., Schwartz, E. M., and Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, **38**(2), 226–252.

[112] Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124.

[113] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**(1), 115–132.

[114] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, **103**(23), 8577–8582.

[115] Nguyen, T. T. and Lauw, H. W. (2014). Dynamic clustering of contextual multi-armed bandits. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1959–1962.

[116] Nourani-Koliji, B., Ghoorchian, S., and Maghsudi, S. (2022). Linear combinatorial semi-bandit with causally related rewards. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization.

[117] Nourani-Koliji, B., Bilaj, S., Balef, A. R., and Maghsudi, S. (2023). Piecewise-stationary combinatorial semi-bandit with causally related rewards. *arXiv preprint arXiv:2307.14138*.

[118] Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K. E., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cattarino, L., Cooper, L. V., *et al.* (2021). Reduction in mobility and covid-19 transmission. *Nature communications*, **12**(1), 1–9.

[119] Oh, M.-H., Iyengar, G., and Zeevi, A. (2021). Sparsity-Agnostic Lasso Bandit. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8271–8280. PMLR.

[120] Ortega, A., Frossard, P., Kova vcević, J., Moura, J. M., and Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, **106**(5), 808–828.

[121] Ortelli, F. and van de Geer, S. (2019). Synthesis and analysis in total variation regularization. *arXiv preprint arXiv:1901.06418*.

[122] Pearl, J. (2009). *Causality*. Cambridge university press.

[123] Peleg, A., Pearl, N., and Meir, R. (2022). Metalearning linear bandits by prior update. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR.

[124] Ranjan, G. and Zhang, Z.-L. (2013). Geometry of complex networks and topological centrality. *Physica A: Statistical Mechanics and its Applications*.

[125] Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70.

[126] Robbins, H. (1952). Some aspects of the sequential design of experiments.

[127] Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., *et al.* (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, **11**(1), 1–96.

[128] Sardellitti, S., Barbarossa, S., and Di Lorenzo, P. (2017). On the graph fourier transform for directed graphs. *IEEE Journal of Selected Topics in Signal Processing*, **11**(6), 796–811.

[129] Sardellitti, S., Barbarossa, S., and Di Lorenzo, P. (2019). Graph topology inference based on sparsifying transform learning. *IEEE Transactions on Signal Processing*, **67**(7), 1712–1727.

[130] Schäfer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**(6), 754–764.

[131] Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pages 3057–3066. PMLR.

[132] Shafipour, R., Segarra, S., Marques, A. G., and Mateos, G. (2021). Identifying the topology of undirected networks from diffused non-stationary graph signals. *IEEE Open Journal of Signal Processing*, **2**, 171–189.

[133] Shen, W., Wang, J., Jiang, Y.-G., and Zha, H. (2015). Portfolio choices with orthogonal bandit learning. In *Twenty-fourth international joint conference on artificial intelligence*.

[134] Shen, Y., Baingana, B., and Giannakis, G. B. (2017). Tensor decompositions for identifying directed graph topologies and tracking dynamic networks. *IEEE Transactions on Signal Processing*, **65**(14), 3675–3687.

[135] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, **30**(3), 83–98.

[136] Spielman, D. (2012). Spectral graph theory. *Combinatorial scientific computing*, **18**, 18.

[137] Stein, S., Eshghi, S., Maghsudi, S., Tassiulas, L., Bellamy, R. K. E., and Jennings, N. R. (2017). Heuristic algorithms for influence maximization in partially observable social networks. In *SocInf@IJCAI*.

[138] Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020). OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, **12**(4), 637–672.

[139] Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, **2009**.

[140] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[141] Tang, S., Zhou, Y., Han, K., Zhang, Z., Yuan, J., and Wu, W. (2017). Networked stochastic multi-armed bandits with combinatorial strategies. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 786–793. IEEE.

[142] Tang, Y., Xiao, X., and Shi, Y. (2014). Influence maximization: Near-optimal time complexity meets practical efficiency. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 75–86.

[143] Thanou, D., Shuman, D. I., and Frossard, P. (2014). Learning parametric dictionaries for signals on graphs. *IEEE Transactions on Signal Processing*, **62**(15), 3849–3862.

[144] Thanou, D., Dong, X., Kressner, D., and Frossard, P. (2017). Learning heat diffusion graphs. *IEEE Transactions on Signal and Information Processing over Networks*, **3**(3), 484–499.

[145] Thompson, W. R. (1933a). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

[146] Thompson, W. R. (1933b). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

[147] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

[148] Toni, L. and Frossard, P. (2018). Spectral mab for unknown graph processes. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 116–120. IEEE.

[149] Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, **16**, 262 – 270.

[150] Trovo, F., Paladino, S., Restelli, M., and Gatti, N. (2020). Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, **68**, 311–364.

[151] Tsitsvero, M., Barbarossa, S., and Di Lorenzo, P. (2016). Signals on graphs: Uncertainty principle and sampling. *IEEE Transactions on Signal Processing*, **64**(18), 4845–4860.

[152] Valko, M. (2016). *Bandits on graphs and structures*. Ph.D. thesis, École normale supérieure de Cachan-ENS Cachan.

[153] Valko, M., Munos, R., Kveton, B., and Kocák, T. (2014). Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54. PMLR.

[154] Van Mieghem, P., Devriendt, K., and Cetinay, H. (2017). Pseudoinverse of the laplacian and best spreader node in a network. *Physical Review E*.

[155] Van Parys, B. and Golrezaei, N. (2024). Optimal learning for structured bandits. *Management Science*, **70**(6), 3951–3998.

[156] Vaswani, S., Lakshmanan, L., Schmidt, M., *et al.* (2015). Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*.

[157] Vaswani, S., Kveton, B., Wen, Z., Ghavamzadeh, M., Lakshmanan, L. V., and Schmidt, M. (2017). Model-independent online learning for influence maximization. In *International Conference on Machine Learning*, pages 3530–3539. PMLR.

[158] Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. (2020a). Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR.

[159] Vernade, C., Gyorgy, A., and Mann, T. (2020b). Non-stationary delayed bandits with intermediate observations. In *International Conference on Machine Learning*, pages 9722–9732. PMLR.

[160] Wang, D., Wu, X., Li, C., Han, J., and Yin, J. (2022). The impact of geo-environmental factors on global covid-19 transmission: A review of evidence and methodology. *Science of the Total Environment*, page 154182.

[161] Wang, Q. and Chen, W. (2017). Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, **30**.

[162] Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016). Trend filtering on graphs. *Journal of Machine Learning Research*, **17**(105), 1–41.

[163] Wang, Z., Xie, J., Yu, T., Li, S., and Lui, J. (2023). Online corrupted user detection and regret minimization. *arXiv preprint arXiv:2310.04768*.

[164] Wen, Z., Kveton, B., Valko, M., and Vaswani, S. (2017). Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in Neural Information Processing Systems*, **30**.

[165] Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T., and Langer, T. (2020). A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, **37**, 1–12.

[166] Wu, Q., Li, Z., Wang, H., Chen, W., and Wang, H. (2019). Factorization bandits for online influence maximization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 636–646.

[167] Wu, X., Fu, L., Zhang, Z., Long, H., Meng, J., Wang, X., and Chen, G. (2020). Evolving influence maximization in evolving networks. *ACM Transactions on Internet Technology (TOIT)*, **20**(4), 1–31.

[168] Yang, K. and Toni, L. (2018). Graph-based recommendation system. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 798–802. IEEE.

[169] Yang, K., Toni, L., and Dong, X. (2020). Laplacian-regularized graph bandits: Algorithms and theoretical analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 3133–3143. PMLR.

[170] Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. J. (2020). Graphical models meet bandits: A variational thompson sampling approach. In *International Conference on Machine Learning*, pages 10902–10912. PMLR.

[171] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

[172] Zhao, H., Zhou, D., and Gu, Q. (2021). Linear contextual bandits with adversarial corruptions. *arXiv preprint arXiv:2110.12615*.

[173] Zhou, H., Wang, L., Varshney, L., and Lim, E.-P. (2020). A near-optimal change-detection based algorithm for piecewise-stationary combinatorial semi-bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6933–6940.

[174] Zhou, Q., Zhang, X., Xu, J., and Liang, B. (2017). Large-scale bandit approaches for recommender systems. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24*, pages 811–821. Springer.

[175] Zimmert, J., Luo, H., and Wei, C.-Y. (2019). Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR.

[176] Zuo, J., Liu, X., Joe-Wong, C., Lui, J. C., and Chen, W. (2022). Online competitive influence maximization. In *International Conference on Artificial Intelligence and Statistics*, pages 11472–11502. PMLR.