Improved Exploration through Transfer Learning in Multi-Armed Bandits

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Steven Bilaj

aus Clausthal-Zellerfeld

Tübingen

2025

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

Tag der mündlichen Qualifikation:	19.03.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr.-Ing. Setareh Maghsudi
2. Berichterstatter:	Prof. Dr. Georg Martius

Für meine Familie, meine Freunde und meine liebevolle Partnerin, die mich auf diesem Weg unterstützt hat

Abstract

Reinforcement learning is a branch of machine learning that focuses on training an agent to interact with a dynamic environment to maximize its expected cumulative reward. In contrast to offline learning problems, where the training data is immediately available, the agent typically assembles its own data set by interacting with the environment. Since model performances heavily depend on the gathered training data, a key challenge is determining an exploration strategy at the potential cost of low immediate rewards to find a policy. This is known as the exploration-exploitation trade-off.
This dissertation examines how to effectively re-balance the exploration-exploitation trade-off by transferring knowledge between tasks with a shared structure to enhance the agent's overall performance. The primary focus is on developing algorithms that reduce uncertainties in the model estimations by transferring information given various assumptions while providing theoretical bounds on the regret. The contributions span single-task transfer, meta-learning, multi-task learning and non-stationary environments in the multi-armed bandit setting.
In a straightforward task-to-task transfer approach, where an expert is assumed to be available to the learner, we propose a dynamic convex combination of the expert and target model. We prove that when the expert's parameter vector is close to the true task related feature vector, the learner can exploit the expert's knowledge in the early steps of the algorithm and reduce the regret with high probability.
A generalization would be to investigate feature vectors that are close in a subspace. We address this idea within the concept of meta learning, where an agent sequentially interacts with multiple tasks sampled from a common meta distribution. Under the assumption of a low-dimensional subspace structure in the meta distribution, we propose a framework to estimate the subspace with projection matrices and exploit it as prior information within an OFUL and Thompson sampling based algorithm. With each task the agent interacts with, it improves its estimation of the projections for exploitation in future tasks. Theoretical guarantees are provided with an emphasis on an improvement on the regret bound with respect to the dimensionality.
In a clustered setting, we assume that tasks are grouped in clusters such that only tasks of the same cluster share the same feature vector. When the number of clusters is lower than the number of dimensions, it can be interpreted as a special case of the low-dimensional subspace setting. We explore the general clustered setting in a multi-task framework, where an agent interacts with a fixed number of tasks in

parallel. The agent has access to a graph, where each node is associated with a different task. We introduce a network lasso based bandit algorithm that exploits the given graph such that it implicitly learns the cluster structure. Theoretical bounds show that, with a well suited graph, this approach offers significant improvements over other baselines.
Finally, we address dynamic environments or piecewise-stationary settings, where the agent typically discards all collected data points upon detecting changes in the environment and retrains its model from scratch. Instead, we propose an algorithm that only discards data points directly associated with the environmental change and retains the rest. We show that intelligent transfer of data from previous segments can reduce exploration after each change and increase overall reward.
This dissertation thus proposes multiple algorithms for transferring information in several multi-armed bandit settings with the purpose of optimizing exploration. We provide both theoretical guarantees and empirical evaluations showcasing significant improvements over existing methods.

Kurzfassung

Verstärkendes Lernen (Reinforcement Learning) ist ein Teilgebiet des maschinellen Lernens, das darauf abzielt, einen Agenten so zu trainieren, dass er mit einer dynamischen Umgebung interagiert, um seine erwartete kumulative Belohnung zu maximieren. Im Gegensatz zu Offline-Lernproblemen, bei denen die Trainingsdaten unmittelbar zur Verfügung stehen, stellt der Agent typischerweise seinen eigenen Datensatz durch die Interaktion mit der Umgebung zusammen. Da die Leistung der Modelle stark von den gesammelten Trainingsdaten abhängt, besteht eine zentrale Herausforderung darin, eine geeignete Explorationsstrategie festzulegen, gegebenenfalls auf Kosten geringer unmittelbarer Belohnungen, um eine optimale Strategie zu lernen. Dies ist als Explorations-Exploitation-Trade-off bekannt.
Diese Dissertation untersucht, wie sich der Explorations-Exploitation-Trade-off durch den Transfer von Wissen zwischen Aufgaben mit gemeinsamer Struktur effektiv neu ausbalancieren lässt, um die Gesamtperformance des Agenten zu steigern. Der Schwerpunkt liegt auf der Entwicklung von Algorithmen, die die Unsicherheit in den Modellen reduzieren, indem Informationen unter verschiedenen Annahmen übertragen werden und dabei theoretische Obergrenzen des Regret liefern. Die Beiträge dieser Dissertation umfassen Single-Task-Transfer, Meta-Lernen, Multi-Task-Lernen sowie nicht-stationäre Umgebungen im Setting des Mehrarmigen Banditen.
In einem einfachen Ansatz für den Transfer von Aufgabe zu Aufgabe, bei dem angenommen wird, dass dem Lernenden ein Experte zur Verfügung steht, schlagen wir eine dynamische konvexe Kombination zwischen dem Experten- und dem Zielmodell vor. Wir beweisen, dass, wenn der Parametervektor des Experten dem wahren, aufgabenbezogenen Feature-Vektor nahekommt, der Lernende das Wissen des Experten in den frühen Schritten des Algorithmus ausnutzen und den Regret mit hoher Wahrscheinlichkeit verringern kann.
Eine Verallgemeinerung besteht darin, Feature-Vektoren zu betrachten, die in einem Unterraum nahe beieinander liegen. Dies greifen wir im Konzept des Meta-Lernens auf, bei dem ein Agent nacheinander mit mehreren Aufgaben interagiert, die aus einer gemeinsamen Meta-Verteilung stammen. Unter der Annahme einer niedrig dimensionalen Unterraumstruktur in der Meta-Wahrscheinlichkeitsverteilung schlagen wir einen Algorithmus vor, um den Unterraum mit Hilfe von Projektionsmatrizen abzuschätzen und als Vorwissen in auf OFUL und Thompson-Sampling basierenden Algorithmen zu nutzen. Mit jeder Aufgabe, mit der der Agent interagiert, verbessert er seine Schätzung der Projektionen zur besseren Verwertung in zukünftigen Aufgaben. Theoretische Garantien werden insbesondere hinsichtlich

einer Verbesserung der Obergrenze des Regret in Bezug auf die Dimensionalität gegeben.
In einem geclusterten Setting nehmen wir an, dass Aufgaben in Cluster gruppiert sind, so dass Aufgaben desselben Clusters den gleichen Feature-Vektor haben. Ist die Anzahl der Cluster kleiner als die Anzahl der Dimensionen, so lässt sich dies als Spezialfall der niedrig dimensionalen Unterraumstruktur interpretieren. Wir untersuchen das allgemeine Cluster-Setting in einem Multi-Task-Setting, in dem ein Agent mit einer festen Anzahl an Aufgaben parallel interagiert. Der Agent hat Zugriff auf einen Graphen, wobei jeder Knoten einer anderen Aufgabe zugeordnet ist. Wir präsentieren einen auf Netzwerk-Lasso basierenden Banditenalgorithmus, der den gegebenen Graphen nutzt, um die Clusterstruktur implizit zu erlernen. Theoretische Grenzen zeigen, dass dieser Ansatz mit einem gut geeigneten Graphen signifikante Verbesserungen gegenüber anderen Algorithmen liefert.
Abschließend widmen wir uns dynamischen oder abschnittsweise-stationären Umgebungen, in denen der Agent typischerweise alle gesammelten Datenpunkte beim Erkennen von Veränderungen verwirft und sein Modell von Grund auf neu trainiert. Stattdessen schlagen wir einen Algorithmus vor, der nur die Datenpunkte verwirft, die direkt mit der Veränderung der Umgebung zusammenhängen, und den Rest beibehält. Wir zeigen, dass ein intelligenter Transfer von Daten aus vorherigen Segmenten die Exploration nach jeder Änderung reduzieren und die Gesamtbelohnung erhöhen kann.
Diese Dissertation schlägt somit mehrere Algorithmen zum Wissenstransfer für die Optimierung der Exploration in verschiedenen Settings des Mehrarmigen Banditen vor. Wir liefern sowohl theoretische Garantien als auch empirische Evaluationen, die signifikante Verbesserungen gegenüber bestehenden Methoden aufzeigen.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor, Prof. Dr.-Ing. Setareh Maghsudi, who gave me the unique opportunity to pursue my doctoral degree in the first place.
I am sincerely grateful to Prof. Dr. Georg Martius, who agreed to serve as my second supervisor and to review my dissertation, thereby accepting me as a Ph.D. candidate.
I also thank Prof. Dr. Claire Vernade, who allowed me to attend her weekly reading group meetings, and invited me to a retreat with her team. The topics during the reading group were always interesting and a welcome addition to my schedule. I thoroughly enjoyed the retreat and its activities, from working on problems to playing Mario Kart in the evening. I got to know the members of her group and greatly appreciated the group dynamics. Of course, I am also very grateful that Claire agreed to be part of the examination committee for my defense.
I thank Prof. Dr. Philipp Hennig for kindly taking the time to serve as the fourth examiner of my defense.
I greatly enjoyed the time spent with my former colleagues from Tübingen and Bochum (listed in no particular order): Michela, Ioannis, Amir, Mariam, Behzad, Xiaotong, Qiang, Melodi, Cheng, Francesco and Sofiane, as well as with the members of Claire's group and my fellow IMPRS-IS scholars: Michela (again), Marko, Onno, Ziyad, Dilsad, Cağatay, Amin and Nico. In hindsight, I have come to understand how important it is to have colleagues with whom one can laugh and share the same hardships, joys, and frustrations.
I am especially grateful to Sofiane for his guidance and for many insightful discussions on scientific and entertainment-media-related topics. In my eyes, he is a brilliant researcher, and his support throughout this journey was invaluable.
Last but certainly not least, I would like to show my appreciation to my mother, father, and brother for their continuous support and confidence in me, my closest friends, Max, Richard and Samuel, whom I have known since my time in school and dominated in competitive SSB, and finally, to my partner, Lucia, who was the closest person to me throughout this journey. We moved together to an unknown city, and I was able to share all my sorrows and successes with her. Her love, patience, and support carried me through many difficult moments, and I could not have wished for a better person by my side.

Contents

Chapter 1

Introduction

Reinforcement Learning (RL)-based models have proven exceptional efficiency in tasks such as board games, control problems (Duan *et al.*, 2016), robotics Kober *et al.* (2013); Levine *et al.* (2016) or autonomous driving (Kiran *et al.*, 2022). While AlphaGo (Silver *et al.*, 2016, 2017) and AlphaZero (Silver *et al.*, 2018) already outperform human top-level players in classic board games like Go and Chess, AlphaStar showed that proficient reinforcement learning agents extend beyond board games with discrete action spaces to complex real-time strategy game environments such as StarCraft II Vinyals *et al.* (2019). Outside of gaming environments, RL is used within the training process of chat bots such as Chat-GPT (Brown *et al.*, 2020; Ouyang *et al.*, 2022) and Deep-Seek (DeepSeek-AI *et al.*, 2024; DeepSeek-AI *et al.*, 2024) or in more subtle ways through recommender systems used for news articles, shopping websites, streaming services or social networks (Li *et al.*, 2010; Silva *et al.*, 2022; Zhu and Van Roy, 2023). RL is usually classified as an online learning problem (Sutton and Barto, 2018), that is, the training data is not readily accessible, but progressively gathered through interactions with the environment. The accuracy of an RL model hinges on the accumulated data points and how well they represent the essential state-action pairs required to find an optimal policy. Employing a greedy strategy may result in data sets that capture irrelevant parts of the state-action space, potentially compromising performance. As a result, agents cannot depend solely on their current model to select an action but need to incorporate exploration at the potential cost of receiving low immediate rewards. This is known as the exploration-exploitation trade-off (March, 1991). In some scenarios, data may be scarce, prompting the need to explore additional sources of external information and incorporate them into novel algorithms to further enhance the agent's performance. This aspect will be addressed in this dissertation.

The multi-armed bandit (MAB) problem, formally introduced in Robbins (1952), is an idealized version of the RL problem, in which the environment is fixed to a single state with no transitions. In each round, the agent selects one of many actions, called arms, receives noisy feedback, and updates its policy. Each arm is associated with an unknown distribution, from which the immediate reward is sampled when the arm is pulled. The goal, as in regular RL, is to maximize the expected reward

over several rounds or equivalently minimize the expected regret. In the case of MABs, this translates into selecting the action yielding the highest mean reward. The lack of state transitions makes the MAB setting more suitable to prove theoretical guarantees and pushes the focus onto the exploration-exploitation trade-off. Several works studied the MAB problem, for the regular stochastic setting, we have Thompson (1933) that introduced the Thompson sampling algorithm even before the MAB problem was formally introduced and Auer *et al.* (2002a) that introduced the upper-confidence bound algorithm based on the optimism principle. The works of Li *et al.* (2010), Chu *et al.* (2011) and Abbasi-Yadkori *et al.* (2011) proposed algorithms for the contextual setting with linear reward functions and Besson and Kaufmann (2019) addressed the non-stationary setting.

The exploration-exploitation trade-off depends on the amount of feasible information available. Leveraging external knowledge allows us to reduce exploration while avoiding potential uncertainty in our models. The concept of utilizing external information during training is generally known as Transfer Learning (TL) (Pan and Yang, 2009). In TL we denote the input data and the marginal distribution over it as domain and the label space together with the true label-predicting function as target. TL methods are often used to supplement traditional machine learning models when labeled data is scarce due to an expensive or unrealistic data generation process. By transferring knowledge from a data-rich domain, called the source, to a sparse one, called the target, TL methods help mitigating data sparsity and improve overall performance.

TL approaches are generally effective when certain conditions between source and target are met. In a worst-case scenario, when the domains or tasks are too disparate, transferring knowledge may degrade the performance of the target model, a phenomenon known as negative transfer. TL strategies therefore require source and target to be similar or share a common structure to be exploited.

Although TL methods have traditionally been applied in offline machine learning settings, there has been significant progress in formulating diverse strategies for online reinforcement learning and bandit problems. Multi-task and meta learning frameworks, in particular, utilize transfer learning across multiple different tasks, to enhance model performance by leveraging knowledge of a common meta structure. Many meta learning frameworks in bandits assume an underlying meta distribution from which a potentially unlimited number of bandit tasks are sampled (Cella *et al.*, 2020; Kveton *et al.*, 2021; Basu *et al.*, 2021; Peleg *et al.*, 2022). Algorithms aim to gain information on the meta distribution by sequentially running bandit algorithms and exploit that knowledge in subsequent tasks. Multi-task learning, on the other hand, deals with a limited set of tasks at once. There is an underlying meta structure that allows the transfer of knowledge between the tasks to accelerate learning, but in some cases this information is readily available to the learner in the form of an undirected graph (Cesa-Bianchi *et al.*, 2013; Yang and Toni, 2018), where each node is associated with a different task. Other common assumptions

for multi-task settings are cluster structures (Gentile *et al.*, 2014; Li *et al.*, 2019), where subsets of tasks yield the same features or low-rank structures, where all tasks are assumed to be located in a low-dimensional subspace (Cella *et al.*, 2023). In piecewise stationary environments, we assume that the reward distributions stay constant for periods of time and may change abruptly (Besson and Kaufmann, 2019). These stationary segments can be interpreted as tasks, implying that each environmental change triggers a transition in the bandit problem's task. Most algorithms deal with non-stationarity by discounting old data (Garivier and Moulines, 2011; Russac *et al.*, 2019) or actively detecting changes in the environment and restarting the training process (Auer *et al.*, 2019; Besson and Kaufmann, 2019; Besson *et al.*, 2022). Collected training data are simply discarded without evaluating them for new segments, thus motivating approaches that transfer knowledge from old to new segments.
This dissertation focuses on the advancement of transfer learning across diverse bandit frameworks such as piecewise stationary environments, meta and multi-task learning or regular single task settings by developing novel algorithms and offering theoretical proofs as well as empirical findings. In particular, we exploit the close proximity of tasks in euclidean space or arbitrary subspaces to reduce uncertainty in the models and therefore, improve exploration.

Outline In Chapter 2, we cover the basics of reinforcement learning, Markov decision processes and multi-armed bandits. In addition, we introduce the field of transfer learning and provide a detailed overview of state-of-the-art models pertinent to transfer learning applications in bandit problems, which are relevant to this dissertation.
Chapter 3 presents the work of Bilaj *et al.* (2023), which considers a transfer scenario for linear bandits where the learner has access to one or more expert feature vectors, also called sources, and uses them to enhance the learning of a target task. We assume that the source stems from a task with a reward function similar to that of the target. By applying a convex weighting strategy in an upper-confidence bound algorithm, we can reduce exploration and improve the accuracy of the current model if the source feature vector is close to the true feature vector of the target. The weights are calculated dynamically through the estimation of confidence set bounds according to Abbasi-Yadkori *et al.* (2011) of the source and target models on the already explored data set. When the confidence set bound of the target model becomes tighter than the bound of the source model, the source-related weight becomes zero, and it gets discarded. Theoretically, we show that with high probability there is a point in time until the source model is viable such that the learner can profit from it. A source continues to be viable for a longer duration when the Euclidean distance between its feature vector and that of the target is smaller. During this initial period, we show an improvement in the regret

bound over the baseline algorithm. The proposed model addresses the cold start problem and improves the regret by an offset.
Chapter 4 covers the work of Bilaj *et al.* (2024) and considers a meta learning setting, that is, the learner interacts with multiple tasks sequentially and the tasks share some underlying structure. In particular, we assume that all task feature vectors are located around an affine subspace, which is a generalization to the low-rank assumption in the multi-task setting and a generalization to the assumption we made in Chapter 3, instead of assuming a low distance between the feature vectors of source and target, we only assume close proximity within a low-dimensional subspace. By interacting with tasks and estimating feature vectors, we calculate projection matrices of the underlying subspace. We propose projection-based UCB and Thompson sampling algorithms that utilize the projection matrix as prior information and reduce uncertainty of the estimators in orthogonal direction to the subspace. As such, we improve the expected transfer regret bound with respect to the dimensionality of the problem.
Within Chapter 5, covered in Dhouib *et al.* (2025), we explore the multi-task bandit problem in environments with a clustered structure. In line with the methodologies of Cesa-Bianchi *et al.* (2013) and Yang and Toni (2018), we employ a graph that represents the relationships between tasks, with each task corresponding to a unique node. This graph is presumed to induce the concealed cluster configuration, which we leverage through network lasso regularization (Hallac *et al.*, 2015). Unlike other approaches, our algorithm does not actively discern the cluster structure. Instead, it implicitly enforces this structure in the estimation of the task feature vectors by solving the network lasso problem. We present a novel oracle inequality, attain a sublinear regret bound, and demonstrate how our findings are enhanced when applying graphs characterized by numerous edges between nodes within the same cluster and sparse edges between nodes of different clusters.
In this dissertation's final study Nourani-Koliji *et al.* (2023), outlined in Chapter 6, we explore a scenario where a combinatorial semi-bandits algorithm (Nourani-Koliji *et al.*, 2022) operates under a piecewise-stationary environment. Specifically, the distributions associated with the arms can shift abruptly at times unknown to the agent. Under the assumption that environmental changes do not apply to all arm distributions at once but only to those that are linked to each other, we adopt a group restart strategy using a GLR-change-point detector (Besson and Kaufmann, 2019). Upon detecting a change, the learner maintains the estimated mean rewards for all arms except those tied to the arm where the change was observed. Unlike other non-stationary algorithms, our method utilizes information of arms from previous segments or tasks and reduces uncertainty in these arms. This lessens the necessary exploration needed for each task and, as a result, lowers the regret bound.
In Chapter 7, we conclude with a summary of our findings and explain their significance to the core research themes of this dissertation. We also propose possible directions for future research.

Chapter 2

Background

2.1 Foundations of Reinforcement Learning Theory

Reinforcement Learning (RL) is a subclass machine learning problems focusing on agents learning to interact with an environment over a series of discrete time steps with the purpose of maximizing a reward signal (Sutton and Barto, 2018). Initially, the agent lacks any prior knowledge, but gathers essential information through interaction with the environment. In each time step, the agent selects an action and receives a corresponding reward as feedback. In the most general case, selected actions determine not only the immediate reward, but also the state of the environment as well as the set of available actions and consequently all possible future rewards. The accumulated data set of actions, rewards and state transitions is used to determine the agent's action-selection strategy, which we refer to as policy. Supervised learning relies on labeled data sets from an external source of knowledge and aims to find a mapping from the input data to the labels, whereas in unsupervised learning the goal is to learn hidden structures in unlabeled data sets. RL is different from both, as it involves learning through dynamic interaction instead of learning from a fixed labeled data set.

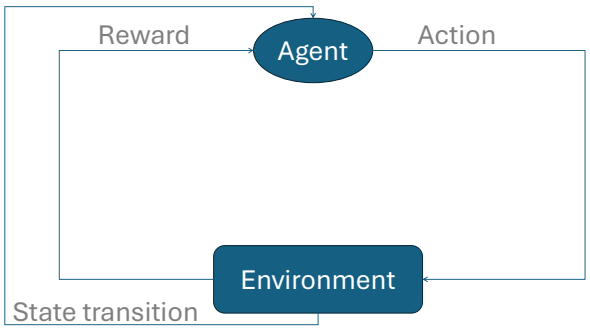


Figure 2.1: General reinforcement learning scheme.

2.1.1 Understanding Markov Decision Processes

Markov Decision Processes (MDP), introduced in Bellman (1957), provide a mathematical framework for RL. A MDP is a tuple $(\mathcal{S}, \mathcal{A}(s), R(s,a,s'), \gamma, \mathcal{T}(s'|s,a))$, with $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ where:

- $\mathcal{S}$ is the state space
- $\mathcal{A}(s)$ is the action space that depends on the current state
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function
- γ is a discount factor
- $\mathcal{T}(s'|s,a)$ yields the probability function of the environment transitioning to state s' after selecting an action a in state s

In each round k, the agent receives information on the state of the environment $s_k \in \mathcal{S}$. Upon selecting an action $a_k \in \mathcal{A}(s_k)$, the environment shifts from state s_k to a new state $s_{k+1} \sim \mathcal{T}(\cdot|s_k, a_k)$, and the agent receives a reward $r_k = R(s_k, a_k, s_{k+1})$. A key trait of MDPs is the Markov property: it implies that the probability of transitioning to the next state only depends on the current state and action taken. The agent selects actions following a policy π. The policy may be deterministically defined as follows:

$$\pi : \mathcal{S} \to \mathcal{A},$$

such that the next action to be selected is determined exactly without randomness $a_k = \pi(s_k)$. Alternatively, we can define stochastic policies:

$$\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$$

where $\pi(a|s) = \mathbb{P}(a_k = a | s_k = s)$ is the probability that the agent selects an action a when the environment is in state s. Sequences of state transitions, actions and immediate rewards are used to update the policy. The objective of the optimization problem is to maximize the expected cumulative discounted sum of rewards also referred to as return G:

$$G = \mathbb{E}_\pi \left[\sum_{k=1}^{\infty} \gamma^{k-1} R(s_k, a_k, s_{k+1}) \right].$$

We define the state-value function $V^\pi(s)$ as the expected return when starting from an initial state s and following a policy π. It can be further derived that the optimal value function, defined as $V^\star(s) = \max_\pi V^\pi(s)$, satisfies the Bellman optimality equation (Bellman (1957)):

$$V^\star(s) = \max_{a \in \mathcal{A}(s)} \sum_{s'} \mathcal{T}(s'|s,a) \left[R(s,a,s') + \gamma V^\star(s') \right].$$

The Bellman optimality equation involves an infinite recursive summation and serves as a cornerstone for modern reinforcement learning algorithms. Given the impracticality of solving it directly due to its need for infinite data, RL models are typically constructed by optimizing approximations of the Bellman equation. The required data for these optimization problems are usually gathered in an online manner, that is, the data sets are only available after the agent selects an action and observes the respective feedback and state transition.
The accuracy of such models depends on the available data set and on its coverage of the state-space. As a result, agents need to balance reward maximization based on the already accumulated data set and exploration to increase information on the environment. This is known as the exploration-exploitation trade-off.

2.1.2 Single-State RL: The Multi-Armed Bandit Problem

The multi-armed bandit (MAB) problem introduced in Robbins (1952) is the special case of the reinforcement learning setting, that does not use any state transitions. Its general setup includes an agent that selects an action also called arm a from a fixed action set $\mathcal{A}$ with $K = |\mathcal{A}|$ available actions at every round k. At the end of each round k, a reward r is sampled from an unknown distribution with some mean μ_{a_k} associated with the selected action a_k. A common assumption in many MAB applications is the sub-Gaussian condition on the random rewards, that is defined as:

Definition 2.1.1 (Definition 5.2 in Lattimore and Szepesvári (2020))**.** *A random variable r is σ-sub-Gaussian if for all $\iota \in \mathbb{R}$, we have:*

$$\mathbb{E}\left[\exp(\iota r)\right] \leq \exp\left(\frac{\iota^2 \sigma^2}{2}\right)$$

The tails of Sub-Gaussian distributions decay at least as fast as a Gaussian. This implies that any bounded random variable satisfies the sub-Gaussian condition. In each round, the agent's goal is to select the action yielding the highest mean reward and thus maximize the expected cumulative reward over time. Equivalently, the performance of a given algorithm after n rounds is measured by the expected cumulative regret $\mathcal{R}(n)$ defined as:

$$\mathcal{R}(n) = \mathbb{E}\left[\sum_{k=1}^{n} \mu_{a^\star} - \mu_{a_k}\right], \tag{2.1}$$

with $a^\star$ as the arm that yields the highest mean reward in every round k. We differentiate between instance-dependent and instance-independent regret: the

instance-dependent regret does depend on the gaps between the mean reward of the optimal arm and the other mean rewards, while the instance-independent regret holds for every problem instance. The basic MAB problem features a stationary environment, that is, the reward distributions and the action set do not change over time, and consequently the arm with the highest mean reward does not change either. The nonstationary case with changing reward distributions will be discussed in a later section. Upon receiving the reward after a round is finished, the agent updates its estimate of the arms' mean rewards and consecutively its policy. If we were to use a random arm selection policy, the resulting regret would grow linearly with respect to n, indicating an inefficient learning process. Therefore, when developing a bandit learning algorithm, we want to guarantee a sublinear bound on the regret. Multiple algorithms have been developed that solve the bandit problem for a constant set of arms and stationary rewards, some of which we will cover next.

Epsilon Greedy: Widely used in bandit tasks and regular RL, an epsilon greedy policy works according to the following rule: Define a probability ε with $0 < \varepsilon < 1$, then the agent selects a random action with probability ε, otherwise with probability $1 - \varepsilon$, it greedily selects an action with the highest estimated mean reward. This version of epsilon greedy yields a linear regret, thus motivating an alternative version called vanishing epsilon greedy, introduced in Auer *et al.* (2002a), where $\varepsilon(k) \propto \sqrt[3]{\log(k)/k}$ decreases with each round. In the multi-armed bandit setting, the vanishing epsilon greedy algorithm achieves a regret bound of $\mathcal{O}(\sqrt[3]{n^2 K})$.

Upper Confidence Bound (UCB): Assume that the action set $\mathcal{A}$ for the agent is fixed at every round and assume a 1-sub-Gaussian reward distribution associated with every arm. For that setting, the UCB algorithm, introduced in Auer *et al.* (2002a), finds a balance between exploitation and exploration depending on the number of times an arm has been pulled. We define the UCB function at each round k as:

$$\mathrm{UCB}(a) = \hat{\mu}_a + \sqrt{\frac{2\log(k)}{N_a}},$$

with $\hat{\mu}_a$ as the estimated mean reward of arm a and N_a as the number of times arm a has been pulled. The agent then selects an action that maximizes the UCB value: $a = \arg\max_{a' \in \mathcal{A}} \mathrm{UCB}\,(a')$. The latter term in the UCB function reflects the exploration contribution for each arm and decreases the more often an action has been selected. After each round, the mean reward estimate is updated. The UCB-algorithm follows the principle of optimism in the face of uncertainty (OFU), which intuitively can be explained as the agent selecting an action with the 'highest

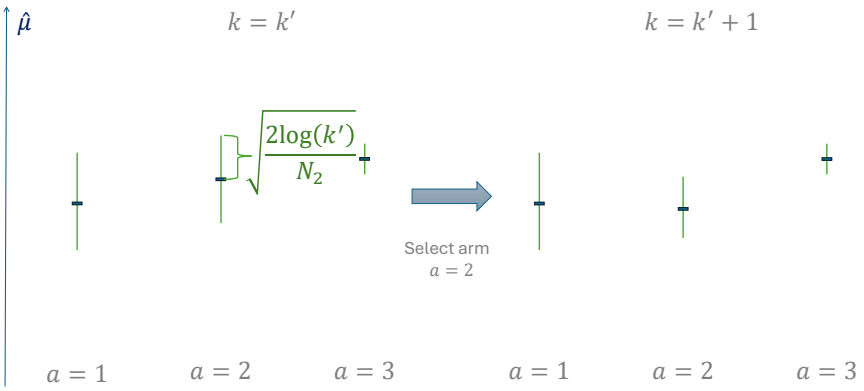


Figure 2.2: Illustration of the UCB algorithm. Mean reward estimations for $K = 3$ arms are shown at two consecutive time steps. At $k = k'$ the third arm yields the highest estimation but arm $a = 2$ was selected due to the higher confidence bound.

potential'. When an action a is selected many times, the agent becomes confident in its reward estimation, and the uncertainty decreases. A different action a' with a lower reward estimate, that has been selected fewer times, yields a greater uncertainty in the estimation, thus a higher 'potential'. If the uncertainty regarding a' is high enough such that it maximizes the UCB function, the agent will select a' over a. The UCB-algorithm is well known and achieves an upper regret bound of $\tilde{\mathcal{O}}\left(\sqrt{Kn}\right)$[1] for any environment (Lattimore and Szepesvári, 2020). An illustration of the UCB arm selection strategy is provided in Figure 2.2.

Thompson Sampling: One of the first algorithms to ever solve the MAB problem is known as Thompson Sampling (TS) and dates back to Thompson (1933). Following a Bayesian approach, the agent is given a prior distribution over the task parameters, in this case a prior belief on the arms' reward distributions $\mathcal{P}_a(\mu)$ for the initial round. According to Bayes' theorem, a posterior on the estimated mean rewards for every arm can be calculated given the priors and the received rewards:

$$\mathcal{P}_a(\mu|\mathcal{D}) \propto \mathcal{P}_a(\mu)\mathcal{P}_a(\mathcal{D}|\mu),$$

where $\mathcal{D}$ contains the selected actions and the respective observed rewards. In each round k, the algorithm samples a bandit environment from the posterior, based on the data $\mathcal{D}_k$ collected up to round k, that is, for each arm $a \in \mathcal{A}$ a value $\tilde{\mu}_{a,k} \sim \mathcal{P}_a(\cdot|\mathcal{D}_k)$ is sampled and the agent selects the arm with the highest sample value:

$$a_k = \arg\max_{a' \in \mathcal{A}} \tilde{\mu}_{a',k}.$$

[1]The $\tilde{\mathcal{O}}(\cdot)$ notation is equivalent to the $\mathcal{O}(\cdot)$ notation but hides polylogarithmic factors.

Exploration occurs as a result of the randomization of the sampling process. Actions that were explored less have a higher variance in the corresponding posterior, leading to a higher chance of sampling the largest value. Thompson sampling is usually applied to Bernoulli rewards, while a beta distribution is used as parametric assumption to model the prior as well as the posterior. Similarly to UCB, TS achieves a regret bound of $\tilde{\mathcal{O}}\left(\sqrt{nK}\right)$.

2.1.3 Select Multiple Actions: Combinatorial Bandits

Combinatorial multi-armed bandits (CMAB) are a generalization to the classic MAB problem where an agent does not select one but up to s actions from an action set $\mathcal{A}$, with $K = |\mathcal{A}|$ within the same round. We will refer to the actions as base arms. As in the classic MAB setting, each base arm is associated with a random feedback variable, where each random variable is sampled i.i.d. from an unknown distribution. In a combinatorial setting, there are up to 2^K combinations of base arms an agent can select. Every subset or combination of base arms is referred to as super arm $\boldsymbol{S} \in \{0,1\}^K$, where each entry is associated with a different base-arm with the entry equal to 1 if a base-arm was selected and 0 otherwise. Upon selecting a super arm $\boldsymbol{S}_k$ in round k, the feedback of all selected base arms becomes available to the agent as well as a reward $r(\boldsymbol{S}_k)$. The objective of algorithms solving the CMAB problem is the minimization of the expected regret:

$$\mathcal{R}(n) = \mathbb{E}\left[\sum_{k=1}^{n} r(\boldsymbol{S}^\star) - r(\boldsymbol{S}_k)\right],$$

with $\boldsymbol{S}^\star$ as the super arm that maximizes the expected reward.

Combinatorial UCB Chen *et al.* (2013) proposes the combinatorial upper confidence bound (CUCB) algorithm to solve the CMAB problem. In each round k, the CUCB algorithm calculates as UCB value for every base arm a:

$$\mathrm{UCB}(a) = \hat{\mu}_a + \sqrt{\frac{3\log(k)}{2N_a}},$$

with $\hat{\mu}_a$ as the mean estimation of the feedback distribution of base arm a and N_a as the number of times base-arm a was selected in a super arm. Chen *et al.* (2013) make the assumption of an existing α, β-approximation oracle that takes the expectation values $\boldsymbol{\mu}$ as input and outputs the optimal super arm $\boldsymbol{S}^\star$ with high probability. The oracle is defined as follows:

Definition 2.1.2 (α, β-approximation oracle)**.** *Let $0 \leq \alpha, \beta \leq 1$. Given an expectation vector $\boldsymbol{\mu}$ as input, the α, β-approximation oracle outputs a super arm $\boldsymbol{S}$, such that $\Pr[r(\boldsymbol{S}) \geq \alpha r(\boldsymbol{S}^\star)] \geq \beta$, where β is the success probability of the oracle.*

Since the true expectation values of the base-arm associated distributions is unavailable to the agent, the calculated UCB-values are given as input to the oracle, which then outputs the super arm that the agent will select in that round. The CUCB algorithm yields an instance-dependent regret bound of $\mathcal{O}(K\log(n))$ and an instance-independent bound of $\mathcal{O}\left(\sqrt{Kn}\right)$.

Structural Equation Model UCB (SEM-UCB) Instead of assuming access to an oracle that outputs the optimal super arm with high probability given the expectation values of the base-arm distributions, the approach of Nourani-Koliji *et al.* (2022) assumes a structural equation model behind the true reward function. That is, they consider a directed acyclic graph (DAG) to model causal relations between the base arms, and consequently the reward function can be expressed as:

$$r(\boldsymbol{S}) = \mathbf{1}^\top (\boldsymbol{I}_K - \boldsymbol{W})^{-1} \mathrm{diag}\,(\boldsymbol{z})\, \boldsymbol{S}, \tag{2.2}$$

with $\boldsymbol{I}_K$ as the K-dimensional identity matrix, $\boldsymbol{W} \in \mathbb{R}^{K\times K}$ as adjacency matrix of the DAG and $\boldsymbol{z} \in \mathbb{R}^K$ as feedback vector that contains immediate sampled feedback of the selected base arms in the corresponding entries and is 0 in all other entries. Note that the output of $r(\cdot)$ yields randomness due to the base-arm feedback z_k. The SEM-UCB algorithm includes an initial exploration phase to learn the adjacency matrix $\boldsymbol{W}$ of the DAG behind the causal relations. Super arms are selected by calculating UCB values for each base arm and inserting them into the reward model in eq. (2.2) in place of the base arm feedback vector. Maximizing the resulting rewards with respect to $\boldsymbol{S}$ yields the super arm to select. Similar to the CUCB algorithm, SEM-UCB achieves an instance-dependent regret bound of $\mathcal{O}(K\log(n))$.

2.1.4 Adding Context: Linear Bandits

Linear Bandits extend the classic MAB framework by associating each arm a with a context vector $\boldsymbol{x}_a \in \mathbb{R}^d$ that is available to the learner. In contrast to the classic stationary MAB setting, the set of arms $\mathcal{A}_k$ and respective context vectors can change arbitrarily. The immediate reward r_k for pulling an arm a_k in round k is assumed to follow a linear relation:

$$r_k = \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^\star + \epsilon,$$

with $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ as unknown task-related feature vector and $\epsilon \in \mathbb{R}$ as sub-Gaussian noise term. The rewards are stored in a vector $\boldsymbol{y}_{k+1} = [r_i]_{i\in\{1,\dots,k\}} \in \mathbb{R}^k$ while the collected context vectors are stored in a data matrix $\boldsymbol{D}_{k+1} = [\boldsymbol{x}_{a_i}^\top]_{i\in\{1,\dots,k\}} \in \mathbb{R}^{k\times d}$. We further define $\boldsymbol{A}_{k,\lambda} = \lambda \boldsymbol{I}_d + \boldsymbol{D}_k^\top \boldsymbol{D}_k$, with $\boldsymbol{I}_d$ as d-dimensional identity matrix and $\lambda \in \mathbb{R}$. The goal is to find a suitable estimate for the true feature vector based

on the collected rewards and the context vectors and use it to minimize the expected regret bound. Given the reward function, we provide an equivalent definition of the expected regret defined in 2.1:

$$\mathcal{R}(n) = \mathbb{E}\left[\sum_{k=1}^{n} \max_{a' \in \mathcal{A}_k} \left\{\boldsymbol{x}_{a'}^\top \boldsymbol{\theta}^\star\right\} - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^\star\right],$$

Due to the inclusion of context vectors and the linearization of the reward function, the number of arms available becomes irrelevant to the problem if $|\mathcal{A}| > d$ for some algorithms. This means that adding context vectors to arms allows one to solve the bandit problem with a large set of actions. Instead, we have the dimensionality d that influences the complexity and the regret bound.
We are able to retrieve an equivalent problem to the context-free case by designing a fixed set of arms of size $|\mathcal{A}| = d$ with linearly uncorrelated context vectors, that is,

$$\boldsymbol{x}_a^\top \boldsymbol{x}_{a'} = 0 \quad \Leftrightarrow \quad a \neq a'.$$

If the context vectors assigned to the actions are all uncorrelated, there is no information shared between the arms, and the mean reward estimation of arm a is independent of any other prior selected arm $a' \neq a$.
There are various approaches to solving the linear bandit problem, some of which we present here.

Exploration-Based Algorithms

Optimism in the Face of Uncertainty Linear Bandit Algorithm: In Abbasi-Yadkori *et al.* (2011) an algorithm called "optimism in the face of uncertainty linear bandit algorithm" (OFUL) is proposed for linear bandits with sub-Gaussian reward distributions and adversarial context vectors. The algorithm finds an estimate $\hat{\boldsymbol{\theta}}_k \in \mathbb{R}^d$ of the true feature vector by solving the ridge regression optimization problem at each round k. The solution can be determined analytically as follows:

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \left\{\|\boldsymbol{y}_k - \boldsymbol{D}_k \boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2\right\} = \boldsymbol{A}_{k,\lambda}^{-1} \boldsymbol{D}_k^\top \boldsymbol{y}_k, \tag{2.3}$$

with some regularization parameter $\lambda \in \mathbb{R}$. A major contribution of Abbasi-Yadkori *et al.* (2011) is the confidence set $\mathcal{C}_k$ provided for the estimator $\hat{\boldsymbol{\theta}}_k$:

Theorem 2.1.3 (Theorem 2 of Abbasi-Yadkori *et al.* (2011))**.** *Let $\{F_k\}_{k=1}^\infty$ be a filtration and let $\{\epsilon_k\}_{k=1}^\infty$ be a real valued stochastic process such that ϵ_k is F_k-measurable and ϵ_k is conditionally σ-sub-Gaussian for some $\sigma \geq 0$. Let $\hat{\boldsymbol{\theta}}_k$ be determined as in eq.* (2.3) *and assume $\|\boldsymbol{\theta}^\star\|_2 \leq V$, $\|\boldsymbol{x}_a\|_2 \leq 1$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $k > 0$, $\boldsymbol{\theta}^\star$ lies in the set*

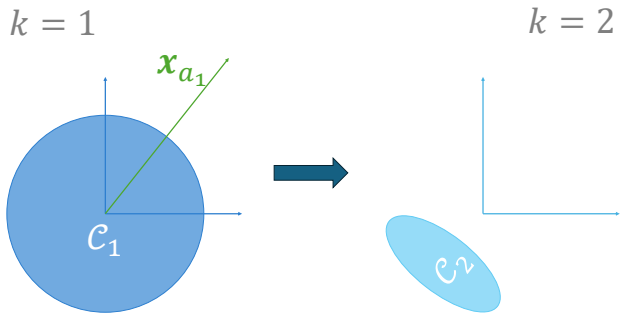


Figure 2.3: The confidence ellipsoid in $d = 2$ dimensions. Left hand side marks the initial ellipsoid at time $k = 1$. Right hand side shows the ellipsoid at $k = 2$ after selecting the first action with context vector $\boldsymbol{x}_{a_1}$. The ellipsoid becomes smaller in direction of the context vector, indicating a reduction in uncertainty.

$$\mathcal{C}_k = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \left\| \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta} \right\|_{\boldsymbol{A}_{k,\lambda}} \leq \sigma \sqrt{d \log \left(\frac{1 + (k-1)/\lambda}{\delta} \right)} + \sqrt{\lambda} V \right\}.$$

Given the confidence set, the agent selects an action according to:

$$\left(a, \tilde{\boldsymbol{\theta}} \right) = \underset{(a', \boldsymbol{\theta}) \in \mathcal{A}_k \times \mathcal{C}_k}{\arg\max} \; \boldsymbol{x}_{a'}^\top \boldsymbol{\theta}.$$

The confidence set takes the shape of a multi-dimensional ellipsoid and reflects the level of certainty regarding the current model $\hat{\boldsymbol{\theta}}$ based on the arms explored so far. This ellipsoid indicates the greatest confidence or minimal uncertainty in the model along the smallest axis and the least confidence or highest uncertainty along the largest axis. Thus, when selecting an arm, context vectors that align with the largest axis of the confidence ellipsoid are favored. The OFUL algorithm achieves a regret bound of $\mathcal{O}\left(d\sqrt{n \log(n)} \right)$

Linear UCB (LinUCB) In Li *et al.* (2010) a similar model was proposed, which is also based on the principle optimism in face of uncertainty, and constructs an upper confidence bound algorithm called LinUCB. The UCB function at each round k is defined as:

$$\text{UCB}(a) = \boldsymbol{x}_a^\top \hat{\boldsymbol{\theta}}_k + \gamma_k \| \boldsymbol{x}_a \|_{\boldsymbol{A}_{k,\lambda}^{-1}}, \tag{2.4}$$

with γ_k as exploration scaling term and $\hat{\boldsymbol{\theta}}_k$ determined as in eq. (2.3). The agent

selects an arm a according to $a = \arg\max_{a' \in \mathcal{A}_k} \text{UCB}(a')$. The first term in eq. (2.4) yields the reward estimate, while the second term is an exploration-scaling quantity. Given the mean reward $\bar{r}_a = \boldsymbol{x}_a^\top \boldsymbol{\theta}^\star$ and the reward estimate $\hat{r}_{a,k} = \boldsymbol{x}_a^\top \hat{\boldsymbol{\theta}}_k$ for any arm a, we can bound the following with high probability using the Cauchy-Schwarz inequality:

$$|\bar{r}_a - \hat{r}_{a,k}| = \left| \boldsymbol{x}_a^\top \left(\boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}_k \right) \right| \leq \left\| \boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}_k \right\|_{\boldsymbol{A}_{k,\lambda}} \|\boldsymbol{x}_a\|_{\boldsymbol{A}_{k,\lambda}^{-1}} \leq \gamma_k \|\boldsymbol{x}_a\|_{\boldsymbol{A}_{k,\lambda}^{-1}}.$$

The exploration scaling term is set to $\gamma_k = 1 + \frac{\log 2/\delta}{2}$ in Li *et al.* (2010) but could also be determined according to the confidence set bound in Theorem 2.1.3, in which case we have $\gamma_k \geq \left\| \boldsymbol{\theta}^\star - \hat{\boldsymbol{\theta}}_k \right\|_{\boldsymbol{A}_{k,\lambda}}$. This choice of exploration scaling would make LinUCB equivalent to the OFUL algorithm and γ_k would reflect the uncertainty of the reward estimation at round k. The LinUCB algorithm proposed in Li *et al.* (2010) achieves, under the assumption of a fixed set of arms $\mathcal{A}_k$ with $K = |\mathcal{A}|$, a regret bound of $\tilde{\mathcal{O}}(\sqrt{dKn})$. Note that this bound still depends on the number of arms, contrary to the result of Abbasi-Yadkori *et al.* (2011).

Linear TS: Typically, for the Bayesian Bandit setting, the agent assumes a prior $\mathcal{P}(\boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{0}, \frac{v^2}{\lambda} \boldsymbol{I}_d\right)$, with some hyperparameter v, over the task parameters before the first round. After selecting an action and receiving feedback in round k, the agent updates its posterior with the accumulated rewards and context vectors summarized in $\mathcal{D}_k$:

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}_k) = \mathcal{N}\left(\hat{\boldsymbol{\theta}}_k, v^2 \boldsymbol{A}_{k,\lambda}^{-1}\right),$$

where $\hat{\boldsymbol{\theta}}_k$ is estimated as in eq. (2.3). The agent selects an action by sampling a bandit environment or vector $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^d$ from the posterior and following a greedy policy:

$$a = \underset{a' \in \mathcal{A}_k}{\arg\max}\, \boldsymbol{x}_{a'}^\top \tilde{\boldsymbol{\theta}}.$$

As in the context-free case, there is no explicit exploration term here, it occurs implicitly during the sampling of $\tilde{\boldsymbol{\theta}}$, where $\boldsymbol{A}_{k,\lambda}^{-1}$ acts as covariance and scales with the confidence ellipsoid of the current estimation. Although Agrawal and Goyal (2013) provided a regret bound of $\tilde{\mathcal{O}}\left(d^2\sqrt{n}\right)$ for the linear TS algorithm, Abeille and Lazaric (2017) was able to prove a tighter bound of $\tilde{\mathcal{O}}\left(d^{\frac{3}{2}}\sqrt{n}\right)$. Note that Thompson sampling makes weak prior assumptions on the environment with $\mathcal{P}(\boldsymbol{\theta})$, that encourages the feature vector solutions to be close to the null vector. In Chapter 4, we show that you can improve your prior distribution through the transfer of knowledge between related settings.

Greedy Algorithms

There is a class of algorithms that do not take the approach of explicitly balancing exploration and exploitation during arm selection, but greedily select arms w.r.t. the current estimator $\hat{\boldsymbol{\theta}}$. Greedy algorithms usually require additional assumptions to guarantee sublinear regret bounds. One major assumption is that at each round, the context vectors for the set of available arm is sampled i.i.d. from a distribution with zero mean. In addition, restricted eigenvalue (RE) conditions or compatibility assumptions (Bühlmann and van de Geer (2011)) are posed on the true gram matrices of the context vector generating distributions. With the help of suitable RE conditions, it is possible to show that oracle inequalities hold with high probability for the estimators such that:

$$\lim_{k\to\infty} \left\| \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^\star \right\|_2 = 0.$$

With the convergence of the oracle inequality, sublinear regret bounds can be proven. For the single task scenario, two notable examples are Oh *et al.* (2021) and Bastani *et al.* (2021).

Sparsity Agnostic Lasso Bandit Oh *et al.* (2021) proposes a lasso-regularized estimator for high-dimensional settings to exploit potential sparse structures in the true feature vector:

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \mathcal{L}_k(\boldsymbol{\theta}) + \lambda_k \|\boldsymbol{\theta}\|_1 \right\},$$

with $\mathcal{L}_k(\cdot)$ as negative log-likelihood loss and λ_k as a time-dependent regularization parameter. The lasso regularization term enforces sparsity on the estimator, and thus the algorithm implicitly learns and exploits any sparsity structure in the true feature vector. The Sparsity-Agnostic Lasso Bandit Algorithm achieves, without prior knowledge of the sparsity, a regret bound of $\mathcal{O}\left(s_0\sqrt{n\log(dn)}\right)$, with s_0 as sparsity index i.e. the number of nonzero elements in $\boldsymbol{\theta}^\star$.

OLS Bandit The approach of Bastani *et al.* (2021) proposes a greedy algorithm that solves the regular ordinary least squares (OLS) regression problem at each round if the gram matrix $\boldsymbol{D}_k^\top \boldsymbol{D}_k$ is invertible.

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{y}_k - \boldsymbol{D}_k\boldsymbol{\theta}\|_2^2 = \left[\boldsymbol{D}_k^\top \boldsymbol{D}_k\right]^{-1} \boldsymbol{D}_k^\top \boldsymbol{y}_k.$$

When the number of arms available for each round is $K = 2$, the greedy algorithm proposed in Bastani *et al.* (2021) achieves an instance-dependent regret bound of $\mathcal{O}\left(d\log(d)^{\frac{3}{2}}\log(n)\right)$.

2.1.5 Adapting to Change: Non-Stationary Bandits

Up until now, we only discussed the stationary setting, i.e. the reward distributions for every arm $a \in \mathcal{A}$ stay fixed, which is an inappropriate assumption if the environment is changing. The nonstationary setting for contextless stochastic bandits can be formalized by assuming that there are at most $M-1$ shifts in the mean values of the feedback distributions or M stationary segments:

$$\sum_{k=1}^{n} \max_{a \in \mathcal{A}} \mathbb{1}\{\mu_a(k) \neq \mu_a(k+1)\} \leq M-1,$$

where we added a time dependency to the mean values of the reward distributions $\mu_a(k)$. With the reward distributions changing, we adapt the regret definition in eq. (2.1) for non-stationary bandits and introduce the expected dynamic regret:

$$\mathcal{R}(n) = \mathbb{E}\left[\sum_{k=1}^{n} \mu_{a_k^\star}(k) - \mu_{a_k}(k)\right],$$

with $a_k^\star$ as the arm yielding the highest mean reward and a_k as the arm selected in round k. If the points in time where the shifts occur were known, one could simply apply any suitable bandit algorithm to each stationary segment separately. Since the location of the change-points is usually unavailable for the agent, different strategies have been developed to deal with the non-stationary setting.

Passively Adapting Algorithms

Discounted UCB In Garivier and Moulines (2011) an algorithm called Discounted UCB is provided. Using an appropriate weight $w < 1$ and assuming that all rewards are bounded by a constant $B \in \mathbb{R}$, we can discount old rewards r_{a_i} for $i < k$ in the mean reward estimation:

$$\hat{\mu}_a(w) = \frac{1}{N_a(w)} \sum_{i=1}^{k} w^{k-i} r_{a_i} \mathbb{1}_{\{a_i = a\}} \quad N_a(w) = \sum_{i=1}^{k} w^{k-i} \mathbb{1}_{\{a_i = a\}}.$$

The UCB function is adapted accordingly, taking the discount into account:

$$\mathrm{UCB}(a) = \hat{\mu}_a(w) + 2B\sqrt{\xi \frac{\log \sum_{a' \in \mathcal{A}} N_{a'}(w)}{N_a(w)}},$$

With ξ as a hyperparameter. Old data is supposed to have less influence on the current estimation of the mean rewards, due to changes in the environment. The discounted UCB achieves a regret bound of $\mathcal{O}\left(\sqrt{Mn}\log(n)\right)$ through fine-tuning of hyper parameters with prior knowledge on M.

Sliding Window UCB In the same work of Garivier and Moulines (2011) another approach is proposed, called sliding window UCB, where only newly received data points are considered in each round k. Rewards r_{a_i} for $i < k - \tau$ are discarded. The maximum number of data points considered in each round is referred to as window size τ. By including only the most recent data for the agent, the probability of old or invalid data decreases at the cost of higher uncertainty in the reward estimations. The estimated means for every arm and the number of times an arm was chosen within the current window are therefore determined as:

$$\hat{\mu}_a(\tau) = \frac{1}{N_a(\tau)} \sum_{i=k-\tau+1}^{k} r_{a_i} \mathbb{1}_{\{a_i=a\}}, \quad N_a(\tau) = \sum_{i=k-\tau+1}^{k} \mathbb{1}_{\{a_i=a\}}.$$

The sliding window UCB achieves a regret bound of $\mathcal{O}\left(\sqrt{Mn\log(n)}\right)$ through fine-tuning.

Active Detection of Environmental Shifts

The previously mentioned approaches work both passively, as they do not react to changes in the environment. In contrast, active methods such as change-point detectors are designed to detect shifts in the mean of the arm's reward distributions with high probability. After detecting the shifts, the agent discards all data regarding the arms and starts learning from scratch. Being an active method, change-point detectors are computationally more expensive than the sliding window or discounted methods but provide a tighter regret bound.

Bernoulli GLR-Change-Point Detector In Besson and Kaufmann (2019) A generalized likelihood ratio (GLR) change-point detector was proposed as an active method to solve the piecewise-stationary MAB problem for sub-Bernoulli reward distributions. In Besson and Kaufmann (2019) sub-Bernoulli distributions are defined as follows:

Definition 2.1.4. *A distribution ν is called sub-Bernoulli if for any $\lambda \in \mathbb{R}$ it satisfies:*

$$\mathbb{E}_{r\sim\nu}\left[\exp(\lambda r)\right] \leq \log\left(1 - \mathbb{E}_{r\sim\nu}\left[r\right](1 - \exp(\lambda))\right)$$

Any random variable supported in an interval $[0, 1]$ satisfies the sub-Bernoulli condition. The GLR change-point detector is designed to detect shifts in the mean of the base-arm distributions. Given binary relative entropy kl function:

$$\mathrm{kl}(x, y) \coloneqq x \log\left(\frac{x}{y}\right) + (1 - x)\log\left(\frac{1 - x}{1 - y}\right),$$

we can define the GLR change-point detector:

Definition 2.1.5. *The Bernoulli GLR change-point detector with threshold function $\beta(k, \delta)$ is*

$$\hat{\tau}_\delta \coloneqq \inf \left\{ k \in \mathbb{N}^\star : \sup_{i \in [1,k]} \left[i \times \mathrm{kl}\left(\hat{\mu}_{1:i}, \hat{\mu}_{1:k}\right) + (k-i) \times \mathrm{kl}\left(\hat{\mu}_{i+1:k}, \hat{\mu}_{1:k}\right) \right] \geq \beta(k, \delta) \right\},$$

where $\hat{\mu}_{j:i}$ is the estimated mean of an arm distribution estimated from the rewards collected in rounds $\{j, ..., i\}$ in which the respective arm was selected.

The change-point detector is triggered when the set defined in Definition 2.1.5 for the stationary segment is not empty. Maillard (2019) proposes for sub-Gaussian distributions and quadratic divergence $\mathrm{kl}(x, y) = 2(x-y)^2$, a threshold function $\beta(k, \delta)$ such that the probability of a false alarm occurring is bounded by δ:

$$\beta(k, \delta) = \left(1 + \frac{1}{k}\right) \log\left(\frac{3k\sqrt{k}}{\delta}\right).$$

For Bernoulli distributions and the relative entropy kl function, Besson and Kaufmann (2019) recommend using $\beta(k, \delta) = \log\Big(3k\sqrt{k}/\delta\Big)$ in practice. When the number of change-points and the horizon n is known by the learner, the algorithm proposed in Besson and Kaufmann (2019) achieves through fine-tuning a regret bound of $\mathcal{O}\left(K\sqrt{Mn\log(n)}\right)$. In Chapter 6, we apply the GLR change-point detector in the combinatorial bandit setting similarly to the work in Zhou *et al.* (2020).

Adaptive Switching Algorithm In Auer *et al.* (2019) an algorithm is proposed that achieves optimal regret without requiring the learner to know the number of change-points in advance. Instead, it utilizes the total variation of the change. The ADSwitch algorithm achieves this by partitioning the arms into "good" and "bad" arms at the beginning, where a "good" arm is indistinguishable from the optimal arm. Good arms are selected regularly by the algorithm, thus detecting shifts in the reward distribution is simple. The main challenge lies in monitoring changes in the "bad" or suboptimal arms, as frequent selection of these arms could lead to significantly increased regret. They propose a consecutive sampling approach, in which a suboptimal arm is selected consecutively for a fixed number of times. The number of consecutive samples is fine-tuned to detect shifts with high probability and at the same time minimize the contribution to the cumulative regret. ADSwitch achieves a regret bound of $\mathcal{O}\left(\sqrt{KMn\log(n)}\right)$.

Non-stationary algorithms deal with dynamic environments by discounting or discarding old data points that belong with high probability to a previous stationary

segment. But in some cases these changes may follow some underlying structure and using data from prior segments could elevate the agent's performance instead of learning all segments independently. We follow that concept in Chapter 6.

2.2 Transfer Learning across Tasks

Transfer Learning (TL) methods are often used to supplement traditional machine learning models that suffer from an insufficient amount of labeled data. Data scarcity might arise because the data generation process is too expensive or simply unrealistic. TL addresses this problem by transferring knowledge from a data-rich setting, called source, to a sparse setting, which we refer to as target.
According to Pan and Yang (2009), a domain can be defined as: $\mathcal{D} = (\mathcal{X}, \mathcal{P}(X))$, with $\mathcal{P}(X)$ being the marginal distribution for $X \in \mathcal{X}$, whereas a task is defined as: $\mathcal{T} = (\mathcal{Y}, R(\cdot))$ with $R : \mathcal{X} \to \mathcal{Y}$ denoting the predictive function and $\mathcal{Y}$ the label space. We are now able to give a definition of TL:

Definition 2.2.1 (Transfer Learning, Definition 1 of Pan and Yang (2009))**.** *Given a source domain and learning task, transfer learning aims to improve the learning of the target predictive function $R_T(\cdot)$ in the target domain using the knowledge of the source domain $\mathcal{D}_S$ and the source task $\mathcal{T}_S$.*

Generally, TL methods are effective when certain conditions between source and target are met. In a worst-case scenario, when the domains are too disparate, transferring knowledge may actually degrade the performance of the target model, a phenomenon known as negative transfer. Therefore, TL strategies require the domains or tasks of the source and target to be similar or share a common structure.

Transductive Transfer Learning In transductive TL, the tasks of source and target are the same, but the domains may differ. Additionally, some unlabeled target domain data must be available during training. This setting can be divided into two subcategories: When the feature spaces of source and target are different with $\mathcal{X}_S \neq \mathcal{X}_T$ and when only the marginal distributions of input data differ with $\mathcal{P}_S(X) \neq \mathcal{P}_T(X)$. The latter setting is also known as domain adaptation (DA).

Inductive TL: When the target and source share the same domain but the tasks differ, the setting is known as inductive TL. Here, some labeled data of the target domain is required to learn the objective predictive function. Since the RL and MAB settings do not work with labels but with reward signals, we use the reward generating function $R(\cdot)$ when referring to the objective predictive function in a task.

Unsupervised TL When both domains and tasks differ between source and target, and the observable data in all domains are unlabeled, we speak of unsupervised TL. This setting focuses on problems such as clustering, dimensionality reduction and density estimations.

Within the context of MABs, transfer learning methods have been studied in meta and multi-task learning frameworks. These cases deal with multiple tasks, i.e. multiple feature vectors that share information between each other. Since feature vectors in a bandit setting determine the reward function, meta and multi-task learning fall into the category of inductive TL. We highlight some concepts and respective models in the next subsection, and throughout this dissertation, we propose different inductive TL approaches applied in different MAB settings.

2.2.1 Learning Shared Structures: Meta Learning in Bandits

Meta Learning in the bandit setting deals with a possible infinite number of tasks sequentially. The general assumption is that there exists an underlying structure that is shared between all tasks. Thus, algorithms focus on learning the underlying structure and leveraging this knowledge to solve future tasks. In Kassraie *et al.* (2022) a non-linear bandit setting is considered where the goal is to find a common kernel function, though we will focus on linear bandits with a shared meta distribution. This can be mathematically formalized by associating each task $t \in \{1, ..., M\}$ with a feature vector $\boldsymbol{\theta}^\star(t) \in \mathbb{R}^d$ and assuming that all feature vectors are sampled from a common Gaussian ρ:

$$\boldsymbol{\theta}^\star(t) \sim \rho$$

In this case, learning the meta structure amounts to finding an estimate of ρ. We now define the expected transfer regret:

$$\mathcal{R}(n) = \mathbb{E}_{\boldsymbol{\theta}^\star \sim \rho} \left[\mathbb{E} \left[\sum_{k=1}^{n} \max_{a' \in \mathcal{A}_k} \left\{ \boldsymbol{x}_{a'}^\top \boldsymbol{\theta}^\star \right\} - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^\star \right] \right],$$

to measure the performance of meta learning bandit algorithms.

Biased OFUL The work of Cella *et al.* (2020) assumes a Gaussian meta distribution with a known total variation σ_{total} and unknown mean $\boldsymbol{\mu}$. While the mean is estimated through the ridge estimators $\hat{\boldsymbol{\theta}}_{\text{Ridge}}(t)$ of M previously solved tasks as $\hat{\boldsymbol{\mu}} = \frac{1}{M} \sum_{t=1}^{M} \hat{\boldsymbol{\theta}}_{\text{Ridge}}(t)$, the total variation is assumed to be small compared to the euclidean bound of each task feature vector $\boldsymbol{\theta}^\star$:

$$\sigma_{\text{total}} \coloneqq \mathbb{E}_{\boldsymbol{\theta}^\star \sim \rho}\left[\|\boldsymbol{\mu} - \boldsymbol{\theta}^\star\|_2^2\right] \ll \|\boldsymbol{\theta}^\star\|_2^2.$$

New tasks are solved by incorporating the estimation $\hat{\mu}$ as a biased regularization term in the optimization problem of eq. (2.3) and subsequently in the confidence set $\mathcal{C}_k$ defined in Theorem 2.1.3:

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \left\{ \|\boldsymbol{y}_k - \boldsymbol{D}_k\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}\|_2^2 \right\},$$

$$\mathcal{C}_k = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \left\| \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta} \right\|_{\boldsymbol{A}_{k,\lambda}} \leq \sigma \sqrt{d \log\left(\frac{1 + (k-1)/\lambda}{\delta} \right)} + \sqrt{\lambda}\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\mu}}\|_2 \right\}.$$

When the estimated mean is close to the true feature vector $\boldsymbol{\theta}^\star$, the value for λ may be increased such that $\lambda \propto \frac{1}{\mathbb{E}\left[\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\mu}}\|_2^2\right]}$, effectively shrinking the confidence set and consequently the regret bound. Since $\mathbb{E}\left[\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\mu}}\|_2^2\right]$ cannot be usually determined directly, Cella *et al.* (2020) provide an upper bound $\mathbb{E}\left[\|\boldsymbol{\theta}^\star - \hat{\boldsymbol{\mu}}\|_2^2\right] \leq \sigma_{\text{total}} + \Delta\mu$, where $\sqrt{\Delta\mu}$ is the mean estimation error. The Biased OFUL algorithm achieves a bound on the expected transfer regret of $\mathcal{O}\left(d\sqrt{n \log\left(1 + \frac{n(\sigma_{\text{total}} + \Delta\mu)}{d}\right)}\right)$.

Updating the Prior in Quasi Bayesian (QB) algorithms In Bayesian bandit settings, a prior on the true task parameter is effectively assumed. As the true prior is usually unknown to the learner, a zero-mean distribution with covariance $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ is assumed, as previously discussed for linear Thompson sampling. Intuitively, the performance of a Bayesian bandit algorithm improves with a better prior distribution at hand. In the work of Peleg *et al.* (2022), the prior is shared by multiple bandit tasks and acts as a meta-distribution, that is, all true feature vectors are sampled from the common prior distribution. They propose a meta learning algorithm, to find estimations on the mean and covariance of the true prior. They rely on random exploration phases at the beginning of each task and use the accumulated data during exploration for their estimations. By updating the prior after each task, it can be exploited for subsequent tasks in Bayesian approaches such as the Thompson sampling algorithm to significantly reduce uncertainty and improve the regret bound. Peleg *et al.* (2022) use the notion of "relative regret" that is defined by the regret of a Bayesian algorithm due to not knowing the true prior of a setting. They provide upper bounds on the relative regret in the order of $\tilde{\mathcal{O}}\left(d^{\frac{5}{2}}\sqrt{nM}\right)$, with M as the number of tasks.

In Chapter 4 we contribute to the meta learning in bandits literature by first generalizing the setting of Cella *et al.* (2020) to learn the mean as well as the

covariance matrix of the meta distribution in the form of projection matrices and exploit them in a UCB-type algorithm as well as in a Bayesian setting by building a projection-based prior distribution.

2.2.2 Multi Task Learning in Bandits

Similarly to the meta learning setting, a shared structure between tasks or users is assumed in the multi-task case, but task learning ensues in parallel rather than sequentially. The number of tasks is fixed from the start and each task is associated with an index $t \in \{1, ..., M\}$, while the respective true feature vector is denoted as $\boldsymbol{\theta}^\star(t)$. Since the set of tasks is fixed in the settings we will discuss, we define the vertical concatenation of all estimated feature vectors $\hat{\boldsymbol{\Theta}}_k = \left[\hat{\boldsymbol{\theta}}_k^\top(t)\right]_{t\in\{1,...,M\}} \in \mathbb{R}^{M\times d}$. We denote the task drawn in round k as t_k and the respective action set as $\mathcal{A}_k$. As in the linear bandit setting, we define context vectors associated with an arm a as $\boldsymbol{x}_a$ and arms selected in round k as a_k. Additionally, we define the expected regret for the multi-task bandit setting as:

$$\mathcal{R}(n) = \sum_{k=1}^{n} \max_{a'\in\mathcal{A}_k} \left\{\boldsymbol{x}_{a'}^\top \boldsymbol{\theta}^\star(t_k)\right\} - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^\star(t_k)$$

There are various multi-task bandit scenarios, which make different assumptions on the underlying structure, some of which we discuss next.

Trace Norm Bandit A simple but effective assumption is that the set of tasks have a low rank structure, that is, the true feature vectors are allocated in a lower dimensional subspace with dimension $p < d$. Cella *et al.* (2023) propose a greedy algorithm called the trace-norm bandit using the trace-norm[2] regularization:

$$\hat{\boldsymbol{\Theta}}_k = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{M\times d}}{\arg\min} \left\{ \frac{1}{k} \sum_{i=1}^{k} \left(r_i - \boldsymbol{x}_{a_i}^\top \boldsymbol{\theta}(t_i)\right)^2 + \lambda \|\boldsymbol{\Theta}\|_* \right\},$$

with r_k as immediate reward received in round k and $\lambda \in \mathbb{R}$ as regularization parameter. Instead of learning the underlying structure directly, the trace norm regularization implicitly enforces a low-rank structure in the solution of the loss function. Actions are greedily selected according to:

$$a_k = \underset{a'\in\mathcal{A}_k}{\arg\max}\, \boldsymbol{x}_{a'}^\top \hat{\boldsymbol{\theta}}_k(t).$$

Similarly to single-task greedy algorithms, this approach requires the stronger assumption on the action sets $\mathcal{A}_k$, that the context vectors are sampled i.i.d. from

[2]We denote the trace norm as $\|\boldsymbol{\Theta}\|_*$, that is the sum of all singular values of a matrix $\boldsymbol{\Theta}$.

a zero-mean sub-Gaussian distribution. In order to guarantee an oracle inequality, a restricted strong convexity condition is imposed on the data-generating gram matrix that serves a similar purpose as compatibility assumptions discussed in Section 2.1.4. The Trace-Norm Bandit regret bound yields: $\tilde{\mathcal{O}}\left(M\sqrt{pn} + \sqrt{pdMn}\right)$.

Graph-Based Multi-Task Learning:

The use of graphs in multitask learning for bandits appears in the works of Cesa-Bianchi *et al.* (2013) and Yang and Toni (2018), where the relations between users or tasks are encoded within a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each task is associated with a different node in the set $\mathcal{V}$ such that $M = |\mathcal{V}|$, while the edges in $\mathcal{E}$ are an indicator of the Euclidean distance of the feature vectors of the tasks.

Gang of Bandits Linear Version In Cesa-Bianchi *et al.* (2013) an algorithm called Gang of Bandits linear version (GOBLin) is proposed, where the Laplacian $\boldsymbol{L}$ of the graph is used to construct modified context vectors for task t:

$$\phi_{\boldsymbol{L},t}(\boldsymbol{x}) = ((\boldsymbol{I}_M + \boldsymbol{L}) \otimes \boldsymbol{I}_d)^{-\frac{1}{2}} \left[\underbrace{\mathbf{0}_d^\top, ...}_{t-1 \text{ times}}, \boldsymbol{x}^\top, \underbrace{..., \mathbf{0}_d^\top}_{M-t \text{ times}} \right]^\top \in \mathbb{R}^{dM},$$

with $\mathbf{0}_d \in \mathbb{R}^d$ being the null vector, $\boldsymbol{I}_M$ and $\boldsymbol{I}_d$ as identity matrices in M and d dimensions respectively. The multitask problem of dimension d with M tasks is transformed to a single task problem of dimension dM, while the agent learns the concatenation of the feature vectors $\left[\boldsymbol{\theta}^\top(1), ..., \boldsymbol{\theta}^\top(M)\right]^\top \in \mathbb{R}^{dM}$ using the LinUCB algorithm. Due to the Laplacian, information of context vectors is spread throughout the graph and enables the update of all feature vectors in parallel. GOBLin achieves a regret bound of $\mathcal{O}\left(dM\sqrt{n} + \sqrt{dn|\mathcal{E}|}\right)$.

Graph UCB In Yang and Toni (2018) the same assumptions are made as in Cesa-Bianchi *et al.* (2013), but the graph is used differently. In their algorithm Graph UCB they apply a Laplacian regularization to their loss function to exploit the knowledge of the graph:

$$\hat{\boldsymbol{\Theta}}_k = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d \times M}}{\arg\min} \left\{ \sum_{i=1}^{k} \left(r_i - \boldsymbol{x}_{a_i}^\top \boldsymbol{\theta}(t_i) \right)^2 + \lambda \operatorname{Tr}\left(\boldsymbol{\Theta}^\top \boldsymbol{L} \boldsymbol{\Theta} \right) \right\},$$

with λ as some regularization parameter and t_i as task, the agent interacted with in round i. Here, the Laplacian is used in the regularization term and directly exploits the smoothness of the feature vectors over the graph. GraphUCB achieves a regret bound of $\tilde{\mathcal{O}}\left(d\sqrt{nM}\right)$

Clustered Bandit Tasks

A common assumption made is that for a set of M tasks, the tasks are grouped into C clusters, where the feature vectors of any two tasks within the same cluster are equal. Here, the clusters are unknown to the agent and have to be learned.

Cluster of Bandits Gentile *et al.* (2014) propose a linUCB-based algorithm called Cluster of Bandits (CLUB), where the agent maintains a dynamic graph in each round and every node is associated with a different user. The graph is initially fully connected and the edges are gradually erased depending on the estimated feature vectors $\hat{\boldsymbol{\theta}}_k(t)$ until only edges within the same estimated cluster are left. The CLUB algorithm has a regret bound of $\tilde{\mathcal{O}}\left(d\sqrt{Cn}\right)$.

Set-Based Clustering of Bandits While CLUB requires that every round a user is sampled i.i.d. from a uniform distribution *rho*, the work of Li *et al.* (2019) generalizes this uniformity assumption and allows the users to be drawn from any unknown distribution ρ. At the same time, their algorithm named set-based clustering of bandits (SCLUB) achieves a regret bound of $\tilde{\mathcal{O}}\left(d\sqrt{Cn}\right)$ that is independent of the distribution ρ.

Motivated by various examples, using graphs or cluster structures, we contribute to the field of multi-task bandits by proposing the network lasso model in Chapter 5 that utilizes a given graph to share information between tasks and implicitly learns the cluster structure for the agent to exploit. As in the work of Cella *et al.* (2023) our proposed algorithm utilizes greedy arm selection and requires a restricted eigenvalue condition on the context-data-generating matrix to hold.

Chapter 3

Hypothesis Transfer in Bandits by Weighted Models

This chapter, along with Appendix A, is a verbatim copy of Bilaj *et al.* (2023). Author contributions are stated as follows:

Author	**Author Position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Steven Bilaj	1	80	100	80	80
Title of paper:		Hypothesis Transfer in Bandits by Weighted Models			
Status in publication process:		Published			

Table 3.1: Chapter 3 author contributions.

Abstract We consider the problem of contextual multi-armed bandits in the setting of hypothesis transfer learning. That is, we assume having access to a previously learned model on an unobserved set of contexts, and we leverage it in order to accelerate exploration on a new bandit problem. Our transfer strategy is based on a re-weighting scheme for which we show a reduction in the regret over the classic Linear UCB when transfer is desired, while recovering the classic regret rate when the two tasks are unrelated. We further extend this method to an arbitrary amount of source models, where the algorithm decides which model is preferred at each time step. Additionally we discuss an approach where a dynamic convex combination of source models is given in terms of a biased regularization term in the classic LinUCB algorithm. The algorithms and the theoretical analysis of our proposed methods substantiated by empirical evaluations on simulated and real-world data.

3.1 Introduction

The *multi-armed bandit* problem (MAB) Thompson (1933); Robbins (1952); Bush and Mosteller (1953) revolves about maximizing the reward collected by playing

actions from a predefined set, with uncertainty and limited information about the observed payoff. At each round, the bandit player chooses an arm according to some rule that balances the exploitation of the currently available knowledge and the exploration of new actions that might have been overlooked while being more rewarding. This is known as the exploration-exploitation trade-off. MAB's find applications in several areas Bouneffouf *et al.* (2020), notably in recommender systems Li *et al.* (2010); Zhou *et al.* (2017); Liu *et al.* (2018a); Labille *et al.* (2021). In these applications, the number of actions to choose from can grow very large, and it becomes provably detrimental to the algorithm's performance to ignore any side information provided when playing an action or dependence between the arms Auer *et al.* (2002b). Considering such information defines the *Stochastic Contextual Bandits* Langford and Zhang (2007); Li *et al.* (2010); Chu *et al.* (2011); Abbasi-Yadkori *et al.* (2011) setting, where playing an action outputs a context-dependent reward, where a context can correspond to a user's profile and/or the item to recommend in recommender system applications. Hence, less exploration is required as arms with correlating context vectors share information, thus further reducing uncertainty in the reward estimation. This ultimately led to lower regret bounds and improved performance Abbasi-Yadkori *et al.* (2011).

While the stochastic contextual bandit problem solves the aforementioned issues, it disregards the possibility of learning from previously trained bandits. For instance, assume a company deploys its services in a new region. Then it would waste the information it has already learned from its previous recommending experience if it is not leveraged to accelerate the recognition of the new users' preferences. Such scenarios have motivated transfer learning for bandits Soare *et al.* (2014); Liu *et al.* (2018a); Suk and Kpotufe (2021); Labille *et al.* (2021), which rely on the availability of contexts of the previously learned tasks to the current learner. However, regarding a setup where context vectors correspond to items which have been selected by a user, privacy issues are encountered in healthcare applications Stark *et al.* (2019); Ras *et al.* (2021) for instance, the aim being to recommend a treatment based on a patient's health state. Indeed, accessing the contexts of the previous tasks entails the history of users' previous activities. Moreover, in engineering applications such as scheduling of radio resources Amrallah *et al.* (2020), storage issues Maiti *et al.* (2021); Liau *et al.* (2018); Xu and Zhao (2021) might arise when needing access to the context history of previous tasks. These problems would render algorithms depending on previous tasks' contexts inapplicable.

In this work, we aim to reduce exploration by exploiting knowledge from a previously trained contextual bandit accessible only through its parameters, thus accelerating learning if such model is related to the one at hand, and ultimately decreasing the regret. We extend this idea by including an arbitrary amount of models increasing the likelihood of including useful knowledge. To summarize our contributions, we propose a variation of the Linear Upper Confidence Bound (*LinUCB*) algorithm, which has access to previously trained models called source models. The

knowledge transfer takes place by using an evolving convex combination of sources models and a LinUCB model, called a target model, estimated with the collected data. The combination's weights are updated according to two different weighting update strategies which minimize the required exploration factor and consecutively the upper regret bound, while also taking a lack of information into consideration. Our regret bound is at least as good as the classic LinUCB one Abbasi-Yadkori *et al.* (2011), where the improvement depends on the quality of the source models. Moreover, we prove that if the source model used for transfer is not related to our problem, then it will be discarded early on and we recover the LinUCB regret rate. In other words, our algorithm is immune against negative transfer. We test our algorithm on synthetic and real data sets and show experimentally how the overall regret improves on the classic model.

The rest of the paper is organized as follows. We discuss related work in Section 3.2 and formulate our problem in Section 3.3, then we provide and analyse our weighting solution in Section 3.4. This is followed by an extension to the case where one has access to more than one trained model in Section 3.5. Finally, the performance of our algorithm is assessed in Section 3.6.

3.2 Related Work

We hereby discuss two families of contributions related to ours, namely transfer for multi-armed bandits, and hypothesis transfer learning.

Transfer for MAB's

To the best of our knowledge, tUCB Azar *et al.* (2013) is the first algorithm to tackle transfer in an MAB setting. Given a sequence of bandit problems picked from a finite set, it uses a tensor power method to estimate their parameters in order to transfer knowledge to the task at hand, leading to a substantial improvement over UCB. Regarding the richer contextual MAB setting, MT-LinUCB Soare *et al.* (2014) reduces the confidence set of the reward estimator by using knowledge from previous episodes. More recently, transfer for MAB's has been applied to recommender systems Liu *et al.* (2018a); Labille *et al.* (2021), motivated by the cold start problem where a lack of initial information requires more exploration at the cost of higher regret. The TCB algorithm Liu *et al.* (2018a) assumes access to correspondence knowledge between the source and target tasks, in addition to contexts, and achieves a regret of $O(d\sqrt{n \log n})$ as in the classic LinUCB case, with empirical improvement. The same regret rate holds for the T-LinUCB algorithm Labille *et al.* (2021), which exploits prior observations to initialize the set of arms, in order to accelerate the training process. The main difference of our formulation

with respect to the previous ones is that we assume having access only the the preference vectors of the previously learnt tasks, without their associated contexts, which goes in line with the Hypothesis Transfer Learning setting. Even with such a restriction, we keep the LinUCB regret rate and we show that the regret is lower in the case source parameters that are close to those of the task at hand.

Hypothesis Transfer Learning

Using previously learned models in order to improve learning on a new task defines the hypothesis transfer learning scenario, also known as model reuse or learning from auxiliary classifiers. Some lines of work consider building the predictor of the task at hand as the sum of a source one (possibly a weighted combination of different models) and the one learned from the available data points Yang *et al.* (2007); Duan *et al.* (2009); Tommasi *et al.* (2014). Such models were thoroughly analyzed in Kuzborskij and Orabona (2013, 2017); Perrot and Habrard (2015) by providing performance guarantees. The previously mentioned additive form of the learned model was further studied and generalized to a large family of transformation functions in Du *et al.* (2017). In online learning, the pioneering work of Zhao *et al.* (2014) relies on a convex combination instead of a sum, with adaptive weights. More recently, the `Condor` algorithm Zhao *et al.* (2020) was proposed and theoretically analyzed to handle the concept drift scenario, relying on biased regularization w.r.t. a convex combination of source models. Our online setting involves transfer with decisions over a large set of alternatives at each time step, thus it becomes crucial to leverage transfer to improve exploration. To this end, we use a weighting scheme inspired by Zhao *et al.* (2014) but that relies on exploration terms rather than on how the models approximate the rewards.

3.3 Problem Formulation

We consider a contextual bandit setting in which at each time k, playing an action a from a set $\mathcal{A}$ results in observing a context vector $\boldsymbol{x}_{a_k} \in \mathbb{R}^d$ assumed to satisfy $\|\boldsymbol{x}_{a_k}\| \leq 1$, in addition to a reward $r(k)$. We further define the matrix induced norm: $\|\boldsymbol{x}\|_{\boldsymbol{A}} \coloneqq \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$ for any vector $\boldsymbol{x} \in \mathbb{R}^d$ and any matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. The classical case aims to find an estimation $\hat{\boldsymbol{\theta}}$ of an optimal bandit parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ which determines the rewards r of each arm with context vector $\boldsymbol{x}_a$ in a linear fashion $r = \boldsymbol{x}_a^\top \boldsymbol{\theta}^* + \epsilon$ up to some σ-sub-Gaussian noise ϵ. The decision at time k is made according to an upper confidence bound (UCB) associated to $\hat{\boldsymbol{\theta}}(k)$:

$$a_k = \arg\max_{a\in\mathcal{A}} \boldsymbol{x}_a^\top \hat{\boldsymbol{\theta}}(k) + \gamma\sqrt{\boldsymbol{x}_a^\top \boldsymbol{A}^{-1}(k)\boldsymbol{x}_a}, \tag{3.1}$$

where $\gamma > 0$ is a hyperparameter estimated through the derivation of the UCB

later and $\boldsymbol{A}(k) \coloneqq \lambda \boldsymbol{I}_d + \sum_{k'=0}^{k-1} \boldsymbol{x}_{a_{k'}} \boldsymbol{x}_{a_{k'}}^\top$. The latter term in the sum equation 3.1 represents the exploration term which decreases the more arms are explored. $\hat{\boldsymbol{\theta}}(k)$ is computed through regularized least-squares regression with regularization parameter $\lambda > 0$: $\hat{\boldsymbol{\theta}}(k) = \boldsymbol{A}^{-1}(k)\boldsymbol{D}^\top(k)\boldsymbol{y}(k)$, with $\boldsymbol{D}(k) = [\boldsymbol{x}_{a_i}^\top]_{i\in\{0,\dots,k-1\}}$ and $\boldsymbol{y}(k) = [r(i)]_{i\in\{0,\dots,k-1\}}$ as the concatenation of selected arms' context vectors and corresponding rewards respectively. We alter this decision making approach with the additional use of a previously trained source bandit. Inspired by Zhao *et al.* (2014), we transfer knowledge from one linear bandit model to another by a weighting approach. We denote the parameters of the source bandit by $\boldsymbol{\theta}_S \in \mathbb{R}^d$. The bandit at hand's parameters are then estimated as:

$$\hat{\boldsymbol{\theta}} = \alpha_S \boldsymbol{\theta}_S + \alpha_T \hat{\boldsymbol{\theta}}_T(k), \tag{3.2}$$

with weights $\alpha_S, \alpha_T \geq 0$ satisfying $\alpha_S + \alpha_T = 1$. More important is how the exploration term changes and how it affects the classic regret bound. From Abbasi-Yadkori *et al.* (2011) we know that the upper bound of the immediate regret in a linear bandit algorithm directly depends on the exploration term of the UCB. We aim to reduce the required exploration with the use of the source bandits knowledge, in order to accelerate the learning process as well as reducing the upper regret bound. For the analysis we consider the pseudo-regret Audibert *et al.* (2009) defined as:

$$R(n) = n \max_{a\in\mathcal{A}} \boldsymbol{x}_a^\top \theta^* - \sum_{k=0}^{n-1} \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^*.$$

Our goal is to prove that this quantity is reduced if the source bandit is related to the one at hand, whereas its rate is not worsened in the opposite case.

3.4 Weighted Linear Bandits

The model we use features dynamic weights, thus at time k, we use the following model for our algorithm:

$$\hat{\boldsymbol{\theta}}(k) = \alpha_S(k)\boldsymbol{\theta}_S + \alpha_T(k)\hat{\boldsymbol{\theta}}_T(k),$$

with $\hat{\boldsymbol{\theta}}_T(k)$ being updated like in the classic LinUCB case Abbasi-Yadkori *et al.* (2011) and $\boldsymbol{\theta}_S$ remaining constant. To devise an update rules of the weights, we first re-write the new UCB expression as:

$$\text{UCB}(a) = \boldsymbol{x}_a^\top \left(\alpha_S(k)\boldsymbol{\theta}_S + \alpha_T(k)\hat{\boldsymbol{\theta}}_T(k)\right) + \left(\alpha_S(k)\gamma_S + \alpha_T(k)\gamma_T\right) \|\boldsymbol{x}_a\|_{\boldsymbol{A}^{-1}}, \tag{3.3}$$

with $\gamma_S \geq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_S\|_{\boldsymbol{A}(k)}$ and $\gamma_T \geq \left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_T(k)\right\|_{\boldsymbol{A}(k)}$ as confidence set bounds

for the source bandit and target bandit respectively. We retrieve the classic case by setting $\alpha_S(k)$ to zero *i.e.*erasing all influence from the source. The confidence set bound γ_T has already been determined in Abbasi-Yadkori *et al.* (2011).
As mentioned in Section 3.3 we aim to reduce the required exploration in order to reduce the upper regret bound. Thus we select the weights such that the exploration term in eq. (3.3) is minimized.

3.4.1 Weighting Update Strategies

We want to determine the weights after each time step such that:

$$\alpha_S, \alpha_T = \underset{\substack{\alpha'_S, \alpha'_T \geq 0 \\ \alpha'_S + \alpha'_T = 1}}{\arg\min}\, \alpha'_S \gamma_S + \alpha'_T \gamma_T. \tag{3.4}$$

The above minimization problem is solved for:

$$\alpha_S = \mathbb{1}_{\gamma_S \leq \gamma_T}, \quad \alpha_T = 1 - \alpha_S. \tag{3.5}$$

This strategy would guarantee an upper regret bound at least as good as the LinUCB bound in Abbasi-Yadkori *et al.* (2011) as will be shown in the analysis section later. However, without any knowledge of the relation between source and target tasks, our upper bound on the confidence set of the source bandit is rather loose:

$$\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_S\|_{\boldsymbol{A}(k)} = \sqrt{\lambda U^2 + \|\boldsymbol{D}(k)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_S)\|_2^2} \quad \leq \sqrt{4\lambda + \|\overline{\boldsymbol{y}}(k) - \boldsymbol{y}_S(k)\|_2^2},$$

with $\|\theta^* - \theta_S\|_2 = U$, $\boldsymbol{y}_S$ as the concatenation of the source estimated rewards and $\overline{\boldsymbol{y}}$ as the concatenation of the observed mean rewards for each arm. Naturally after every time step, each entry in $\overline{\boldsymbol{y}}$ corresponding to the latest pulled arm needs to be updated to their mean value. The mean values are taken in order to cancel out the noise term in the observations. Also, we have $U \leq 2$ in case the vectors show in opposing directions and we additionally assume that $\|\boldsymbol{\theta}^*\|, \|\boldsymbol{\theta}_S\| \leq 1$. An upper bound on the confidence set γ_T of the target bandit has been determined in Abbasi-Yadkori *et al.* (2011):

$$\gamma_T = \sqrt{d \log\left(1 + \frac{k}{d\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)}. \tag{3.6}$$

As such, γ_T grows with $\sqrt{\log(k)}$ and later on in the analysis we show if $\boldsymbol{\theta}_S \neq \boldsymbol{\theta}^*$ then an upper bound on γ_S grows with at least $\sqrt{k}$. Consequently, in theory there is some point in time where γ_S will outgrow γ_T, meaning that the source bandit will be discarded. As already mentioned, our estimation of γ_S can be loose due to our lack of information on the euclidean distance term U, thus we potentially waste a

good source bandit with this strategy. Additionally we would only use one bandit at a time this way instead of the span of two bandits for example. Alternatively we can adjust the strategy in equation 3.4 by adding a regularization term in the form of KL-divergence. By substituting $\alpha_T = 1 - \alpha_S$ we get:

$$\alpha_S(k+1) = \underset{\alpha_S \in [0,1]}{\arg\min} \left\langle \begin{pmatrix} \alpha_S \\ 1-\alpha_S \end{pmatrix}, \begin{pmatrix} \gamma_S \\ \gamma_T \end{pmatrix} \right\rangle + \frac{\mathrm{KL}(\boldsymbol{\alpha} \| \boldsymbol{\alpha}(k))}{\beta}, \tag{3.7}$$

with $\boldsymbol{\alpha} \coloneqq (\alpha_S, 1-\alpha_S)^\top$ being a vector containing both weights. The addition of the KL divergence term forces both weights to stay close to their previous value, where $\beta > 0$ is a hyper parameter controlling the importance of the regularization. Problem equation 3.7 is solved for:

$$\alpha_S(k+1) = \frac{1}{1 + \frac{1-\alpha_S(k)}{\alpha_S(k)} \exp(\beta(\gamma_S - \gamma_T))}, \tag{3.8}$$

which is a softened version of our solution in equation 3.5, but in this case the source bandit will not be immediately discarded if the upper bound on its confidence set becomes larger than the target bandit's.

3.4.2 Analysis

We are going to analyse how the upper regret bound changes, within our model in comparison to Abbasi-Yadkori *et al.* (2011). All proofs are given in the appendix. First we bound the regret for the hard update approach, not including the KL-divergence term in equation 3.5:

Theorem 3.4.1. *Let $\{\boldsymbol{x}_{a_k}\}_{k=0}^{n-1}$ be sequence in $\mathbb{R}^d$, $U \coloneqq \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|$ and R_T be the classic regret bound of the linear model Abbasi-Yadkori* et al. *(2011). Let $m \coloneqq \min(\kappa, n)$ and $\delta \leq \exp(-2\lambda)$. Then, with a probability at least $1-\delta$, the regret of the hard update approach for the weighted LinUCB algorithm is bounded as follows:*

$$R(n) \leq U\sqrt{8md \log\Big(1 + \frac{m}{d\lambda}\Big)(\lambda + m)} + R_T(n) - R_T(m) \leq R_T(n) \tag{3.9}$$

with κ satisfying:

$$\kappa = \left\lfloor 2\left[d\left(\frac{1}{U^2} - \lambda\right) + \lambda\left(\frac{2}{U^2} - \frac{1}{2}\right)\right]\right\rfloor. \tag{3.10}$$

The value for κ essentially gives a threshold such that we have $\gamma_S < \gamma_T$ for every $k < \kappa$. As expected, for better sources *i.e.*low values U, κ increases meaning the source is viable for more time steps. Also notable is how we see an increasing value for κ at high dimensional spaces. This is most likely due to the fact, that at higher

dimensions the classic algorithm requires more time steps, in order to find a suitable estimation, thus having a larger confidence set bound. In these instances a trained source bandit would be viable early on. The regret is reduced for lower values of U and the time κ at which a source is discarded is extended. For source bandits satisfying $\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 = 2$, we would retrieve the classic regret bound, preventing negative transfer.
Next we show what happens in case of a negative transfer for the softmax update strategy, *i.e.*the source does not provide any useful information at all and worsens the regret rate with $\gamma_S > \gamma_T$ at all time steps.

Theorem 3.4.2. *Let $\{\boldsymbol{x}_{a_k}\}_{k=0}^{n-1}$ be sequence in $\mathbb{R}^d$ and the minimal difference between confidence set bounds given as $\Delta_{\min} = \min_{k\in\{0,\dots,n-1\}}(\gamma_S(k) - \gamma_T(k))$, with $\gamma_S > \gamma_T$ for all time steps and the initial target weight denoted by $\alpha_T(0)$. Then with probability of at least $1-\delta$ an upper regret bound $R(n)$ in case of a negative transfer scenario is given by:*

$$R(n) \leq \frac{(1-\alpha_T(0))}{e\beta\alpha_T(0)(1-\exp(-\beta\Delta_{\min}))} + R_T(n) \tag{3.11}$$

Theorem 3.4.2 shows that in case of a negative transfer, the upper regret bound is increased by at most a constant term and vanishes in the case of $\beta \to \infty$ retrieving the hard update rule.

3.5 Weighted Linear Bandits with Multiple Sources

Up until now we only used a single source bandit, but our model can easily be extended to an arbitrary amount of different sources. Assuming we have M source bandits $\{\theta_{S,j}\}_{j=1}^{M}$, we define $\hat{\theta}$ as:

$$\hat{\boldsymbol{\theta}} = \sum_{j=1}^{M} \alpha_{S,j}\boldsymbol{\theta}_{S,j} + \alpha_T\hat{\boldsymbol{\theta}}_T, \tag{3.12}$$

with $\alpha_{S,j}, \alpha_T \geq 0\ \forall 1 \leq j \leq M$ and $\alpha_T + \sum_{j=1}^{M}\alpha_{S,j} = 1$. With this each source bandit yields its own confidence set bound $\gamma_{S,j}$. Similarly to eq. (3.3) we retrieve for the UCB with multiple sources:

$$\text{UCB}(a) = \boldsymbol{x}_a^\top \left(\sum_{j=1}^{M} \alpha_{S,j}(k)\boldsymbol{\theta}_{S,j} + \alpha_T(k)\hat{\boldsymbol{\theta}}_T(k)\right) + \boldsymbol{\alpha}^\top(k)\boldsymbol{\gamma}\|\boldsymbol{x}_a\|_{\boldsymbol{A}^{-1}(k)}, \tag{3.13}$$

with $\boldsymbol{\alpha}(k) = (\alpha_{S,1}(k), ..., \alpha_{S,M}(k), \alpha_T(k))^\top$ and $\boldsymbol{\gamma} = (\gamma_{S,1}, ..., \gamma_{S,M}, \gamma_T)^\top$. As for the weight updates the same single source strategies apply *i.e.*the minimization of the exploration term in the UCB function:

$$\boldsymbol{\alpha}(k+1) = \underset{\boldsymbol{\alpha} \in \mathcal{P}_{M+1}}{\arg\min}\, \boldsymbol{\alpha}^\top(k)\boldsymbol{\gamma} + \frac{1}{\beta}\mathrm{KL}(\boldsymbol{\alpha}||\boldsymbol{\alpha}(k)), \tag{3.14}$$

where $\mathcal{P}_{M+1}$ is the $(M+1)-$dimensional probability simplex. The solution of the previous problem is:

$$\alpha_{S,m}(k+1) = \frac{\alpha_{S,m}(k)\exp\left(-\beta\gamma_{S,m}\right)}{\sum_{j=1}^{M}\alpha_{S,j}(k)\exp(-\beta\gamma_{S,j}) + \alpha_T(k)\exp(-\beta\gamma_T)}. \tag{3.15}$$

This is basically the solution of eq. (3.8) generalized to multiple sources. In the decisions making it favors the bandit with the lowest upper bound γ of their confidence set. When we take the limit $\beta \to \infty$ in eq. (3.14) the KL-divergence term vanishes and we retrieve the hard case:

$$\alpha_{S,j} = \mathbb{1}_{\gamma_{S,j} = \min\left(\min_i \gamma_{S,i}, \gamma_T\right)} \tag{3.16}$$

which forces the weights to satisfy $\alpha_{S,m}, \alpha_T \in \{0, 1\}$ for every source index and for all time steps. Thus decision making is done by selecting one single bandit in each round with the lowest value of their respective confidence set bound γ. The regret of hard update strategy for multiple sources is given by the following theorem:

Theorem 3.5.1. *Let* $\{\boldsymbol{x}_{a_k}\}_{k=0}^{n-1}$ *be sequence in* $\mathbb{R}^d$ *and* $\min_m \|\boldsymbol{\theta}_{S,m} - \boldsymbol{\theta}^*\| = U_{\min}$ *and the classic regret bound of the linear model up to time step n given by* $R_T(n)$ *Abbasi-Yadkori* et al. *(2011). Let* $m \coloneqq \min(\kappa, n)$ *and* $\delta \leq \exp(-2\lambda)$. *Then with probability of at least* $1-\delta$ *the regret of the hard update approach for the weighted LinUCB algorithm with multiple sources is bounded by:*

$$R(n) \leq 4U_{\min}\sqrt{\kappa d \log\left(1 + \frac{\kappa}{d\lambda}\right)(\lambda + \kappa)} - R_T(m) + R_T(n) \leq R_T(n),$$

with κ *as:*

$$\kappa = \left\lfloor 2\left[d\left(\frac{1}{U_{\min}^2} - \lambda\right) + \lambda\left(\frac{2}{U_{\min}^2} - \frac{1}{2}\right)\right]\right\rfloor.$$

depending on $U_{\min}$ the multiple source approach benefits from the additional information as the upper bound corresponds to the best source overall. In case of the softmax update strategy, we need to show how the regret changes in case of a negative transfer scenario, *i.e.*the confidence set bounds of any source bandit is larger than the target bound at any time.

Theorem 3.5.2. *Let $\{\boldsymbol{x}_{a_k}\}_{k=0}^{n-1}$ be sequence in $\mathbb{R}^d$, a total of M source bandits being available indexed by j and the minimal difference between confidence set bounds set as $\Delta_{\min,j} = \min_{k\in\{0,\dots,n-1\}}(\gamma_{S,j}(k) - \gamma_T(k))$ for every source j with $\gamma_{S,j} > \gamma_T \ \forall j$ at every time step. Additionally the initial target weight is denoted by $\alpha_T(0)$. Then with probability $1-\delta$ an upper regret bound $R(n)$ in case of a negative transfer scenario is given by:*

$$R(n) \leq \frac{(1-\alpha_T(0))}{e\beta M \alpha_T(0)} \sum_{j=1}^{M} \frac{1}{(1-\exp(-\beta\Delta_{\min,j}))} + R_T \tag{3.17}$$

In comparison to the single source result, the additional constant is averaged over all sources. Depending on the quality, it can be beneficial to include more source bandits as potentially bad sources would be mitigated.

Algorithm 1: Weighted LinUCB

Initialize: $\hat{\boldsymbol{\theta}}_T(0)$ from $\mathcal{U}([0,1]^d)$, $\alpha_{S,j}(0) = (1-\alpha_T(0))/M = \frac{1}{2M}$, $U_j > 0$ $\gamma_{S,j} > 0 \ \forall j \in \{1,\dots,M\}$, $\delta \in [0,1]$, $\gamma_T > 0$, $\lambda > 0$, $\beta > 0$, $\boldsymbol{A}(0) = \lambda \boldsymbol{I}$, $\boldsymbol{b}(0) = \boldsymbol{0}$;

for $k \in \{0,\dots,n-1\}$ **do**

- Pull arm $a_k = \arg\max_a \mathrm{UCB}(a)$ taken from eq. (3.13);
- Receive estimated rewards from sources and real rewards: $r_{S,j}(k)|_{j\in\{1,\dots,M\}}, r(k)$;
- $\boldsymbol{A}(k+1) = \boldsymbol{A}(k) + \boldsymbol{x}_{a_k}\boldsymbol{x}_{a_k}^\top$;
- $\boldsymbol{b}(k+1) = \boldsymbol{b}(k) + r(k)\boldsymbol{x}_{a_k}$;
- $\hat{\boldsymbol{\theta}}_T(k+1) = \boldsymbol{A}^{-1}(k+1)\boldsymbol{b}(k+1)$;
- Store rewards $r_{S,j}(k)|_{j\in\{1,\dots,M\}}, r(k)$ in vectors $\boldsymbol{y}_{S,j}(k)|_{j\in\{1,\dots,M\}}, \boldsymbol{y}(k)$ respectively;
- Calculate $\overline{\boldsymbol{y}}(k)$ from $\boldsymbol{y}(k)$ such that each entry r corresponding to the latest arm a_k pulled is updated to the mean reward $\overline{r}$ of the respective arm;
- Update $U_j = \max_{i\in\{0,\dots,k\}} \frac{\left|\overline{r}(i) - r_{S,j}(i)\right|}{\left\|\boldsymbol{x}_{a_i}\right\|}$ for every j;
- $\gamma_{S,j} = \sqrt{\lambda U_j + \|\boldsymbol{y}_{S,j}(k) - \overline{\boldsymbol{y}}(k)\|}$;
- $\gamma_T = \sqrt{\lambda} + \sqrt{\log \frac{|\boldsymbol{A}(k)|}{\lambda^d \delta^2}}$;
- update source weights $\alpha_{S,j}(k+1)$ according either to softmax rule in eq. (3.15):
- or to the hard update rule in eq. (3.16);
- update target weight as: $\alpha_T(k+1) = 1 - \sum_{j=1}^{M} \alpha_{S,j}(k+1)$;

end

For the practical implementation we use $\gamma_T = \sqrt{\lambda} + \sqrt{\log \frac{|\boldsymbol{A}(k)|}{\lambda^d \delta^2}}$ which is also taken from Abbasi-Yadkori *et al.* (2011) and gives a tighter confidence set bound on the target estimator. Also we give an estimation for U_j by taking the maximum value of the lower bound induced by the Cauchy-Schwartz inequality $U_j = \|\boldsymbol{\theta}_{S,j} - \boldsymbol{\theta}^*\| \geq \max_{i \in \{0,\dots,k\}} \frac{\left|\bar{r}(i) - r_{S,j}(i)\right|}{\|\boldsymbol{x}_{a_i}\|}$ at each time step.

3.5.1 Biased Regularization

In Zhao *et al.* (2020) a similar approach of model reuse was used in a concept drift scenario for linear classifiers via biased regularization. In Kuzborskij and Orabona (2017) the risk generalization analysis for this approach was delivered in a supervised offline learning setting. Their mathematical formulation is stated as following: A classifier is about to be trained given a target training set $(\boldsymbol{D}, \boldsymbol{y})$ and a source hypothesis $\boldsymbol{\theta}_{src}$, which is specifically used for a biased regularization term. In contrast to our approach the weighting is only applied the source model, giving an alternate solution to the target classifier. Adapted to a linear bandit model, the optimization problem can be formulated as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{D}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_{src}\|^2. \tag{3.18}$$

$\boldsymbol{\theta}_{src}$ is a convex combination of an arbitrary amount of given source models $\{\boldsymbol{\theta}_j\}_{j \in \{1,\dots,M\}}$:

$$\boldsymbol{\theta}_{src} = \sum_{j=1}^{M} \alpha_j \boldsymbol{\theta}_j, \tag{3.19}$$

As in our model, these weights are not static and are updated after each time step. The update strategy is not chosen to minimize the upper regret bound but can be chosen such that the convex combination is as close as possible to the optimal bandit parameter. The UCB function is then simply given by:

$$\mathrm{UCB}(a) = \boldsymbol{x}_a^\top \hat{\boldsymbol{\theta}} + \gamma \|\boldsymbol{x}_a\|_{\boldsymbol{A}^{-1}(k)}, \tag{3.20}$$

with $\gamma = \sqrt{d \log\left(1 + \frac{k}{d\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} \|\boldsymbol{\theta}_{src} - \boldsymbol{\theta}^*\|_2$ and the solution to eq. (3.18):

$$\hat{\boldsymbol{\theta}} = \boldsymbol{A}^{-1}\boldsymbol{D}^\top \boldsymbol{y} - (\boldsymbol{A}^{-1}\boldsymbol{D}^\top \boldsymbol{D} - \boldsymbol{I})\boldsymbol{\theta}_{src}. \tag{3.21}$$

At some point in time we expect the weights to converge to a single source bandit closest to the optimal bandit. But contrary to our original model it is not possible for the model to discard all sources once the target estimation yield better upper bounds for their confidence sets. The upper regret bound is similar to the classic bound with the difference being in one term.

Theorem 3.5.3. *Let* $\{\boldsymbol{x}_{a_k}\}_{k=0}^{n-1}$ *be sequence in* $\mathbb{R}^d$ *and the upper bound of the biggest euclidean distance between any of the* M *source bandit indexed by* m *and optimal bandit parameter given by* $\max_m \|\boldsymbol{\theta}_{S,m} - \boldsymbol{\theta}^*\| \leq U_{\max}$*, then with probability of at least* $1-\delta$ *the regret of the biased LinUCB algorithm with multiple sources is upper bounded by:*

$$R(n) \leq \sqrt{8nd\log(\lambda + n/d)}\left(\sqrt{d\log\left(1+\frac{n}{d\lambda}\right)+\log\left(\frac{1}{\delta^2}\right)}+\sqrt{\lambda}U_{\max}\right) \tag{3.22}$$

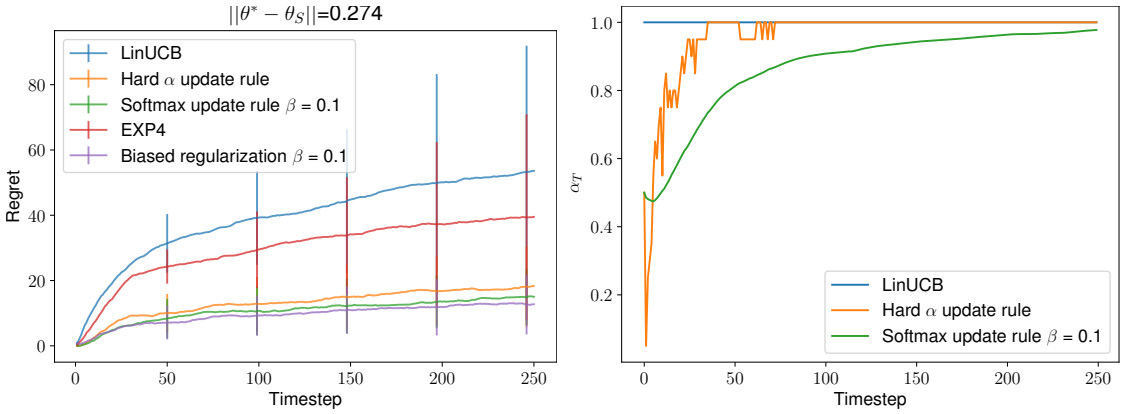
Since we are looking for an upper bound, U is dominated by the largest euclidean distance between the optimal bandit parameter and all given source bandits. Theorem 3.5.3 differs from the classic case in the regularization related parameters where we have $\sqrt{\lambda}U_{\max}$ instead of $\sqrt{\lambda}\|\boldsymbol{\theta}^*\|$. For sources with low values of U, we improve the overall regret.

3.6 Experimental Results

We test the presented algorithms, *i.e.*the weighted model algorithm as well as the biased regularization algorithm, for single source and multiple source transfers on synthetic and real data sets. The plots include the results from the classical LinUCB approach as well as the EXP4 approach from Lattimore and Szepesvári (2020) with target and source models acting as expert, for comparison purposes. Additionally to the regret plots we also showcase the mean of the target weight as a function of time to see how the relevancy of the target estimation evolved.

3.6.1 Synthetic Data Experiments

Our synthetic experiments follow a similar approach to Liu *et al.* (2018a). The target context feature vectors $\boldsymbol{x}_a$ are drawn from a multivariate Gaussian with variances sampled from a uniform distribution. We chose the number of dimensions $d = 20$ and the number of arms to be 1000. Our optimal target bandit parameter is sampled from a uniform distribution and scaled such that $\|\boldsymbol{\theta}^*\| \leq 1$, thus the rewards are implicitly initialized as well with $r = \boldsymbol{x}_a^\top \boldsymbol{\theta}^* + \epsilon$, with some Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1/\sqrt{2\pi}$. The source bandit parameters $\boldsymbol{\theta}_{S,m}$ are initialized by adding a random noise vector $\boldsymbol{\eta}_m$ to the optimal target bandit parameters for every source bandit to be generated $\boldsymbol{\theta}_{S,m} = \boldsymbol{\theta}^* + \boldsymbol{\eta}_m$. This way we ensure that there is actual information of the target domain in the source bandit parameter. We could also scale $\boldsymbol{\eta}_m$ to determine how much information the respective source yields about the target domain. The regularization parameter was constantly chosen to be $\lambda = 1$ and the initial weights are equally distributed among all available bandit



(a) Regret evolution plot labeled by confidence set bound.

(b) Evolution of the target weight α_T.

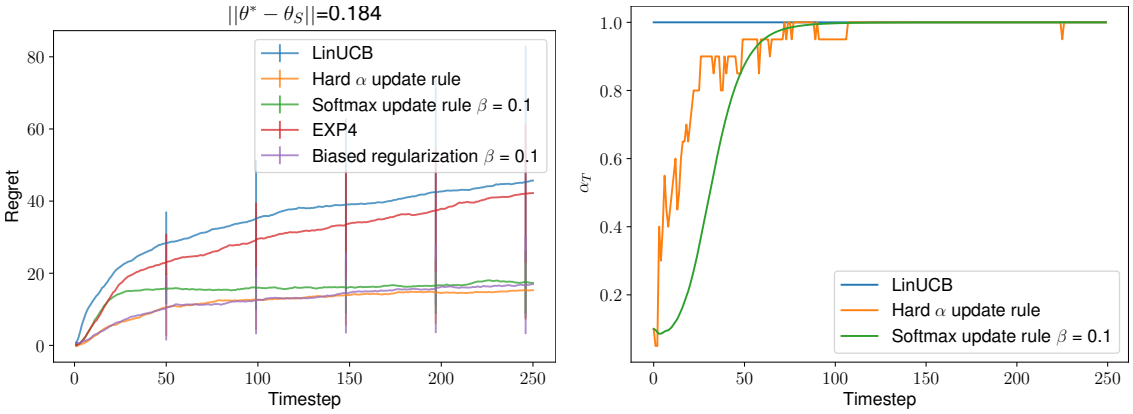
Figure 3.1: Regret and weight evolution for single source transfer scenario on synthetic data sets. The blue lines showcase the classic LinUCB results. The vertical lines indicate the standard deviation.

parameters: $\alpha_T = \alpha_{S,m} = \frac{1}{M+1}$. The shown results are the averaged values over 20 runs.

As we showed in Section 3.4 the upper regret bound is lower for $\beta \to \infty$ *i.e.*the hard update rule which ignores the KL-divergence in the optimization, but we see overall better results than in the classic case with the softmax update strategy as well. The inclusion of eight more source bandits in Figure 3.2 improves the sources slightly, though it should be mentioned that all sources generated were similar in quality. Thus we would expect higher improvements in the regret when including significantly better sources. The EXP4 algorithm on the other hand does not perform as well when increasing the number of experts.

3.6.2 Real Data Experiments

The real data sets used for our purposes are taken from the MovieLens sets. Their data include an assemble of thousands of users and corresponding traits such as age, gender and profession as well as thousands of movies and their genres. Every user has a rating from 1 to 5 given to at least 20 different movies. The movies, rated by a user, function as the available arms for that particular user. The information of the movies apart from the title itself are solely given by their genres. Each movie may have up to three different genres and there are 18 different genres in total. Arms, which are linked to the movies, have context vectors depending on the movies genre only. We design 18-dimensional context vector with each dimension representing a genre. If the movie is associated with a particular genre, the respective dimensional feature is set as $x_i = \frac{1}{\sqrt{S}}$ with S as the total number

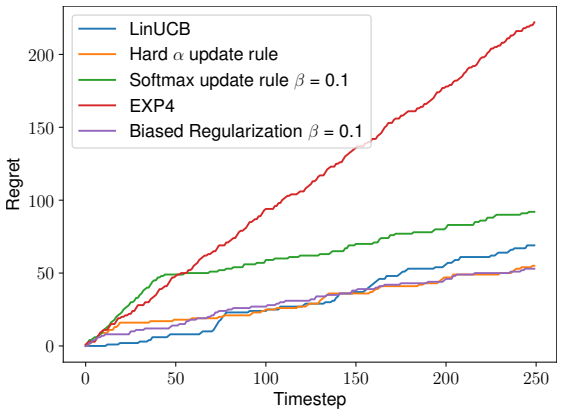


(a) Regret evolution plot labeled by the lowest confidence set bound of all available sources.

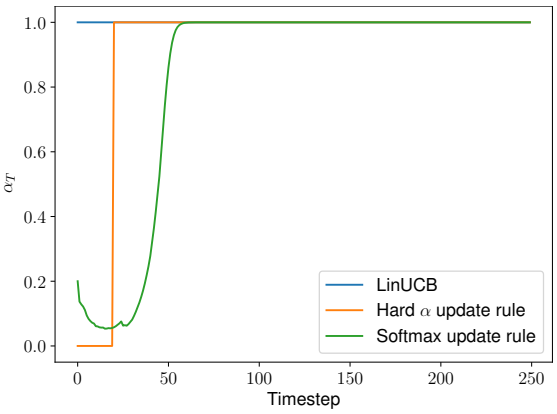
(b) Evolution of the target weight α_T. Since multiple sources are present, the initial weight is reduced

Figure 3.2: Regret and weight evolution for multiple source transfer scenario (9 sources) on synthetic data sets. The blue lines showcase the classic LinUCB results. The vertical lines indicate the standard deviation.

of genres the movie is associated with. This way we guarantee that every context vector is bounded by 1. the reward of an arm in our bandit setting is simply given by the user rating.
For our purposes we require source bandits for the transfer learning to take place. Therefore we pretrained a bandit for every single user, given all of the movie information, with the classic LinUCB algorithm and stored the respective parameters. This way every single user can function as a potential source for a different user. With all of the users available we grouped them according their age, gender and profession. We enforce every user to only act as source to other users with similar traits. This stems from a general assumption that people with matching traits may also have similar interests. This is a very general assumption made but given all of the information, it is the easiest way to find likely useful sources for every user. In Figure 3.3 the results for two individuals of two different groups of users respectively are showcased. Instead of only using one source, we used the multiple source strategy and made use of every user of the same group the individuals are located in, since this way we have a higher chance to find good sources. Even though the real data is far from guaranteed to have a linear reward structure, as well as the fact that important information on the arms' contexts are not available, since ratings usually not only depend on the movie genre, we find satisfying results with converging regrets as well as improved learning rates when including sources.



(a) Regret evolution plot with user data taken from the group of 35 to 44 years old female lawyers.



(b) Target weight evolution plot with the respective algorithms labeled with user data taken from the group of 35 to 44 years old female lawyer.

Figure 3.3: Regret evolution for multiple source transfer scenario on real data sets taken from Movielens data. A group of users are shown with one bandit trained for a random user of each group, while the rest of the users act as source to the respective user. The blue lines showcase the classic LinUCB results.

3.7 Discussion and Outlook

This work shows that our approach to make use of information from different tasks, without having actually access to concrete data points, is efficient, given the improved regrets. We have proven an upper regret bound of our weighted LinUCB algorithm with the hard update strategy at least as good as the classic LinUCB bound with a regret rate of $O(d\sqrt{n \log n})$, and a converging sub-linear negative-transfer term when using the softmax update strategy. Further argument for the utility of our model was given with synthetic and real data experiments. The synthetic data sets showed promising results especially with the softmax update strategy, even without having a guaranteed improved regret bound. The softmax approach uses a convex combination of models, which might be more practical than using one model at a time especially when it comes to high quality sources. This further raises the question whether different weighting update rules, which yield solutions consisting of a span of source models, might be more efficient for transfer. The inclusion of multiple sources further improved the results, indicating that using information from multiple different tasks is more effective then just one, which aligns with our theoretical result in Theorem 3.5.1. The real-world data experiments showed improvements as well, even when considering that the rewards did not necessarily follow a linear model and that the available features for the context vector were rather sparse, the transfer of information from similar users almost

always led to lower regrets.

In upcoming projects we intend to adapt our approach to non-linear models such as kernelized bandits, since the convex weighting is not limited to just linear models, as well as give a proper regret bound for the softmax update strategy. There is potential in using our transfer model to non stationary bandits, such that each prior estimation of the bandit parameter may act as source for the current setting, thus making use of the information collected in prior instances of the bandit setting. In this case we would need to make assumptions of the change rate of the tasks after a certain amount of time steps. Previous algorithms on non-stationary bandits Russac *et al.* (2019) perform weighting on data points and discard them after some time steps, without evaluating the benefit of the data beforehand. In our setting, previously trained bandit parameters would be used according to their performance.

Chapter 4

Meta Learning in Bandits within Shared Affine Subspaces

This chapter, along with Appendix B, is a verbatim copy of Bilaj *et al.* (2024). Author contributions are stated as follows:

Author	**Author Position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Steven Bilaj	1	80	100	80	80
Title of paper:		Meta Learning in Bandits within Shared Affine Subspaces			
Status in publication process:		Published			

Table 4.1: Chapter 4 author contributions.

Abstract We study the problem of meta learning several contextual stochastic bandit tasks by leveraging their concentration around a low-dimensional affine subspace, which we learn via online principal component analysis to reduce the expected regret over the encountered bandits. We propose and theoretically analyze two strategies that solve the problem: One based on the principle of optimism in the face of uncertainty and the other via Thompson sampling. Our framework is generic and includes previously proposed approaches as special cases. Besides, the empirical results show that our methods significantly reduce the regret on several bandit tasks.

4.1 Introduction

In several real-world applications, such as website design and healthcare, the system recommends an item to a user upon observing some side information depending on the user and the corresponding item. Upon receiving the recommendation, the user sends feedback to the system that captures his interest in the recommendation (Glowacka *et al.*, 2019; Bouneffouf *et al.*, 2020; Atan *et al.*, 2023). One can interpret the feedback as a reward that characterizes the suitability of the selected

recommendation or action with the final objective of maximizing the cumulative payoff over time. At the same time, such a selection might be suboptimal due to the incomplete knowledge of the environment. This *exploration-exploitation* tradeoff, along with the side information, is formalized by the *contextual multi-armed bandit (CMAB)* problem (Langford and Zhang, 2007; Li *et al.*, 2010; Chu *et al.*, 2011; Abbasi-Yadkori *et al.*, 2011; Nourani-Koliji *et al.*, 2022), a notable extension of the *multi-armed bandit (MAB)* problem (Thompson, 1933; Robbins, 1952).

In the applications mentioned above, the tasks often relate to each other despite being different. For instance, subgroups of patients have comparable features. As another example, holidays or discount periods promote similar interests in the products of an e-commerce website. That observation motivates us to look beyond a single task to uncover a relation between different ones to accelerate learning on newly encountered tasks. That problem, referred to as *meta learning* or *learning-to-learn (LTL)*, has mainly appeared in the offline learning literature so far (Hutter *et al.*, 2019). Nevertheless, an emergent body of literature combines LTL and MAB to accelerate learning and reduce the average regret per task (Cella *et al.*, 2020; Cella and Pontil, 2021; Bilaj *et al.*, 2023). In the linear contextual setting, an assumption about the preference vectors captures the relation between the tasks.

In this work, we assume that the feature vectors stem from a distribution that concentrates around a low dimensional subspace, *i.e.,* its variance is explained by a limited number of principal components. We propose learning this structure using online *principal component analysis (PCA)*. We then exploit that knowledge to develop two decision-making policies. The first policy relies on the principle of optimism in the face of uncertainty for linear bandits (OFUL) (Chu *et al.*, 2011; Abbasi-Yadkori *et al.*, 2011), and the second is a Thompson sampling policy (Russo *et al.*, 2018; Agrawal and Goyal, 2013). Analytically, we establish per-task regret upper bounds for both strategies that theoretically prove the benefit of learning such a structure. Moreover, our empirical evaluations of our methods using simulated and real-world data sets confirm their benefits.

Our paper is organized as follows. We review meta learning and related themes for bandit problems in Section 4.2. Then we formulate our problem in Section 4.3. We describe the subspace learning procedure in Section 4.4 to use in our proposed algorithms in Sections 4.5 and 4.6. Finally, we empirically assess our algorithms in Section 4.8.

4.2 Related Work

Learning to learn was first developed for offline learning (Thrun, 1998; Baxter, 2000; Hutter *et al.*, 2019) as a sub-field of transfer learning. In this paradigm, one seeks to learn a structure shared by many tasks to generalize to new ones. That structure can be encoded in several ways such as a prior over the task distribution

(Amit and Meir, 2018; Rothfuss *et al.*, 2021), a kernel (Aiolli, 2012), a common mean around which tasks concentrate (Denevi *et al.*, 2018), or an approximate low dimensional manifold (Jiang *et al.*, 2022), to name a few.

Recently, meta learning received attention in the online setting (Finn *et al.*, 2019), more precisely, in the case of bandit feedback. The main idea is that the learner interacts sequentially with bandit problems, so the meta learned shared structure accelerates exploration for upcoming tasks. In this setting, the objective is to improve the regret guarantees compared to those achievable by considering each task separately. The notion of regret can capture such guarantees; nevertheless, it has several definitions depending on the line of work. We distinguish mainly two regret types in a multi-task scenario: *transfer regret* and *meta regret*. The former depends on the number of learned tasks, whereas the latter takes an expectation on a possibly infinite number of tasks.

Concerning transfer regret, the goal is to prove sub-linear regret in the number of tasks. If the learner considers each task independently, the total regret over tasks is linear in the number of tasks. Within a task, the expected transfer regret is linear in the number of rounds. References Cella and Pontil (2021) and Cella *et al.* (2022) prove that if preference vectors have a low-rank structure, then learning it improves performance.

In the setting of Bayesian bandits, instead of assuming that the agent knows the true prior over tasks, a recent line of work proposes to learn that distribution. For example, Bastani *et al.* (2019) studies the dynamic pricing problem and proposes a Thompson sampling approach. Reference Kveton *et al.* (2021) generalizes the scope of the stochastic MAB problem by developing a meta-Thompson Sampling (meta-TS) algorithm. Basu *et al.* (2021) improves the guarantees of Kveton *et al.* (2021) via a modification of meta-TS. It also generalizes the core idea to other bandit settings, such as linear and combinatorial bandits. While Basu *et al.* (2021) and Kveton *et al.* (2021) study learning the mean of the tasks with a known covariance matrix, Peleg *et al.* (2022) relaxes that assumption. It proposes a general multivariate Gaussian prior learning framework that applies to several prior-update-based bandit algorithms. In the nonlinear contextual bandit case, Kassraie *et al.* (2022); Schur *et al.* (2022) investigate learning a shared kernel. Concerning the second type of guarantees, Cella *et al.* (2020) proves that the regret expectation over a potentially infinite number of tasks shrinks to 0 provided that the ridge regularization parameter is inversely proportional to the tasks' variance, and that said variance approaches 0.

Another line of work (Boutilier *et al.*, 2020; Kveton *et al.*, 2020; Yang and Toni, 2020) takes inspiration from the policy gradient methods (Williams, 1992) and aims to learn hyperparameters of policies to maximize the expected cumulative reward. Besides, meta learning is also applicable to solve problems in other settings concerning the reward generating mechanism, such as the non-stationary case (Azizi *et al.*, 2022), and more generally the adversarial case (Balcan *et al.*, 2022).

Multi-task learning is a field closely related to meta learning. The main difference between the two is the following: The former is about simultaneously learning over a finite family of bandit tasks without being concerned with generalization over future ones. That method is applied to solve the unstructured stochastic bandit case (Azar *et al.*, 2013), where although the interaction with tasks is sequential, they are finite. Therefore, the agent might encounter the same bandit problem more than once and can leverage the previous experience. Besides, In the case of contextual bandits, a low dimensional structure (Cella and Pontil, 2021; Cella *et al.*, 2023; Yang *et al.*, 2020a) or prior knowledge of the relations between tasks (Yang *et al.*, 2020b) provably reduces the regret

In this work, we borrow the concept of low dimensional structure from multi-task learning and leverage it with the concentration of tasks around some space region to improve the regret bound over a family of contextual linear bandits tasks. Indeed, assumptions such as high task concentration around a mean or strictly belonging to a low-dimensional subspace are restrictive. Thus, we aim at relaxing them. Finally, our approach is interpretable as learning an approximation of the covariance matrix of the tasks where the total variance is dominated by the contributions of a few principal components that span the subspace so it tightly relates to Peleg *et al.* (2022); Nevertheless, one of our proposed algorithms does not rely on the prior update.

4.3 Problem Formulation

We consider an agent (learner, interchangeably) that sequentially interacts with several contextual bandit tasks. While learning one task over n rounds, at each round k, the learner selects an arm a_k from a dynamic set of arms $\mathcal{A}_k$ with associated context vector $\boldsymbol{x}_{a_k} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{x}_{a_k}\| \leq 1$. Then it receives a reward $r_k = \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^* + \epsilon_k$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the true task parameter to estimate. For different tasks, $\boldsymbol{\theta}^*$ is independently drawn from a probability distribution ρ over $\mathbb{R}^d$ (i.i.d.) with mean $\boldsymbol{\mu}$. Besides, they are bounded, formally, $\|\boldsymbol{\theta}^*\| \leq V$ for some $V > 0$.[1] Moreover, ϵ_k is the zero-mean 1-sub-Gaussian noise such that $\{\epsilon_k\}_{k=1}^{n-1}$ are independent and identically distributed (i.i.d).

Our main assumption is that the distribution ρ has low variance along certain directions in space which ought to be learnt. Assumption 4.3.1 states this requirement formally. Besides, an illustration of a sampling from such a task distribution in 3 dimensions appears in Figure 4.1. Finally, we denote the covariance of ρ as $\boldsymbol{\Sigma}$ with ordered eigenvalues $\sigma_1 \geq \cdots \geq \sigma_d$.

Assumption 4.3.1. *There exists an orthogonal projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ with*

[1] Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm.

rank p such that:

$$\begin{aligned}\mathrm{Var}_\rho &\coloneqq \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\left(\boldsymbol{\theta}^*-\boldsymbol{\mu}\right)\right\|^2\right]\\ &\ll \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{P}\left(\boldsymbol{\theta}^*-\boldsymbol{\mu}\right)\right\|^2\right]\\ &\leq \mathrm{Var}_{\max} \coloneqq \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{\theta}^*-\boldsymbol{\mu}\right\|^2\right],\end{aligned}$$

$$\mathrm{Var}_\rho \ll \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{P}\boldsymbol{\theta}^*\right\|^2\right].$$

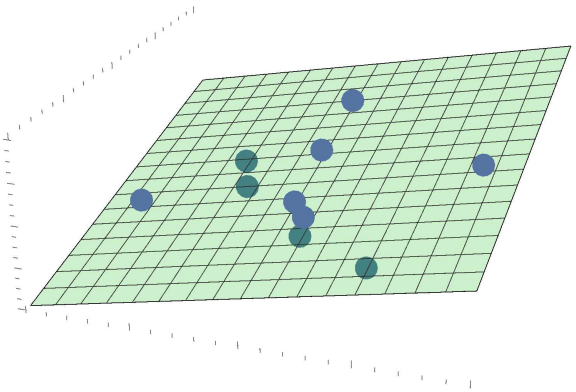


Figure 4.1: Sample of task parameters (blue points) from a distribution with low variance along one dimension.

Our goal is to learn $\boldsymbol{P}$ and $\boldsymbol{\mu}$ as well as bound the expected transfer regret $\mathcal{R}(n)$ adapted from Cella *et al.* (2020) defined as:

$$\mathcal{R}(n) = \mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\mathbb{E}\left[\sum_{k=0}^{n-1}\left(\boldsymbol{x}_{a_k^*}-\boldsymbol{x}_{a_k}\right)^\top\boldsymbol{\theta}^*\right]\right], \tag{4.1}$$

where $a_k^* \coloneqq \arg\max_{a\in\mathcal{A}_k}\boldsymbol{x}_a^\top\boldsymbol{\theta}^*$ is the optimal arm at round k. We propose two different approaches to exploit the knowledge of $\boldsymbol{P}$. First we present a variation of the standard LinUCB algorithm (Abbasi-Yadkori *et al.*, 2011) by adjusting the regularization term in the regularized least squares optimization problem. Our second approach is a variation of the linear Thompson Sampling algorithm (Agrawal and Goyal, 2013), where we adjust the covariance term of the normal distribution from which a task parameter is sampled from after every task according to the learned projection.

4.4 Subspace Learning

We use an online PCA version, namely, Candid Covariance-Free Incremental Principal Component Analysis (CCIPCA) Cardot and Degras (2015), to learn the underlying subspace structure from estimated task parameters. The core idea is to find an approximation of a set of orthonormal vectors that represent the principal components of a vertically concatenated data set $\boldsymbol{\Theta} = [(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}})^\top]_{i\in\{1,\dots,t\}}$, with $\bar{\boldsymbol{\theta}} \coloneqq \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}(i)$. Vector $\boldsymbol{\theta}(i)$ denotes the ith task, which was estimated after a total of at least n rounds. Upon finishing a task after n rounds, the agent updates the learned projection matrix. Nevertheless, applying PCA is costly in the long run, whereas an online estimation mitigates the costs while offering sufficient estimations on the learned projection. Starting with a set of orthonormal eigenvectors $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_d\}$ and their corresponding eigenvalues $\{\sigma_1, \dots, \sigma_d\}$ based on the covariance matrix $\frac{1}{t}\boldsymbol{\Theta}^\top\boldsymbol{\Theta}$, we define $\boldsymbol{v}_j \coloneqq \sigma_j\boldsymbol{u}_j$ for $j \in \{1, \dots, d\}$ as the set of scaled principal components. Here we assume $\bar{\boldsymbol{\theta}} = \mathbf{0}$, for the general case, the task parameters have to be centralized. Each additional task parameter $\boldsymbol{\theta}(i)$ adjusts the estimation of $\boldsymbol{v}_j$ *i.e.* after every round, every principal component $\boldsymbol{v}_j$ will be updated as

$$\boldsymbol{v}_{j,i+1} = \frac{i}{i+1}\boldsymbol{v}_{j,i} + \frac{1}{i+1}\boldsymbol{z}_{i+1}\boldsymbol{z}_{i+1}^\top\frac{\boldsymbol{v}_{j,i}}{\|\boldsymbol{v}_{j,i}\|}, \tag{4.2}$$

with $\boldsymbol{z}_i$ determined to ensure orthogonality of the eigenvector estimations. Formally, to compute $\boldsymbol{v}_{j,i+1}$ we have:

$$\boldsymbol{z}_{i+1} = \boldsymbol{\theta}(i+1) - \sum_{j'=1}^{j-1}\left(\boldsymbol{\theta}^\top(i+1)\boldsymbol{u}_{j',i}\right)\boldsymbol{u}_{j',i}.$$

CCIPCA is especially beneficial as it is hyperparameter-free. Besides, it estimates the eigenvalues and the corresponding eigenvectors of all principal components. The eigenvalue estimations are essential when choosing the rank p of the projection, which is generally unknown. The vectors $\boldsymbol{u}_i$ with the p highest values σ_i are selected as principal components. We define their horizontal concatenation as $\boldsymbol{U} = [\boldsymbol{u}_j]_{j\in\{1,\dots,p\}} \in \mathbb{R}^{d\times p}$.

Remark 4.4.1. *The choice of p depends on the respective eigenvalues, a common choice would be to maximize the eigengap, thus $p = \arg\max_{p'} \sigma_{p'} - \sigma_{p'+1}$.*

The projection matrix $\boldsymbol{P}$ with rank p as well as the orthogonal projection $\boldsymbol{P}^\perp$ with rank $q = d - p$ can then be constructed using of the principal components as

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top, \quad \boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P} \tag{4.3}$$

We will use the learned projections to exploit the low dimensional subspace structure in both LinUCB and Thompson sampling setting.

4.5 Projection Meta learning with LinUCB

In this section, we present our contextual bandit algorithms based on LinUCB.

4.5.1 Basics of LinUCB

In classic LinUCB, at each round k, the agent uses the collection of previously selected actions $\boldsymbol{D}_k = [\boldsymbol{x}_{a_i}^\top]_{i\in\{0,\dots,k-1\}}$ and the corresponding rewards $\boldsymbol{y}_k = [r_i]_{i\in\{0,\dots,k-1\}}$ to estimate the task parameter $\boldsymbol{\theta}_k$ by solving the following regularized least squares optimization problem:

$$\boldsymbol{\theta}_k = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min} \|\boldsymbol{D}_k\boldsymbol{\theta} - \boldsymbol{y}_k\|^2 + \lambda\|\boldsymbol{\theta}\|^2, \tag{4.4}$$

with $\lambda > 0$ being the regularization parameter. The solution of equation 4.4 is the *ridge estimator*. Given that, the learner selects an action that maximizes the UCB index

$$\text{UCB}(a) = \boldsymbol{x}_a^\top\boldsymbol{\theta}_k + \gamma_k\|\boldsymbol{x}_a\|_{\boldsymbol{A}_k^{-1}},^2 \tag{4.5}$$

with $\boldsymbol{A}_k \coloneqq \lambda\boldsymbol{I} + \boldsymbol{D}_k^\top\boldsymbol{D}_k$, $\boldsymbol{\theta}_k = \boldsymbol{A}_k^{-1}\boldsymbol{D}_k^\top\boldsymbol{y}_k$ and $\gamma_k > 0$ as an upper bound on the confidence set radius proposed in Abbasi-Yadkori *et al.* (2011). The additional term scaling is essential for exploration as $\|\boldsymbol{x}\|_{\boldsymbol{A}_k^{-1}}$ is maximized for context vectors that have the least correlation with already explored arms.

4.5.2 LinUCB with Projection Bias

In our first proposal, we enhance the LinUCB by including the knowledge of the projection matrix $\hat{\boldsymbol{P}}$. The agent learns $\hat{\boldsymbol{P}}$ by an online PCA algorithm using the parameters of t already learned tasks. To enforce the knowledge of the affine subspace during learning, we formulate the following optimization problem for a given task, where we define $\hat{\boldsymbol{\theta}}_k$ as the minimizer over $\boldsymbol{\theta} \in \mathbb{R}^d$ in the following objective:

$$\|\boldsymbol{D}_k\boldsymbol{\theta} - \boldsymbol{y}_k\|^2 + \lambda_1\left\|\hat{\boldsymbol{P}}^\perp(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\right\|^2 + \lambda_2\left\|\hat{\boldsymbol{P}}\boldsymbol{\theta}\right\|^2, \tag{4.6}$$

with $\hat{\boldsymbol{P}}^\perp \coloneqq \boldsymbol{I} - \hat{\boldsymbol{P}}$, $\lambda_1 > 0$ and $\lambda_2 > 0$. Besides,
$\bar{\boldsymbol{\theta}} \coloneqq \frac{1}{t}\sum_{i=1}^t \boldsymbol{\theta}(i)$ is the mean of the ridge regression estimators of the t previous

[2]Throughout the paper, $\|\cdot\|_{\boldsymbol{A}}$ denotes the weighted norm: $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}}$.

tasks. We justify the explicit choice of the regularization parameters in the analysis. Problem equation 4.6 has a closed form solution given by

$$\hat{\boldsymbol{\theta}}_k = (\boldsymbol{D}_k^\top \boldsymbol{D}_k + \lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}})^{-1} \left(\boldsymbol{D}_k^\top \boldsymbol{y}_k + \lambda_1 \boldsymbol{w} \right), \tag{4.7}$$

with $\boldsymbol{w} \coloneqq \hat{\boldsymbol{P}}^\perp \bar{\boldsymbol{\theta}}$. The second regularization term in eq. (4.6) scaling with λ_2 is necessary so that our closed form solution in eq. (4.7) is well defined i.e., it enables us to determine the inverse of

$$\boldsymbol{B}_k \coloneqq \boldsymbol{D}_k^\top \boldsymbol{D}_k + \lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}^\perp. \tag{4.8}$$

The case $\hat{\boldsymbol{P}} = \boldsymbol{I}$, which implies that all tasks are highly concentrated around the vector $\boldsymbol{w} = \bar{\boldsymbol{\theta}}$, would correspond to the setting of Cella *et al.* (2020).

Action selection is based on the principle of optimism in the face of uncertainty (OFUL), we propose an alternative UCB index by estimating the difference between mean reward r and estimated reward $\hat{r}$:

$$\begin{aligned} |\hat{r} - \mathbb{E}(r|\boldsymbol{x})| &= |\boldsymbol{x}^\top (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)| \\ &\leq \left\| \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \right\|_{\boldsymbol{B}_k} \|\boldsymbol{x}\|_{\boldsymbol{B}_k^{-1}} \leq \gamma_k \|\boldsymbol{x}\|_{\boldsymbol{B}_k^{-1}}, \end{aligned}$$

with $\gamma_k \geq \left\| \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \right\|_{\boldsymbol{B}_k}$. We provide an upper bound on the confidence set of the current estimation $\hat{\boldsymbol{\theta}}_k$ in Section 4.5.3. The UCB function is then given by

$$\mathrm{UCB}(a) = \boldsymbol{x}_a^\top \hat{\boldsymbol{\theta}}_k + \gamma_k \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}}. \tag{4.9}$$

4.5.3 Analysis

We start by providing a confidence set bound on the current estimation of our task parameter. We make use of an adapted concentration inequality provided by Abbasi-Yadkori *et al.* (2011) in the following lemma.

Lemma 4.5.1 (Self-normalized bound for vector-valued martingales)**.** *Let τ be a stopping time with respect to a filtration $\{\mathcal{F}_k\}_{k=0}^\infty$ and define $\boldsymbol{\eta}_k = \boldsymbol{D}_k^\top \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \in \mathbb{R}^k$ as sub-Gaussian noise vector. Then, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have*

$$\|\boldsymbol{\eta}_k\|_{\boldsymbol{B}_k^{-1}}^2 \leq \log\left(\frac{\det(\boldsymbol{B}_k)}{\delta^2 \lambda_1^q \lambda_2^p} \right).$$

To emphasize the dimension of the subspace and the residual, in the lemma below, we bound the ratio of determinants in Lemma 4.5.1.

Lemma 4.5.2. *Let $\lambda_1, \lambda_2 > 0$ and $\boldsymbol{B}$ be defined as in eq. (4.8). Then*

$$\log\left(\frac{\det(\boldsymbol{B}_k)}{\det\left(\lambda_1\hat{\boldsymbol{P}}^{\perp} + \lambda_2\hat{\boldsymbol{P}}\right)}\right) \leq S_k^{\lambda_1,\lambda_2} \coloneqq p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right).$$

In the following lemma, we formulate the confidence set bound in our setting.

Lemma 4.5.3. *At round k, and with probability of at least $1-\delta$, the confidence set bound for $\hat{\boldsymbol{\theta}}_k$ is given by*

$$\left\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*\right\|_{\boldsymbol{B}_k} \leq \sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda_2}V + \frac{\lambda_1}{\sqrt{\lambda_2}}W,$$

where $W \coloneqq \left\|\hat{\boldsymbol{P}}^{\perp}(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\|$.

Ideally, we want to show that the confidence set bound is tightened with the knowledge of the shared subspace and the corresponding projection matrix. That can be observed in the regularization terms scaling with $\left\|\hat{\boldsymbol{P}}^{\perp}\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\|$, which is small with high probability due to Assumption 4.3.1. The second regularization based term scales with $\left\|\hat{\boldsymbol{P}}\boldsymbol{\theta}^*\right\|$ and λ_2 and guarantees our problem to be well posed. In addition, the choice of $\lambda_2 \ll \lambda_1$ enforces leveraging our assumption on ρ.

Before establishing an upper bound for the expected transfer regret, we deliver an error estimation on our projection matrices. For this purpose, we use the eigengap $\Delta_\sigma \coloneqq \sigma_p - \sigma_{p+1}$, which is assumed to be positive, where p is the dimension of the low dimensional subspace. A projection $\boldsymbol{P}$ depends on the number p of selected eigenvectors, thus it can be assigned a specific eigengap. The following results shows the benefit of large eigengaps.

Lemma 4.5.4. *Let $\bar{\boldsymbol{\theta}} = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}(i)$ be the empirical mean of L_2 regularized task parameter estimations $\boldsymbol{\theta}(i)$ of true parameters $\boldsymbol{\theta}^*(i) \sim \rho$. Assume that each $\boldsymbol{\theta}(i)$ was estimated after the selection of at least n arms. Let $\hat{\boldsymbol{P}}^{\perp}$ and $\Delta_\sigma > 0$ be the estimation of $\boldsymbol{P}^{\perp}$ and the eigengap of $\boldsymbol{\Sigma}$, respectively. We have*

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\mathbb{E}\left[\left\|\hat{\boldsymbol{P}}^{\perp}\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\|\right]\right] = \mathcal{O}\left(\sqrt{\mathrm{Var}_\rho + b^2\beta_d^2 + \epsilon_{\boldsymbol{\mu}}^2 + c\epsilon_{\boldsymbol{\Sigma}}^2}\right),$$

with $\epsilon_{\boldsymbol{\mu}} = \frac{2\log(2t)}{t} + \sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}$, $\epsilon_{\boldsymbol{\Sigma}}^2 = \frac{C\log(2t)}{t}$, C is an absolute constant, $b = 1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}$, $c = \frac{128pV^2}{\Delta_\sigma^2}$, $\beta_d = \frac{1}{\sqrt{\lambda_{\min}}}\left[\sqrt{d\log\left(1 + \frac{n^2V^2}{d}\right) + 2} + \sqrt{\frac{1}{n}}\right]$ and $\lambda_{\min}$ is the smallest of the minimal eigenvalues of matrices $\boldsymbol{A}_n$.

In what follows, we define

$$Y \coloneqq \mathrm{Var}_\rho + \beta_d^2 \left(1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}\right)^2 + \epsilon_{\boldsymbol{\mu}}^2 + \frac{128p\epsilon_{\boldsymbol{\Sigma}}^2 V^2}{\Delta_\sigma^2}.$$

The mean concentration error is $\epsilon_{\boldsymbol{\mu}}$ that converges to zero for a sufficiently large number of tasks. Besides, $\epsilon_{\boldsymbol{\Sigma}}$ gives us the concentration error bound of the covariance estimated by the true task parameters $\boldsymbol{\theta}^*$ and converges to zero. The bound depends heavily on the eigengap of the true covariance. Larger eigengaps reduce the expected error term and increase the reliability of the projection estimation. For the analysis, we assume $\Delta_\sigma > 0$ for the chosen value of p. By assumption, Var_ρ is relatively small. Thus, the complete term is mostly dominated by β_d^2, which is an upper bound on the mean squared error (MSE) of the ridge estimator from the standard linUCB case. By selecting $\lambda \sim \frac{1}{n}$, the MSE of the ridge estimator converges to the estimator's variance. This variance, in turn, scales with the sub-Gaussian noise term added on the rewards and also depends on the singular values of the respective data covariance matrix $\boldsymbol{D}^\top \boldsymbol{D}$. Ideally, we would prefer non-zero singular values. That implies that the set of context vectors yield information along any dimension, which would minimize the variance of the ridge estimator; Nevertheless, our setting does not guarantee this.

We establish an upper bound on the transfer regret in the following theorem.

Theorem 4.5.5. *Assuming that $\boldsymbol{P}$ and $\boldsymbol{\mu}$ are known, the expected transfer regret of the projected LinUCB algorithm is upper bounded by*

$$\mathcal{R}(n) = \mathcal{O}\left(\sqrt{n}\left(p\log\left(1 + \frac{nV^2}{p}\right) + q\log\left(1 + \frac{n\sqrt{\mathrm{Var}_\rho}}{q}\right)\right)\right).$$

If the assumptions of Lemma 4.5.4 hold, the expected transfer regret is upper bounded by

$$\mathcal{R}(n) = \mathcal{O}\left(\sqrt{n}\left(p\log\left(1 + \frac{nV^2}{p}\right) + q\log\left(1 + \frac{n\sqrt{Y}}{q}\right)\right)\right).$$

Remark 4.5.6. *The case $p = d, q = 0$ yields the expected transfer regret $\mathcal{O}\left(\sqrt{n}d\log\left(1 + \frac{nV^2}{d}\right)\right)$, when no actual meta learning takes place and each task is learnt independently by the LinUCB algorithm.*

The results show that our approach is at most beneficial when p is as low as possible such that Assumption 4.3.1 still holds. In that case, increasing λ_1 in the algorithm reduces the overall regret bound, further supporting the argument that

we made while discussing the confidence set bound in Lemma 4.5.3. By setting $\lambda_1 = \frac{1}{\sqrt{Y}}$, the term $S_n^{\lambda_1,\lambda_2}$ defined in Lemma 4.5.2 changes such that only the p dependent term becomes relevant as Y significantly decreases and in turn $\log\left(1 + \frac{n\sqrt{Y}}{q}\right)$ as well, essentially reducing the effective dimension of the problem to p and indicating that less exploration is required within the q-dimensional subspace.

4.6 Projection Meta Learning with Linear Thompson Sampling

4.6.1 Basics of Linear Thompson Sampling

LinUCB and linear Thompson sampling have the same requirements and assumptions concerning the linear relation between expected rewards and context vectors. Their difference lies in the decision-making process: In the former, the learner maximizes a UCB function by selecting the action at every round, whereas in the latter, it utilizes a Gaussian posterior calculated as $\mathcal{N}(\boldsymbol{\theta}_k, v^2\boldsymbol{A}_k^{-1})$, with $\boldsymbol{\theta}_k$ estimated through solving the regularized least squares as done in LinUCB. From which, the learner then samples a parameter vector $\tilde{\boldsymbol{\theta}}_k$. It then selects the actions as

$$a = \arg\max \boldsymbol{x}_a^\top \tilde{\boldsymbol{\theta}}_k.$$

The posterior is built from the prior of the previous instance given by $\mathcal{N}(\boldsymbol{\theta}_{k-1}, v^2\boldsymbol{A}_{k-1}^{-1})$. This means that at $k = 0$, during the initialization, we have $\boldsymbol{A}_0 = \boldsymbol{I}$. The sampling process reflects the uncertainty of the current estimation $\boldsymbol{\theta}_k$ and directly indicates the exploration behaviour of the learner. A low variance across a specified dimension indicates a high confidence of the current estimation and vice versa. Thus during initialization with $\boldsymbol{A}_0 = \boldsymbol{I}$, there is equal exploration potential along any direction.

4.6.2 Thompson Sampling with Linear Payoffs within an Affine Subspace

Our second proposal is a variation of the linear Thompson sampling: We change the posterior from which $\tilde{\boldsymbol{\theta}}$ is sampled. The mean of the new distribution is the biased regularization solution $\hat{\boldsymbol{\theta}}$ of eq. (4.6), and its covariance matrix is $\boldsymbol{B}^{-1}$. Thus,

$$\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, v^2\boldsymbol{B}^{-1}). \tag{4.10}$$

In eq. (4.10), v is a hyper-parameter that we determine in the analysis. During initialization, we have $\boldsymbol{B}_0 = \lambda_1\hat{\boldsymbol{P}}^\perp + \lambda_2\hat{\boldsymbol{P}}$ and its inverse as covariance for the prior distribution. By choosing $\lambda_1 \gg \lambda_2$ similar to the projected LinUCB setting, we

embed our knowledge of the affine subspace into the prior. That way, the sampling process of $\tilde{\boldsymbol{\theta}}$ incorporates the low variance along the orthogonal subspace.

4.6.3 Analysis

The analysis in inspired by Agrawal and Goyal (2013). First, we define the following two events:

Definition 4.6.1. *The event E_r occurs if*

$$\forall a \in \mathcal{A}_k : \left| \boldsymbol{x}_a^\top \left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \right) \right| \leq l_n \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}},$$

with $l_n \coloneqq \sqrt{2\log\left(\frac{1}{\delta}\right)(d+2)\log(n) + 2K^2}$ and $K \coloneqq \frac{\lambda_2}{\sqrt{\lambda_1}} W + \sqrt{\lambda_1} V$.
The event $E_{\boldsymbol{\theta}}$ occurs if

$$\forall a \in \mathcal{A}_k : \left| \boldsymbol{x}_a^\top \left(\hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k \right) \right| \leq \sqrt{2d + 6\log(n)} v \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}},$$

with $v \coloneqq 4\sqrt{\log\left(\frac{1}{\delta}\right)\frac{d+2}{\alpha}}$ and $\alpha \in (0,1)$.

The event E_r essentially reflects the confidence set bound discussed for the projected LinUCB case. It gives the probability that our current estimation $\hat{\boldsymbol{\theta}}_k$ lies within the bound. The event $E_{\boldsymbol{\theta}}$ is directly linked to the sampling procedure of $\tilde{\boldsymbol{\theta}}_k$. It gives the probability that the reward estimation of the sampled $\tilde{\boldsymbol{\theta}}_k$ is within some limited range of the estimated reward of $\hat{\boldsymbol{\theta}}_k$. Below, we define a filtration, containing all necessary information for the algorithm.

Definition 4.6.2. *We define the filtration $\{\mathcal{F}_k\}_{k\in\{0,\dots,\infty\}}$ with sub-σ-algebras $\mathcal{F}_{k-1}$ at round $k-1$ generated by the current action set and the history up to round $k-1$: $\mathcal{F}_{k-1} = \{\mathcal{A}_k, \mathcal{H}_{k-1}\}$, with the history being recursively defined as:*

$$\mathcal{H}_k = \{\mathcal{A}_k, \hat{\boldsymbol{\theta}}_k, \boldsymbol{B}_k, \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}}, \mathcal{N}(\hat{\boldsymbol{\theta}}_k, \boldsymbol{B}_k^{-1})\} \cup \mathcal{H}_{k-1}.$$

The next lemma states the probability of the events E_r and $E_{\boldsymbol{\theta}}$.

Lemma 4.6.3. *For all $\delta \in (0,1)$, the probability of event E_r is bounded from below as follows: $\Pr(E_r) \geq 1 - \frac{\delta}{n^2}$. Moreover, for all possible filtrations $\mathcal{F}_{k-1}$, the probability of event $E_{\boldsymbol{\theta}}$ is bounded from below as follows: $\Pr(E_{\boldsymbol{\theta}} | \mathcal{F}_{k-1}) \geq 1 - \frac{1}{n^2}$.*

In the following theorem, we establish an upper-bound for the transfer regret of the projected Thompson sampling algorithm

Theorem 4.6.4. *The expected transfer regret of the projected Thompson sampling algorithm verifies*

$$\mathcal{R}(n) = \mathcal{O}\left(\left(d^{\frac{3}{2}}\log(n) + \sqrt{d}\log(n)^2\right)\sqrt{nS_n^{\frac{1}{\sqrt{Y}},\frac{1}{V^2}}}\right).$$

Remark 4.6.5. *With $p = d, q = 0$, meta learning does not take place, i.e., the agent learns each task independently by the linear TS algorithm. As such, the expected transfer regret yields $\mathcal{O}\left(\left(d^2\log(n) + d\log(n)^2\right)\sqrt{n\log\left(1+\frac{nV^2}{d}\right)}\right)$.*

The results shows a the dependency on the dimensions p and q, and the variance related term Y. For a sufficiently small Y, the terms scaling with p would dominate the regret, so we expect greater improvements with decreasing p. The term scaling with q would benefit from the low variance within the respective subspace. As suggested in Agrawal and Goyal (2013), we chose $\alpha = \frac{1}{\log(n)}$ in the proofs.

4.7 Algorithms

The projected LinUCB and projected TS algorithms share many steps. Thus, we unify them and use sub-procedures. We introduce an initialization phase for learning the subspace, as it may only be well-defined after including sufficient task parameters. Enforcing the subspace learning already from the first task might lead to zero-dimensional subspace with $\hat{\boldsymbol{P}} = \mathbf{0}$ that would degrade the overall performance. In the projected LinUCB algorithm, we require the estimation of γ_k taken from Lemma 4.5.2, which in turn requires the value of $W = \left\|\hat{\boldsymbol{P}}^{\perp}\left(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}}\right)\right\|$, which is intractable but we work around this issue by using the simple bound $W \leq 2V$. Since γ_k acts more as exploration scaling factor, we do not lose any benefit from the meta learning as the actual knowledge transfer becomes relevant in the calculations of $\boldsymbol{B}_k$ and $\hat{\boldsymbol{\theta}}_k$

4.8 Numerical Experiments

We test our algorithms experimentally on synthetic data and on real world data taken from the MovieLens data set.

4.8.1 Synthetic Data Experiments

We sampled the context vectors from a zero mean normal distribution with a diagonal covariance matrix whose elements followed a uniform distribution. Following

Algorithm 2: Projected LinUCB/Thompson Sampling

Initialize: $v > 0$, $\delta \in (0,1)$, $\lambda_1 > \lambda_2 > 0$, $\lambda > 0$, $\delta \in (0,1)$;
for $t \in \{1, ..., T\}$ **do**
  Initialize: $\boldsymbol{A}_0 = \lambda \boldsymbol{I}$, $\boldsymbol{b}'_0 = \boldsymbol{0}$;
  Sample new task $\boldsymbol{\theta}^* \sim \rho$;
  if $t < d$ **then**
    $\hat{\boldsymbol{P}} = \boldsymbol{I}$, $\hat{\boldsymbol{P}}^\perp = \boldsymbol{0}$, $\boldsymbol{w} = \boldsymbol{0}$;
  end
  else
    Determine principal components and calculate $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}^\perp$ with $[\boldsymbol{\theta}(i)]_{i \in \{1,...,t\}}$ according to eqs. (4.2) and (4.3) and $\boldsymbol{w} = \frac{1}{t-1} \hat{\boldsymbol{P}}^\perp \sum_{i=1}^{t-1} \boldsymbol{\theta}(i)$;
  end
  Initialize $\boldsymbol{B}_0 = \lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}$, $\boldsymbol{b}_0 = \lambda_1 \hat{\boldsymbol{P}}^\perp \boldsymbol{w}$, $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{B}_0^{-1} \boldsymbol{b}_0$;
  for $k \in \{0, ..., n-1\}$ **do**
    Select arm a_k according to respective arm selection strategy (Algorithm 3 or Algorithm 4);
    Collect immediate reward r_k;
    $\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \boldsymbol{x}_{a_k} \boldsymbol{x}_{a_k}^\top$;
    $\boldsymbol{A}_{k+1} = \boldsymbol{A}_k + \boldsymbol{x}_{a_k} \boldsymbol{x}_{a_k}^\top$;
    $\boldsymbol{b}_{k+1} = \boldsymbol{b}_k + r_k \boldsymbol{x}_{a_k}$;
    $\boldsymbol{b}'_{k+1} = \boldsymbol{b}'_k + r_k \boldsymbol{x}_{a_k}$;
    $\hat{\boldsymbol{\theta}}_{k+1} = \boldsymbol{B}_{k+1}^{-1} \boldsymbol{b}_{k+1}$;
  end
  $\boldsymbol{\theta}(t) = \boldsymbol{A}_n^{-1} \boldsymbol{b}'_n$;
end

Mezzadri (2006), we used a randomly generated orthogonal matrix to define a subspace. We project the randomly generated task parameters onto the subspace and add a multivariate Gaussian noise term in the orthogonal direction to the given subspace to simulate the variance of the task distribution. One drawback of this approach is that it misses the benefits of subspace learning during the first tasks. That is because a subspace with dimension p that ought to be learned requires at least $q = d - p$ data points or task parameters to use the PCA algorithms successfully. Thus, we also implement an initialization phase to prevent subspace learning until learning at least d task parameters. Note that we require at least d tasks as we do not use our knowledge of p. We consider a task as finished after at least $n = 250$ rounds.

Figure 4.2a shows the expected transfer regret for $d = 30$ and $p = 15$, with the oracle and the algorithms of Cella *et al.* (2020) (B-OFUL), Peleg *et al.* (2022) (M-TS)

Algorithm 3: Projected LinUCB Arm Selection Routine

Input: $\hat{\boldsymbol{\theta}}_k$, $\boldsymbol{B}_k$;
$\gamma_k = \log\left(\frac{\det(\boldsymbol{B}_k)}{\delta^2 \lambda_1^q \lambda_2^p}\right) + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W$;
Select arm $a_k = \arg\max_a \mathrm{UCB}(a)$ from equation 4.9;

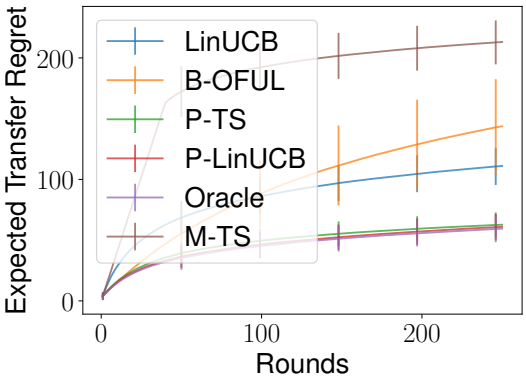
Algorithm 4: Projected TS arm selection routine

Input: $\hat{\boldsymbol{\theta}}_k$, $\boldsymbol{B}_k$, $\alpha \in (0, 1)$;
$v = 4\sqrt{\log\left(\frac{1}{\delta}\right)\frac{d+2}{\alpha}}$;
sample $\tilde{\boldsymbol{\theta}}_k \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_k, v^2 \boldsymbol{B}_k^{-1})$;
Select arm $a_k = \arg\max_a \boldsymbol{x}_a^\top \tilde{\boldsymbol{\theta}}_k$;

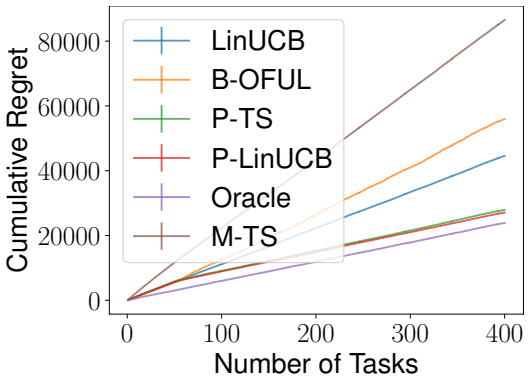
as benchmarks. The projected Thompson sampling (P-TS) approach performs as well as projected LinUCB (P-LinUCB), while the oracle using the true projection and mean is the most efficient one. Our algorithms are significant improvements the other baselines. The superiority of our approach mainly stems from its generality compared to Cella *et al.* (2020). The algorithm provided by Peleg *et al.* (2022) has the worst performance mainly due to the regret contributed by the required forced exploration within a task. Additionally, Figure 4.2b shows the total cumulative regret over tasks, which does not suffer from overlapping error bars as they become negligible. To further emphasize the benefit of exploiting the knowledge on any dimensional low variance, Figure 4.3a shows the total accumulated regret of the projected LinUCB algorithm after T tasks with n rounds of learning each as a function of $q = \mathrm{rank}(\hat{\boldsymbol{P}}^\perp)$. Note that at $q = 0$, the plot shows the total regret using classic LinUCB. As expected, the regret reaches its minimum when q is equal to the rank of the true projection, which is $q = 15$ in this case. Nevertheless even for different values of q, there is a clear benefit over the classic approach. In Figure 4.3b we plot $|\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[W]]^2 / \mathrm{Var}_\rho - 1|$ as a function of number of tasks. As expected the curves for P-LinUCB and P-TS imply that $\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[W]]^2$ is close to Var_ρ. The curve for the B-OFUL algorithm assumes $\boldsymbol{P}^\perp = \boldsymbol{I}$, disregarding the covariance and thus resulting into higher values. Note that lower values imply greater transfer in between sequential tasks as the projection matrix would be well estimated.

4.8.2 Real Data Experiments

We use MovieLens data to test our algorithms in a real-world environment. MovieLens data contains information about over 6000 users that represent the tasks in our setting. Besides, it includes over 3000 movies, which are the arms with their corresponding context vectors. The context vectors are 18-dimensional, each denoting a possible genre. If a movie has a label for a specific genre, the corresponding



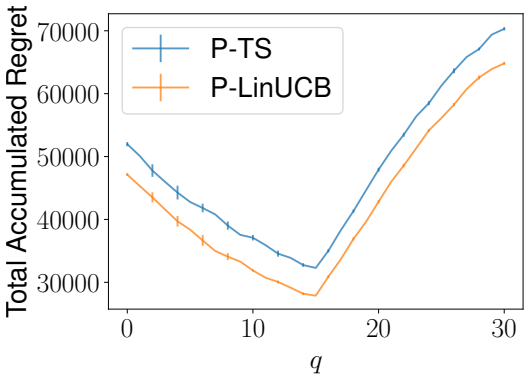
(a) Synthetic data plots of the expected transfer regret as a function of number of rounds.



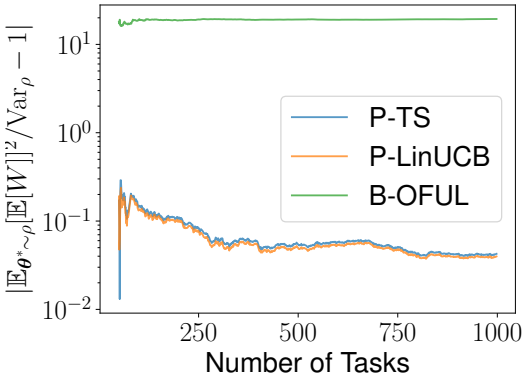
(b) Synthetic data plots of the cumulative regret over the number of tasks.

Figure 4.2: Expected transfer and total cumulative regret plots of the LinUCB and Thompson sampling methods compared to their projection counterparts and additional baselines.

entry for that genre in the context vector is 1. With at most six different genres assigned to a single movie, we normalize the context vectors such that we have $\|\boldsymbol{x}_a\| \leq 1$. Each movie has some available ratings between 1 and 5, given by a user who has watched that movie. Each rating represents a reward for our algorithm. We normalize all such ratings so that $r \in [0, 1]$. We further process the data by grouping the users by their profession or gender and run the algorithm within that set of users. That method stems from the assumption that groups of similar users might share an affine subspace. For every user (task), we run the algorithms for at least $n = 250$ rounds. We do not include the algorithm developed in Peleg *et al.* (2022) as baseline for experimentation using the real data set as it requires contexts from a distribution with an invertible covariance. The reason is that the authors do not use a regularizer on the minimum least squares solution for $\boldsymbol{\theta}$, and thus find the inverse of $\boldsymbol{D}^\top \boldsymbol{D}$ after every finished task. However, in the MovieLens data set, that condition does not hold for many users, making the estimation of $\boldsymbol{\theta}$ ill-posed. In the real data experiments, we observe significant improvements of our models over the baselines in Figure 4.4a showcasing the expected transfer regret, and in Figure 4.4b, showcasing the total cumulative regret over users, which does not suffer from overlapping error bars. A significant point is that we did not perform data preprocessing besides normalizing the rewards and dividing the users into male and female. That would also explain the performance gap to the algorithm of Cella *et al.* (2020), as our assumption is more general and widely applicable.

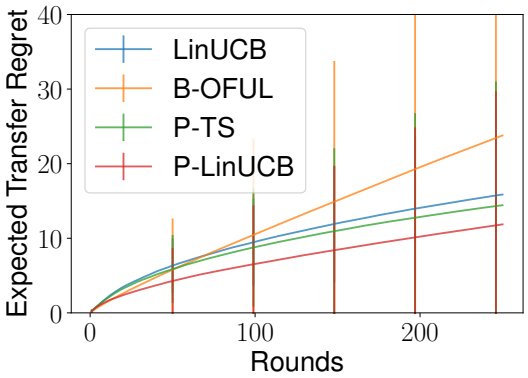


(a) Total accumulated regret after 400 tasks as a function of rank($\hat{\boldsymbol{P}}^{\perp}$).

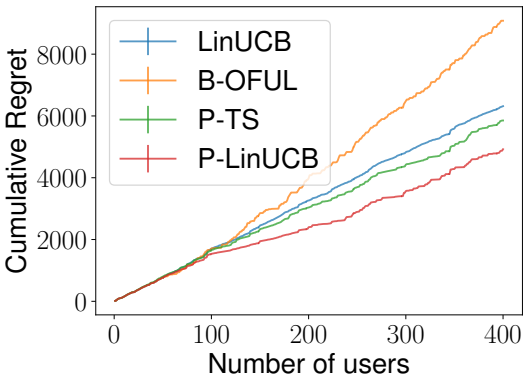


(b) Relative error of $\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}[\mathbb{E}[W]]^2$ as a function of number of tasks in a logarithmic scale.

Figure 4.3



(a) Real data plots of the expected transfer regret as a function of number of rounds.



(b) Real data plots of the cumulative regret over tasks/users.

Figure 4.4: Expected transfer regret and total regret plots of our algorithms and baselines applied to the MovieLens data set. We have included 400 users in the simulations.

4.9 Discussion and Outlook

Our work shows that obtaining knowledge about the underlying subspace structure in a meta learning setting improves sequential task learning. More precisely, assuming a low variance along certain dimensions in the task distribution, we proposed two decision-making policies that exploit the knowledge of the subspace structure for sequential arm selection and significantly improve the performance of widely used algorithms, namely LinUCB and linear Thompson sampling. We provided an improved regret bound that manifests the dependency on the lower dimension, the low variance term, and the eigengap at the considered low dimension. We evaluated our methods numerically through experimentations on synthetic and real-world datasets, confirming their better performance than traditional benchmarks. The results are significant in the real data environments as the rewards do not necessarily follow a linear relation.
Possible extensions of this work include further generalization of our model by learning the variance of the task distribution along all dimensions. Another direction is to generalize our methods to non-linear settings, *i.e.*, when tasks concentrate around a low dimensional manifold.

Chapter 5

Cluster Agnostic Network Lasso Bandits

This chapter, along with Appendix C, is a verbatim copy of Dhouib *et al.* (2025). Author contributions are stated as follows:

Author	**Author Position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Sofien Dhouib	1	50	65	45	45
Steven Bilaj	2	30	10	40	40
Behzad Nourani-Koliji	3	15	25	15	15
Setareh Maghsudi	4	5	0	0	0
Title of paper:		Cluster Agnostic Network Lasso Bandits			
Status in publication process:		Published			

Table 5.1: Chapter 5 author contributions.

Abstract 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 (Su and Khoshgoftaar, 2009), 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 (Li *et al.*, 2010) have been popularly used to formalize and address this exploration-exploitation trade-off.

However, the classical form of contextual bandits (Li *et al.*, 2010; Chu *et al.*, 2011; Abbasi-Yadkori *et al.*, 2011) 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. Second, the partial feedback of the bandit settings exposes the algorithms to having weak estimations and impairing their decision-making ability (Yang *et al.*, 2020b). Consequently, to improve bandit algorithms' performance for large-scale applications, structural assumptions that link the different users are usually integrated within bandit algorithms (Cesa-Bianchi *et al.*, 2013; Gentile *et al.*, 2014; Li *et al.*, 2019; Herbster *et al.*, 2021).

Cesa-Bianchi *et al.* (2013) and Yang *et al.* (2020b) use 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 (Wang *et al.*, 2016). On the other hand, algorithms of online clustering of bandits (Gentile *et al.*, 2014; Li *et al.*, 2019) 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 (Hallac *et al.*, 2015), 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 paper 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 (Tibshirani, 1996) problem under different assumptions (Bastani and Bayati, 2019; Kim and Paik, 2019; Oh *et al.*, 2021; Ariu *et al.*, 2022). They all rely on a previously established compatibility or RE condition (Bühlmann and van de Geer, 2011), 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 (2021) with a Group Lasso regularization (Yuan and Lin, 2006), and to impose a low-rank structure on the task preference vectors in Cella *et al.* (2023). 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.* (2014) 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.* (2019) generalizes CLUB via including merging operations in addition to splitting. In contrast to these approaches, Nguyen and Lauw (2014) groups users via K-means clustering, and Cheng *et al.* (2023) rely on hedonic games for online clustering of bandits. Furthermore, Yang and Toni (2018) make use of community detection techniques on graphs to find user clusters. Gentile *et al.* (2017) 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 Cella and Pontil (2021), 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 Cella *et al.* (2023), where the tasks are assumed to be confined to an arbitrary unknown low-dimensional subspace. That work improves upon Hu *et al.* (2021) 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 (Cesa-Bianchi *et al.*, 2013; Yang and Toni, 2018), which is our case. However, while they assume smoothness as a prior, we assume piecewise constant behavior.

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 (McPherson *et al.*, 2001; Easley *et al.*, 2010). 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 (Dong *et al.*, 2019). Exploiting

this information and integrating them into bandit algorithms can lead to a significant increase in performance Yang *et al.* (2020b). 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 to integrate information from social networks is that, social networks show large *modularity* measures (Newman, 2006; Borge-Holthoefer *et al.*, 2011). 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 (Gentile *et al.*, 2014; Nguyen and Lauw, 2014; Yang and Toni, 2018; Li *et al.*, 2019; Nourani-Koliji *et al.*, 2023; Cheng *et al.*, 2023).

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. Bilaj *et al.* (2024) assume that the tasks are sampled from a common distribution and concentrated around an affine subspace learned 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 Cella *et al.* (2020); Kveton *et al.* (2021); Basu *et al.* (2021), which learns the mean of the distribution under the assumption that the covariance of the prior is known or Peleg *et al.* (2022) which generalizes this assumption and attempts to learn the covariance as well.

5.3 Problem Setting

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 $\boldsymbol{x}(t)$ 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 $\{m(1), \boldsymbol{x}(1), y(1), \cdots, m(t), \boldsymbol{x}(t), y(t), m(t+1)\}$, the received reward $y(t)$ is given by $y(t) = \left\langle \boldsymbol{\theta}_{m(t)}(t), \boldsymbol{x}(t) \right\rangle + \eta(t)$, where $\eta(t)$ is $\mathcal{F}_t-$measurable and $\forall t \geq 1, \forall s \in \mathbb{R}$,

$$\begin{aligned} &\mathbb{E}\left[\eta(t) | \mathcal{F}_{t-1}\right] = 0, \\ &\mathbb{E}\left[\exp(s\eta(t)) | \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{1}{2}\sigma^2 s^2\right). \end{aligned} \tag{5.1}$$

The performance of our policy is assessed by the expected regret over the T interaction rounds for all tasks:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \max_{\tilde{\boldsymbol{x}} \in \mathcal{A}(t)} \left\langle \boldsymbol{\theta}_{m(t)}, \tilde{\boldsymbol{x}} \right\rangle - \left\langle \boldsymbol{\theta}_{m(t)}, \boldsymbol{x}(t) \right\rangle\right]. \tag{5.2}$$

The Optimization problem in eq. (5.4) is an instance of the Network Lasso (Hallac *et al.*, 2015). Several instances of the same type were studied by Jung *et al.* (2018); Jung and Vesselinova (2019); Jung (2020); He *et al.* (2019). 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 (Wang *et al.*, 2016). According to Wang *et al.* (2016), 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 strength of the latter 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 5.3.1 (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 $\|\boldsymbol{x}\| \leq 1 \forall\ \boldsymbol{x} \in \mathcal{A}(t)\ \forall t \geq 1$.*

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

Assumption 5.3.2 (Relaxed symmetry and balanced covariance)**.** *There exists a constant $\nu \geq 1$ such that for all $\boldsymbol{X} \in \mathbb{R}^{K\times d}$, $p(-\boldsymbol{X}) \leq \nu p(\boldsymbol{X})$. Furthermore, there exists $\omega > 0$, such that for any permutation $(a_1, \cdots, a_K)$ of $[K]$, for any $i \in \{2, \cdots, K-1\}$, $\boldsymbol{w} \in \mathbb{R}^d$, we have*

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

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

This assumption was introduced in Oh *et al.* (2021), and has already been used in a multi-task setting by Cella *et al.* (2023). 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 (Oh *et al.*, 2021). The piecewise constant behavior of the graph signal $\boldsymbol{\Theta}$ is formalized in the next assumption.

Assumption 5.3.3 (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 5.3.3 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 Algorithm

Our policy in Algorithm 5 follows a greedy arm selection rule in a multi-task setting, in the same vein as those presented in Oh *et al.* (2021); Cella *et al.* (2023). Indeed, as pointed out in Oh *et al.* (2021), exploration is implicitly incorporated into regularization parameter $\alpha(t)$'s time dependence. It has the following expression

$$\begin{aligned}
\alpha(t) &\coloneqq \frac{\alpha_0\sigma}{t}\sqrt{t + \alpha_1(t) + \alpha_2(t)}, \\
\alpha_1(t) &\coloneqq \sqrt{2\sum_{m\in\mathcal{V}} |\mathcal{T}_m(t)|^2 \log\frac{1}{\delta(t)}}, \\
\alpha_2(t) &\coloneqq 2\max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \log\frac{1}{\delta(t)},
\end{aligned} \tag{5.3}$$

where $\alpha_0 > 0$. The set of time steps a task m has been selected up to time t is denoted by $\mathcal{T}_m(t)$. 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 network lasso optimization problem:

$$\hat{\boldsymbol{\Theta}}(t) = \underset{\tilde{\boldsymbol{\Theta}} \in \mathbb{R}^{|\mathcal{V}| \times d}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \Big(\Big\langle \tilde{\boldsymbol{\theta}}_{m(\tau)}, \boldsymbol{x}(\tau) \Big\rangle - y(\tau) \Big)^2 + \alpha(t) \sum_{(m,n) \in \mathcal{E}} w_{mn} \Big\| \tilde{\boldsymbol{\theta}}_m - \tilde{\boldsymbol{\theta}}_n \Big\|, \quad (5.4)$$

where $\|\cdot\|$ denotes the Euclidean norm for vectors. At each time step the network Lasso problem is solved via the primal-dual algorithm (Jung, 2020).

Algorithm 5: Network Lasso Policy

Input: $T, \alpha_0 > 0, \mathcal{G}, \delta$
Initialization: $\hat{\boldsymbol{\Theta}}(0) = \mathbf{0} \in \mathbb{R}^{|\mathcal{V}| \times d}$
for $t \in \{1, ..., T\}$ **do**
  Draw a user $m(t) \in \mathcal{V}$ uniformly at random
  Observe context set $\mathcal{A}(t)$
  Select $\boldsymbol{x}(t) \in \arg\max_{\tilde{\boldsymbol{x}} \in \mathcal{A}(t)} \Big\langle \hat{\boldsymbol{\theta}}_{m(t-1)}, \tilde{\boldsymbol{x}} \Big\rangle$
  Receive payoff $y(t)$
  Update $\alpha(t)$ via eq. (5.3)
  Update $\hat{\boldsymbol{\Theta}}(t)$ via eq. (5.4)
end

5.5 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 $\boldsymbol{\Theta}$. 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 Notation 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 5.3.3), 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 $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}$, we denote by $\overline{\boldsymbol{Z}}_{\mathcal{P}}$ the signal obtained by setting row vectors of $\boldsymbol{Z}$ 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 Problem eq. (5.4) is equal to $\|\boldsymbol{\Theta}\|_{\mathcal{E}}$. Also, we define the incidence matrix $\boldsymbol{B}_I \subset \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$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 $w_{mn}(\boldsymbol{e}_m - \boldsymbol{e}_n)$, where $\boldsymbol{e}_m$ is the m^{th} canonical basis vector of $\mathbb{R}^{|\mathcal{V}|}$. We define $\boldsymbol{A}_{\mathcal{V}}(t) \coloneqq \operatorname{diag}\left(\boldsymbol{X}_1(t)^\top \boldsymbol{X}_1(t), \ldots, \boldsymbol{X}_{|\mathcal{V}|}(t)^\top \boldsymbol{X}_{|\mathcal{V}|}(t)\right) \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$, and subsequently the empirical multi-task Gram matrix up to time step t is given by $\frac{1}{t}\boldsymbol{A}_{\mathcal{V}}(t)$. 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 5.5.1 (Cluster content constants)**.** *Let $\mathcal{C} \in \mathcal{P}$ be a cluster.*

- *We denote by $\partial_v\mathcal{C}$ 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 $\boldsymbol{B}_{\mathcal{C}}$ the incidence matrix restricted to edges linking vertices of $\mathcal{C}$, its associated Laplacian matrix by $\boldsymbol{L}_{\mathcal{C}} \coloneqq \boldsymbol{B}_{\mathcal{C}}^\top \boldsymbol{B}_{\mathcal{C}}$, and its pseudo-inverse by $\boldsymbol{L}_{\mathcal{C}}^\dagger$. The topological centrality index of node $m \in \mathcal{C}$ w.r.t $\mathcal{C}$ is equal to $(\boldsymbol{L}_{\mathcal{C}}^\dagger)_{mm}^{-1}$. We define the topological centrality index of $\mathcal{C}$ by $c_{\mathcal{G}}(\mathcal{C}) \coloneqq \min_{m\in\mathcal{C}}(\boldsymbol{L}_{\mathcal{C}}^\dagger)_{mm}^{-1}$.*

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 (Cheeger, 1970).

The topological centrality index measures the overall connectedness of a vertex in a network and indicates how robust a node is to edge failures (Ranjan and Zhang, 2013). Also, it can be tied to electricity spreading in a network according to Van Mieghem *et al.* (2017). 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.

Remark 5.5.2. *Both the topological centrality index and inner isoperimetric ratio are key parameters of the cluster structure and the graph. They determine the 'quality' of the given graph. An optimal graph and cluster structure yield many intra-cluster connections and few inter cluster connections i.e. a high topological*

centrality index and low inner isoperimetric ratio for any cluster. This will later be highlighted in the oracle inequality and the regret bound.

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

Definition 5.5.3 (Restricted Eigenvalue (RE) condition and norm)**.** *A PSD matrix* $\boldsymbol{M} \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$ *verifies the RE condition with constants* $\kappa \geq 1$*,* $\psi > 0$ *and* $\phi > 0$ *if*

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

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

$$\mathcal{S} \coloneqq \left\{ \boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}; a_1\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right) \|\boldsymbol{Z}\|_{\partial\mathcal{P}^c} \leq a_2\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right) \left\|\overline{\boldsymbol{Z}}_{\mathcal{P}}\right\|_F \right\},$$

$$a_1(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \coloneqq 1 - \frac{\frac{1}{\alpha_0} + 2\kappa w(\partial\mathcal{P})}{\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}}, \quad a_2(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \coloneqq \frac{1}{\alpha_0} + \sqrt{2}\kappa w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})},$$

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

For the rest of the paper, when we use a_1 and a_2 without arguments, we set $\alpha_0 = \frac{1}{w(\partial\mathcal{P})\psi}$ in order to reduce clutter. 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 need to assume that the graph satisfies $\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} > 2w(\partial\mathcal{P})$, such that we can always find constants $\kappa \geq 1$ and $\psi > 0$ that guarantee $a_1 > 0$

To explain the RE condition, if we had $\mathcal{S} = \mathbb{R}^{|\mathcal{V}|\times d}$ and $\|\cdot\|_{\mathrm{RE}} = \|\cdot\|_F$, then $\boldsymbol{M}$ would be invertible with minimum eigenvalue at least ϕ^2. In comparison, our requirement is weaker since it only needs to hold for signals $\boldsymbol{Z} \in \mathcal{S}$ and for the $\|\cdot\|_{\mathrm{RE}}$ semi-norm. It has the same form as the compatibility assumption for the Lasso problem in (Bühlmann and van de Geer, 2011; Oh *et al.*, 2021) or the restricted strong convexity assumption (Cella *et al.*, 2023).

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

Assumption 5.5.4 (RE condition for the true multi-task Gram matrix)**.** *For* $k \in [K]$*, let* $\boldsymbol{\Sigma}_k \coloneqq \mathbb{E}\left[\boldsymbol{x}_k \boldsymbol{x}_k^\top\right]$ *be the Gram matrix of the* k^{th} *context vector's marginal distribution, let* $\boldsymbol{\Sigma}_{\mathcal{V}}$ *be the true multi-task Gram matrix of the context vector generating distribution, given by*

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

We assume that $\mathbf{\Sigma}_{\mathcal{V}}$ *verifies RE condition (Definition 5.5.3) with some problem dependent constants* $\kappa \in \left[1, \frac{1}{2w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}\right)$, $\psi \in \left(0, \frac{1}{w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\right)$ *and* $\phi > 0$.

This assumption is common to several Lasso-like bandit problems (Oh *et al.*, 2021; Ariu *et al.*, 2022; Cella *et al.*, 2023).

We will later show that it can be transferred to the empirical multi-task Gram matrix.

Remark 5.5.5. *The previous assumption implies the invertibility of the true mean gram matrix of any cluster. For the special case that all clusters are known, i.e. the graph has no inter cluster connections, we can fully exploit this with* $\alpha \to \infty$*, which amounts to the task learning of* $|\mathcal{P}|$ *instead of* $|\mathcal{V}|$ *independent OLS problems, since all users within a cluster share observed context vectors and can be viewed as a singular user. In order for the individual OLS problems to be solvable the true mean gram matrix per cluster needs to be invertible.*

We provide further intuition on the constant ϕ within the RE condition. We can show that ϕ has an upper bound:

Proposition 5.5.6 (On the RE constant ϕ)**.** *Let* $\boldsymbol{M}_i \in \mathbb{R}^{d\times d}$ *be the true multi-task gram matrix of user* i*. Assume* $\kappa \geq 1$*. Then the constant* ϕ *of the RE condition can be upper bounded as:*

$$\phi \leq \sqrt{\lambda_{\min}\left(\frac{\sum_{i\in\mathcal{C}} \boldsymbol{M}_i}{|\mathcal{C}|}\right)},$$

where $\lambda_{\min}(\cdot)$ *yields the minimum eigenvalue of a given matrix.*

Since the true multi-task gram matrix per cluster is always invertible, we always have a non-null minimal eigenvalue.

Remark 5.5.7. *The minimal eigenvalue in Proposition 5.5.6 could be further bounded using the trace of the covariances i.e. the sum of all the eigenvalues over the dimension. This would result into an upper bound of* $\phi^2 \leq \frac{1}{d}$.

5.5.2 Oracle Inequality

This section is dedicated to provide a bound on the estimation error of the Network Lasso problem given in eq. (5.4) at a particular step t of Algorithm 5.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 5.5.8 (Oracle inequality)**.** *Assume that the RE assumption holds for the empirical multi-task Gram matrix* $\frac{1}{t}\boldsymbol{A}_{\mathcal{V}}(t)$ *with constants* $\kappa \in \left[1, \frac{1}{2w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}\right)$, $\psi \in \left(0, \frac{1}{w(\partial\mathcal{P})} \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\right)$ *and* $\phi > 0$*. Suppose that* $\max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \leq bt$ *for some* $b > 0$ *and* $\alpha_0 \geq \frac{1}{\psi w(\partial\mathcal{P})}$*. Then, with a probability at least* $1 - \delta(t)$*, we have*

$$\left\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}(t)\right\|_F \leq 2\frac{\sigma\alpha_0}{\phi^2\sqrt{t}} f(\mathcal{G}, \mathbf{\Theta}, \alpha_0) \sqrt{1 + 2b\sqrt{|\mathcal{V}| \log \frac{1}{\delta(t)}} + 2b \log \frac{1}{\delta(t)}},$$

where

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

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 sub-Gaussian case (Hsu *et al.*, 2012, Theorem 2.1). Compared to the bound of Jung (2020, Theorem 1), we bound a norm of the estimation error rather than just the total variation semi-norm. Besides, due to the expressions of $a_1(\mathbf{\Theta}, \mathcal{G}, \alpha_0)$ and $a_2(\mathbf{\Theta}, \mathcal{G}, \alpha_0)$, 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 and ψ, thus it is most beneficial if κ is close to 1 and ψ close to zero.

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 Oh *et al.* (2021); Cella *et al.* (2023). 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 Assumption 5.3.1, 5.3.2 and 5.5.4.

Theorem 5.5.9 (RE condition holding for the empirical multi-task Gram matrix)**.** *Under assumptions 5.3.2 and 5.5.4, let* $t \geq 1$*, and let* κ, ϕ *be the constants from Assumption 5.5.4. Assume that* $\max_{m \in \mathcal{V}} |\mathcal{T}_m(t)| \leq bt$*. Then, for any* $\gamma \in \left(0, \left(1 + \frac{a_2}{a_1}\right)^{-2}\right)$*, the empirical multi-task Gram matrix* $\frac{1}{t}\boldsymbol{A}_{\mathcal{V}}(t)$ *verifies the RE condition with constants* κ*,* ψ *and* $\hat{\phi}$*, where*

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

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

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 a similar approach as in Oh *et al.* (2021); Cella *et al.* (2023); we prove that the RE condition transfers from the true multi-task Gram matrix to its adapted counterpart $\boldsymbol{V}_{\mathcal{V}}(t)$, defined as follows:

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

where

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

This transfer relies on the work of Oh *et al.* (2021, lemma 10). The other step of the proof is showing that the empirical multi-task Gram matrix and $\boldsymbol{V}_{\mathcal{V}}(t)$ 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 5.5.3). 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 5.5.8 and Theorem 5.5.9, these two conditions are ensured.

Theorem 5.5.10. *Let the mean horizon per node be* $\overline{T} = \frac{T}{|\mathcal{V}|}$. *Under assumptions 5.3.1 to 5.3.3 and 5.5.4, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

$$\mathcal{R}(\overline{T}) \leq \mathcal{O}\Bigg(\frac{\alpha_0 \nu \omega f(\mathcal{G}, \mathbf{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\phi^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right) + \frac{1}{A}\log(d|\mathcal{V}|) + \sqrt{|\mathcal{V}|}\Bigg),$$

with

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

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.* (2021). It has a logarithmic dependence in the graph size and in the dimension, which is a characteristic of regret bound of the "lasso type". Also noteworthy is that the regret grows explicitly with $\log(d)$ only in the time-independent term, making our policy useful in high-dimensional settings. Though from Proposition 5.5.6 we can expect an implicit dependency on the dimension in the RE constant ϕ. Specifically, the lower bound on ϕ is an open problem that appears unsolved in other lasso based works such as Oh *et al.* (2021); Cella *et al.* (2023).

Both the regret bound and the oracle inequality presented in Theorem 5.5.8 hold only for the set of graphs that at least satisfy the condition $\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} > 2w(\partial\mathcal{P})$ and even though our results hold for a large set of graphs, the individual role of graph-related constants, encapsulated in $f(\mathcal{G}, \mathbf{\Theta}, \alpha_0)$, is not obvious. By further restricting the set of graphs, we are able to provide a simplified bound

Corollary 5.5.11. *Assume* $\frac{w(\partial\mathcal{P})(\psi+2\kappa)}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \leq \Omega$, *with some positive constant* $\Omega < 1$, *then under assumptions 5.3.1 to 5.3.3 and 5.5.4, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

$$\mathcal{R}(\bar{T}) = \mathcal{O}\Bigg(\frac{1}{\phi^2(1-\Omega)} \frac{w(\partial\mathcal{P}) \max_{\mathcal{C}\in\mathcal{P}} \iota_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \\ \times \sqrt{\bar{T}} \left(\sqrt{|\mathcal{V}|} + \sqrt{\log\big(\bar{T}|\mathcal{V}|\big)} + \sqrt[4]{|\mathcal{V}| \log\left(\bar{T}|\mathcal{V}|\right)} \right) \\ + \frac{w(\partial\mathcal{P})^2 \max\limits_{\mathcal{C}\in\mathcal{P}} \iota_{\mathcal{G}}(\mathcal{C})}{(1-\Omega)^2 \min_{\mathcal{C}\in\mathcal{P}} (\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))} \log(d|\mathcal{V}|) + \sqrt{|\mathcal{V}|} \Bigg)$$

The simplified bound in Corollary 5.5.11 exhibits the typical multi-task learning dependency $\sqrt{\overline{T}|\mathcal{V}|}$ rather than the independent task learning case $|\mathcal{V}|\sqrt{\overline{T}}$ and highlights the role of graph related properties such as the total weight of the boundary, the maximal inner isoperimetric ratio and the minimal topological centrality index. Furthermore with Ω we can see the influence on the regret bound, when $w(\partial\mathcal{P})$ changes relative to $\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}$. We present two additional specific instances: a graph with no boundary (i.e., $w(\partial\mathcal{P}) = 0$), and a graph where full connectivity is achieved within each individual cluster.

Corollary 5.5.12. *Assume $w(\partial\mathcal{P}) = 0$, then under assumptions 5.3.1 to 5.3.3 and 5.5.4, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

$$\mathcal{R}(\bar{T}) = \mathcal{O}\Bigg(\frac{1}{\phi^2} \sqrt{\bar{T}} \left(\sqrt{|\mathcal{V}|} + \sqrt{\log\big(\bar{T}|\mathcal{V}|\big)} + \sqrt[4]{|\mathcal{V}| \log\left(\bar{T}|\mathcal{V}|\right)} \right) \\ + \frac{1}{\min_{\mathcal{C}\in\mathcal{P}} (\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))} \log(d|\mathcal{V}|) + \sqrt{|\mathcal{V}|} \Bigg)$$

Corollary 5.5.13. *Assume $\frac{w(\partial\mathcal{P})(\psi+2\kappa)}{\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} \leq \Omega$, with some positive constant $\Omega < 1$ and assume any two nodes within the same cluster are connected through an edge in the given graph, then under assumptions 5.3.1 to 5.3.3 and 5.5.4, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

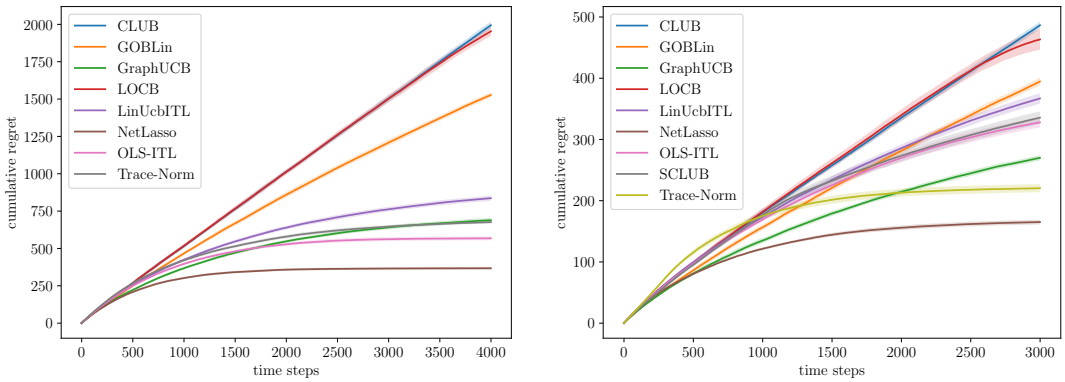
$$\mathcal{R}(\bar{T}) = \mathcal{O}\Bigg(\frac{1}{\phi^2(1-\Omega)}\frac{w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\iota_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{|\mathcal{C}|}} \\ \times\sqrt{\bar{T}}\left(\sqrt{|\mathcal{V}|}+\sqrt{\log\big(\bar{T}|\mathcal{V}|\big)}+\sqrt[4]{|\mathcal{V}|\log\left(\bar{T}|\mathcal{V}|\right)}\right) \\ +\frac{w(\partial\mathcal{P})^2\max\limits_{\mathcal{C}\in\mathcal{P}}\iota_{\mathcal{G}}(\mathcal{C})}{(1-\Omega^2)\min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|}\log(d|\mathcal{V}|)+\sqrt{|\mathcal{V}|}\Bigg)$$

5.6 Experiments

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 (Cesa-Bianchi *et al.*, 2013) and GraphUCB (Yang *et al.*, 2020b) that use the Laplacian to smooth the preference vectors. On the other hand, we compare to clustering of bandits baselines, namely CLUB (Gentile *et al.*, 2014), SCLUB (Li *et al.*, 2019), OLS-ITL (Bastani *et al.*, 2021) and LOCB (Ban and He, 2021). 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 (Cella *et al.*, 2023), which is relevant when the number of clusters is smaller than d. Indeed, the cluster structure of $\mathbf{\Theta}$ can be mathematically written as $\mathbf{\Theta} = \sum_{C\in\mathcal{P}} \mathbf{1}_C \boldsymbol{\theta}_C^\top$, where $\mathbf{1}_C$ is the indicator vector of cluster C (coordinates equal to 1 on the nodes belonging to C and zeros elsewhere) and $\boldsymbol{\theta}_C$ is the true vector of every node in C. The range of $\mathbf{\Theta}$ is equal to the span of $\mathbf{1}_C; C \in \mathcal{P}$, implying that its rank is at most equal to $\min(d, |\mathcal{P}|)$. It will then satisfy the low-rank assumption for $|\mathcal{P}| < d$. 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.

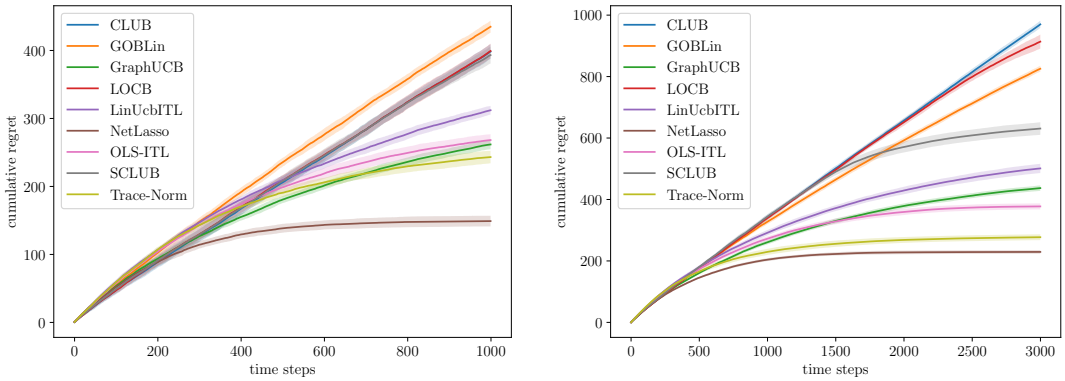
Experimentally, we found that normalizing the weights as $w_{mn} = (\deg(m)\deg(n))^{-\frac{1}{2}}$, where $\deg(m)$ denotes the degree of node m, yields significantly better results. Indeed, such a normalization makes the algorithm focus more on edges between low-degree nodes, which improves the propagation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

Our results clearly demonstrate an improvement compared to the other baselines. 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 between nodes.



(a) $|\mathcal{V}| = 200, |\mathcal{P}| = 25, d = 10, p = 0.5, q = 0.05$

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



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

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

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.

5.7 Conclusion and Future Perspectives

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 (Ariu *et al.*, 2022). Another possible extension would be to use regularization with higher order total variation terms that impose a piecewise polynomial signal on a graph, as explained for scalar signals in Wang *et al.* (2016); Ortelli and van de Geer (2019).

Chapter 6

Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards

This chapter, along with Appendix D, is a verbatim copy of Nourani-Koliji *et al.* (2023). Author contributions are stated as follows:

Author	**Author Position**	**Scientific ideas %**	**Data generation %**	**Analysis & interpretation %**	**Paper writing %**
Behzad Nourani-Koliji	1	70	90	25	65
Steven Bilaj	2	20	10	50	25
Amir Rezaei Balef	3	5	0	25	10
Setareh Maghsudi	4	5	0	0	0
Title of paper:		Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards			
Status in publication process:		Published			

Table 6.1: Chapter 6 author contributions.

Abstract 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.

6.1 Introduction

Multi-armed bandit (MAB) Robbins (1952) 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 Maghsudi and Hossain (2016), recommender systems Li *et al.* (2010), and clinical trials Aziz *et al.* (2021).

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 Auer *et al.* (2002a) 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, Chen *et al.* (2016), Wang and Chen (2017), and Kveton *et al.* (2015) assume that some arms may probabilistically be triggered based on the outcome of other arms. In Lattimore *et al.* (2016), prior knowledge about the causal structure that affects the rewards is available. The authors in Nourani-Koliji *et al.* (2022) 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 apriori knowledge concerning the structural relationships as it can learn the structure from the streaming 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 Auer *et al.* (2002b). 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) Zhou *et al.* (2020) and the *dynamic case* (smoothly changing) Chen *et al.* (2021) Trovo *et al.* (2020). 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* Zhou *et al.* (2020). 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 Besbes *et al.* (2014). In this paper, we focus on the switching case, also referred to as *piecewise* stationary bandit model Besson and Kaufmann (2019). 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 Shen *et al.* (2017). 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 streaming 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* Giannakis *et al.* (2018).

In general, there are two main approaches to follow the piecewise stationary behaviour of base arms distributions; *passively adaptive* approach Garivier and Moulines (2011) and *actively adaptive* approach Besson and Kaufmann (2019) Cao *et al.* (2019) Hartland *et al.* (2007) Liu *et al.* (2018b) Zhou *et al.* (2020) Cheng and Maghsudi (2023). 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 Besson and Kaufmann (2019). Some studies show the superior performance of actively adaptive approaches Mellor and Shapiro (2013). 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 Newman (2006) Borge-Holthoefer *et al.* (2011). Secondly, in some real-world scenarios changes within a network are not completely independent, but they are rather the result of the local spread of a change-seed within the network structure through mechanisms such as contagion Easley *et al.* (2010), social influence Easley *et al.* (2010), or diffusion Thanou *et al.* (2017), e.g. media-based marketing campaigns, or rumor diffusion over social networks Shafipour *et al.* (2021). 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 6.2**, we introduce the piecewise stationary combinatorial semi-bandit problem with causally related rewards. In **Section 6.3**, we develop our decision-making policy, namely, PS-SEM-UCB-Gr. **Section 6.4** presents the theoretical analysis of the regret performance of PS-SEM-UCB-Gr. **Section 6.5** includes the numerical experiments. **Section 6.6** concludes the paper with some suggestions for future works.

6.2 Problem Formulation

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 $\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\boldsymbol{x}_t)$ to denote the agent's expected reward from the decision vector $\boldsymbol{x}_t$ given $\boldsymbol{\mu}_t$ and $\mathcal{A}_t$. Consequently, we characterize a PSCSB with the tuple $(\mathcal{K}, \mathcal{D}, \mathcal{T}, \{\theta_{k,t}\}_{k \in \mathcal{K}, t \in \mathcal{T}}, \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\boldsymbol{x}_t))$. Vector of base arms' instantaneous rewards at time t is represented by $\boldsymbol{b}_t = [b_{1,t}, \ldots, b_{K,t}] \in [0, 1]^K$ 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 $[\zeta] = \{1, \ldots, \zeta\}$. 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

$$N_g = 1 + \sum_{t=1}^{T-1} \mathbb{1}\left\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\right\}. \tag{6.1}$$

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* $\boldsymbol{x}_t = [x_{1,t}, \ldots, x_{K,t}] \in \{0,1\}^K$. 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 $\mathcal{X} = \left\{\boldsymbol{x} \mid \boldsymbol{x} \in \{0,1\}^K \wedge \|\boldsymbol{x}\|_0 \leq m\right\}$ 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), $\mathcal{A}_t = (\mathcal{V}, \mathcal{E}_t, \boldsymbol{W}_t)$. $\mathcal{V}$ represents the set of K vertices, i.e., $|\mathcal{V}| = K$, $\mathcal{E}_t$ and $\boldsymbol{W}_t$ 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 $\boldsymbol{W}_t$ changes only in the elements $\boldsymbol{W}_t[i,j], \forall i, j \in \mathcal{K}, \forall t \in \mathcal{T}$, as far as the underlying graph structure remains a DAG without self-loop, i.e., $\boldsymbol{W}_t[i,i] = 0, \forall i \in \mathcal{K}, \forall t \in \mathcal{T}$. $N_{\boldsymbol{W}}$ represents the number of graph stationary segments. An error-free piecewise static Structural Equation Model (SEM) Giannakis *et al.* (2018) is used to model the generation of reward in the environment. At each time t, $\boldsymbol{z}_t = [z_{1,t}, \ldots, z_{K,t}]$ is used to represent the exogenous input vector while $\boldsymbol{y}_t = [y_{1,t}, \ldots, y_{K,t}]$ denotes the endogenous output vector of the SEM. We write,

$$\boldsymbol{z}_t = \text{diag}(\boldsymbol{b}_t)\boldsymbol{x}_t, \tag{6.2}$$

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

$$y_{k,t} = \sum_{j=1}^{K} \boldsymbol{W}_t[k,j] y_{j,t} + z_{k,t}, \quad \forall k \in \mathcal{K}, \tag{6.3}$$

At each time t, the endogenous output $y_{k,t}$ represents the *overall reward* of base arm $k \in \mathcal{K}$. The element $\boldsymbol{W}[k,j]$ 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 $\boldsymbol{W}_t, \forall t \in \mathcal{T}$ are unknown a priori and $\boldsymbol{W}_t[i,j] \geq 0, \forall i,j \in \mathcal{K}, \forall t \in \mathcal{T}$. The matrix form of eq. (6.3) at time t is given as

$$\boldsymbol{y}_t = \boldsymbol{W}_t \boldsymbol{y}_t + \boldsymbol{z}_t. \tag{6.4}$$

As a result, we write $\boldsymbol{y}_t = (\boldsymbol{I} - \boldsymbol{W}_t)^{-1}\text{diag}(\boldsymbol{b}_t)\boldsymbol{x}_t$ by solving eq. (6.4) for $\boldsymbol{y}_t$, where $\boldsymbol{I}$ is the identity matrix. We assume that the agent is able to observe both the instantaneous semi-bandit feedback vector $\boldsymbol{z}_t$ and the overall reward feedback vector $\boldsymbol{y}_t$. The *payoff* received by the agent upon choosing the decision vector $\boldsymbol{x}_t$ is defined as

$$r_t(\boldsymbol{x}_t) = \boldsymbol{c}^\top \boldsymbol{y}_t = \boldsymbol{c}^\top (\boldsymbol{I} - \boldsymbol{W}_t)^{-1} \,\text{diag}(\boldsymbol{b}_t)\boldsymbol{x}_t, \tag{6.5}$$

where $\boldsymbol{c} = [c_1, \ldots, c_K] \in \{0,1\}^K$ 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 $\boldsymbol{W}_t$ is nilpotent. This property guarantees that the matrix $(\boldsymbol{I} - \boldsymbol{W}_t)$ is invertible. Given a decision vector $\boldsymbol{x}_t \in \mathcal{X}$, the expected payoff at time t is calculated as

$$\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\boldsymbol{x}_t) = \mathbb{E}\left[r_t(\boldsymbol{X}) | \boldsymbol{X} = \boldsymbol{x}_t\right], \tag{6.6}$$

where the expectation concerns the randomness in the reward generating process. We denote by $\boldsymbol{x}_t^* = \underset{\boldsymbol{x} \in \mathcal{X}}{\text{argmax}} \psi_{\boldsymbol{\mu}_t, \boldsymbol{A}_t}(\boldsymbol{x})$ the decision vector with maximum expected reward at time t. The agent minimizes the cumulative piecewise stationary regret defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} (\psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\boldsymbol{x}_t^*) - \psi_{\boldsymbol{\mu}_t, \mathcal{A}_t}(\boldsymbol{x}_t))\right]. \tag{6.7}$$

6.3 The Learning Algorithm

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.

6.3.1 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 Valko *et al.* (2014), 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 Gentile *et al.* (2014), it is suggested that the nodes of the graph can be clustered according to some apriori unknown clustering and the arms within the same cluster exhibit similar behaviours. Also, in Yang *et al.* (2020b), 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 Gentile *et al.* (2014) Li *et al.* (2016). 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.

6.3.2 Piece-wise Static Graph Learning.

Considering the required knowledge of $\boldsymbol{W}_t$ in finding the optimal decision vector, we propose an online graph learning framework that uses the collected feedback $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$ 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 $\boldsymbol{Z}_t$ and $\boldsymbol{Y}_t$, hence, $\boldsymbol{Y}_t = \boldsymbol{W}_t\boldsymbol{Y}_t + \boldsymbol{Z}_t$. We use the collected feedback history, $\boldsymbol{Y}_t$ and $\boldsymbol{Z}_t$, as the input to a parametric graph learning algorithm for a static SEM Giannakis *et al.* (2018). Formally, the adjacency matrix at time t is the solution to the following optimization problem:

$$\begin{aligned} \hat{\boldsymbol{W}}_t = \underset{\boldsymbol{W}\in\mathbb{R}^{K\times K}}{\operatorname{argmin}} \quad & \|\boldsymbol{Y}_t - \boldsymbol{W}\boldsymbol{Y}_t - \boldsymbol{Z}_t\|_{\mathrm{F}}^2 + \lambda_1 \|\boldsymbol{W}\|_1 \\ \text{s.t.} \quad & \boldsymbol{W}[k,k] = 0, \forall k \in \mathcal{K} \end{aligned} \tag{6.8}$$

Algorithm 6: Graph Learning Data Generation (GLDG)

init: $\boldsymbol{Y}_0 = []$, $\boldsymbol{Z}_0 = []$.
for $t' = 1, 2, \ldots, K$ **do**
  $\boldsymbol{x}_t := \textbf{init}[:, t']$
  Play $\mathcal{I}_t$, receive reward $r(\boldsymbol{x}_t)$, $s_{I_t, n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$
  for ***all*** $I_t \in \mathcal{I}_t$ **do**
    update: $\hat{\mu}_{I_t,t}$ using eq. (6.9), $n_{I_t,t}$ using eq. (6.10)
    if $\boldsymbol{GLR}(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **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
  end
  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
  end
  $[\boldsymbol{Y}_t] \leftarrow [\boldsymbol{Y}_{t-1}, \boldsymbol{y}_t]$, $[\boldsymbol{Z}_t] \leftarrow [\boldsymbol{Z}_{t-1}, \boldsymbol{z}_t]$, $t \leftarrow t + 1$
end
Solve eq. (6.8) to get $\hat{\boldsymbol{W}}_{t-1}$
flag $= 0$

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 $\hat{\boldsymbol{W}}_t$. We use the notation $\hat{\boldsymbol{W}}^{(i)}$ 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 $\lambda_2 \left\| \hat{\boldsymbol{W}}^{(i+1)} - \hat{\boldsymbol{W}}^{(i)} \right\|_1$ in eq. (6.8) and have a form of the optimization problem in eq. (6.8) 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.

6.3.3 The 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 Bazerque *et al.* (2013) Nourani-Koliji *et al.* (2022). 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 Nourani-Koliji *et al.* (2022), Section 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 Line 16 of Algorithm 7. The forced uniform exploration guarantees that the GLR change-point detectors receive sufficient 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

$$\hat{\mu}_{k,t_1} = \frac{\sum_{t=\tau_k+1}^{t_1} z_{k,t}}{n_{k,t_1}}, \tag{6.9}$$

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,

$$n_{k,t_1} = \sum_{t=\tau_k+1}^{t_1} x_{k,t}. \tag{6.10}$$

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 5, otherwise the agent plays the next super arm according to the result of the combinatorial optimization in Line 9. 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 $\boldsymbol{U}_t = [U_{1,t}, \ldots, U_{K,t}]$. Therefore, the combinatorial optimization for

Algorithm 7: PS-SEM-UCB-Gr: Piecewise Stationary - Structural Equation Model - Upper Confidence Bound - Group Restart

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 $\Omega \neq \varnothing$ **then**
    Pick $a \in \Omega$, Randomly choose $\mathcal{I}_t$ with $a \in \mathcal{I}_t$
    Remove a from Ω
  end
  else
    Solve eq. (6.11) for $\boldsymbol{x}_t$
  end
  Play $\mathcal{I}_t$, receive reward $r(\boldsymbol{x}_t)$, $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}, \forall I_t \in \mathcal{I}_t$
  for ***all*** $I_t \in \mathcal{I}_t$ **do**
    update: $\hat{\mu}_{I_t,t}$ using eq. (6.9), $n_{I_t,t}$ using eq. (6.10)
    if $\boldsymbol{GLR}(s_{I_t,1}, \ldots, s_{I_t,n_{I_t,t}}; \delta) = 1$ **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
  end
  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
  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
  end
  $[\boldsymbol{Y}_t] \leftarrow [\boldsymbol{Y}_{t-1}, \boldsymbol{y}_t]$, $[\boldsymbol{Z}_t] \leftarrow [\boldsymbol{Z}_{t-1}, \boldsymbol{z}_t]$
  if $\left\| \boldsymbol{y}_t - \hat{\boldsymbol{W}}_{t-1}\boldsymbol{y}_t - \boldsymbol{z}_t \right\|_2^2 > \epsilon$ **then**
    $flag = 1$, $\boldsymbol{Y}_t = []$, $\boldsymbol{Z}_t = []$
  end
  else
    Solve eq. (6.8) to get $\hat{\boldsymbol{W}}_t$
  end
  $t \leftarrow t + 1$
end

finding the best decision vector yields

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}^\top (\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1} \operatorname{diag}(\boldsymbol{U}_{t-1})\boldsymbol{x} \tag{6.11}$$

Let $\boldsymbol{M}^\top = \boldsymbol{c}^\top (\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1} \operatorname{diag}(\boldsymbol{U}_{t-1})$. The elements of $\hat{\boldsymbol{W}}_{t-1}$, $\boldsymbol{c}$, and $\boldsymbol{U}_{t-1}$ are non-negative, then the optimization problem eq. (6.11) can be solved by finding a subset of elements in $\boldsymbol{M}$ such that $\boldsymbol{x} \in \mathcal{X}$. Therefore, it is solvable by using an efficient sorting algorithm that ranks the elements of $\boldsymbol{M}$. Consequently, the agent plays $\boldsymbol{x}_t$, collects the reward in Line 10 according to eq. (6.5), and updates the vectors $\hat{\boldsymbol{\mu}}_t$ and $\boldsymbol{n}_t$ in Line 12. The notation $s_{I_t,n_{I_t,t}} \leftarrow z_{I_t,t}$ in Line 10 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 Besson and Kaufmann (2019) defined as

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

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 Besson and Kaufmann (2019). However, in all our numerical experiments, we follow practical considerations in Besson and Kaufmann (2019), 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 14. The algorithm updates the UCB indices in Line 20. In Line 22, the graph-change detection mechanism uses two vectors of $\boldsymbol{z}_t$ and $\boldsymbol{y}_t$ to test the validity of the last estimate of the graph adjacency matrix, $\hat{\boldsymbol{W}}_{t-1}$. 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 23, we have flag $= 1$, and the previously collected sets of feedback in $\boldsymbol{Z}_t$ and $\boldsymbol{y}_t$ are dropped. Parameter ϵ represents the error we accept in the graph estimation process. In this paper, we take $\epsilon = 0$. However, assuming $\epsilon \neq 0$, the effects of ϵ should be considered in the regret analysis. In case the collected feedback vectors $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$ satisfy the SEM formulation for $\hat{\boldsymbol{W}}_{t-1}$, in Line 21, PS-SEM-UCB-Gr uses the newly updated matrices $\boldsymbol{y}_t$ and $\boldsymbol{Z}_t$, 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.

6.4 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 $\boldsymbol{x}^*$ and an arbitrary decision vector $\boldsymbol{x}$: $\Delta_t(\boldsymbol{x}) = \psi_t(\boldsymbol{x}_t^*) - \psi_t(\boldsymbol{x})$, where $\psi_t(\boldsymbol{x})$ is the mean reward of $\boldsymbol{x}$, with only subscript t used to parameterize it for better readability. The largest gap is denoted as $\Delta_{\max} = \max_t \max_{\boldsymbol{x}:\psi_t(\boldsymbol{x})<\psi_t(\boldsymbol{x}_t^*)} \Delta_t(\boldsymbol{x})$, and the smallest $\Delta_{\min} = \min_t \min_{\boldsymbol{x}:\psi_t(\boldsymbol{x})<\psi_t(\boldsymbol{x}_t^*)} \Delta_t(\boldsymbol{x})$. 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 Nourani-Koliji *et al.* (2022), as our result does not scale with the number of layers in the causal graph but only with the total number of arms;

Lemma 6.4.1. *Let* $\omega_t^T = \boldsymbol{c}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1}\operatorname{diag}(\boldsymbol{x}_{t+1})$ *and* $\omega_{\max} = \max_t \max_k \omega_{k,t}, k \in \mathcal{K}$. *In the stationary case* ($N_\Theta = 1 \wedge N_{\boldsymbol{W}} = 1$) *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 6.4.2 states a bound on the regret in the non-stationary case of our proposed decision-making policy for any grouping of arms.

Theorem 6.4.2. *Let* $\omega_t^T = \boldsymbol{c}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1}\operatorname{diag}(\boldsymbol{x}_{t+1})$ *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:*

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

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

Proof. See Appendix D.1. □

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 6.4.3. *We rewrite the regret upper bound in Theorem 6.4.2 as* $\mathcal{R}(T) \leq \Sigma_{g\in G}[C_1 N_g K_g + C_2 N_g] + C_3$ *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* $\mathfrak{K}_\nu$ *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* $\Delta R(\nu) \leq C_1 \Sigma_{g\in G} K_g \mathbb{1}\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\} + C_2 \Sigma_{g\in G} \mathbb{1}\{\exists k \in g \text{ s.t. } \theta_{k,t} \neq \theta_{k,t+1}\}$. *Consequently, we have the followings;*

- *In the case of Local restart, we have* $\Delta R(\nu) \leq C_1 \mathfrak{K}_\nu + C_2 \mathfrak{K}_\nu$.
- *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* $\mathfrak{K}_\nu$*, for Group restart, we have* $\Delta R(\nu) \leq C_1 \mathfrak{K}_\nu + C_2 \eta_\nu$.
- *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* $\Delta R(\nu) \leq C_1(\mathfrak{K}_\nu + s) + C_2 \eta_\nu$.

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 $\mathfrak{K}_\nu$ 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 6.4.4. *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,*

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

Our regret bound shows an improvement in comparison to the result of Zhou *et al.* (2020) 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., $N_{\boldsymbol{W}} = 1$.

6.5 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** Huyuk and Tekin (2019) is a Thompson sampling-based algorithm for stationary environments; **GLR-CUCB** Zhou *et al.* (2020) 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** Chen *et al.* (2021) 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 eq. (6.11); 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 $\boldsymbol{z}_t$ at each time t. We implemented the exact optimization oracles for *CTS*, *GLR-CUCB*, *GLR-CUCB-Gr*, *GLR-CUCB-Lo*, and *CUCB-SW*.

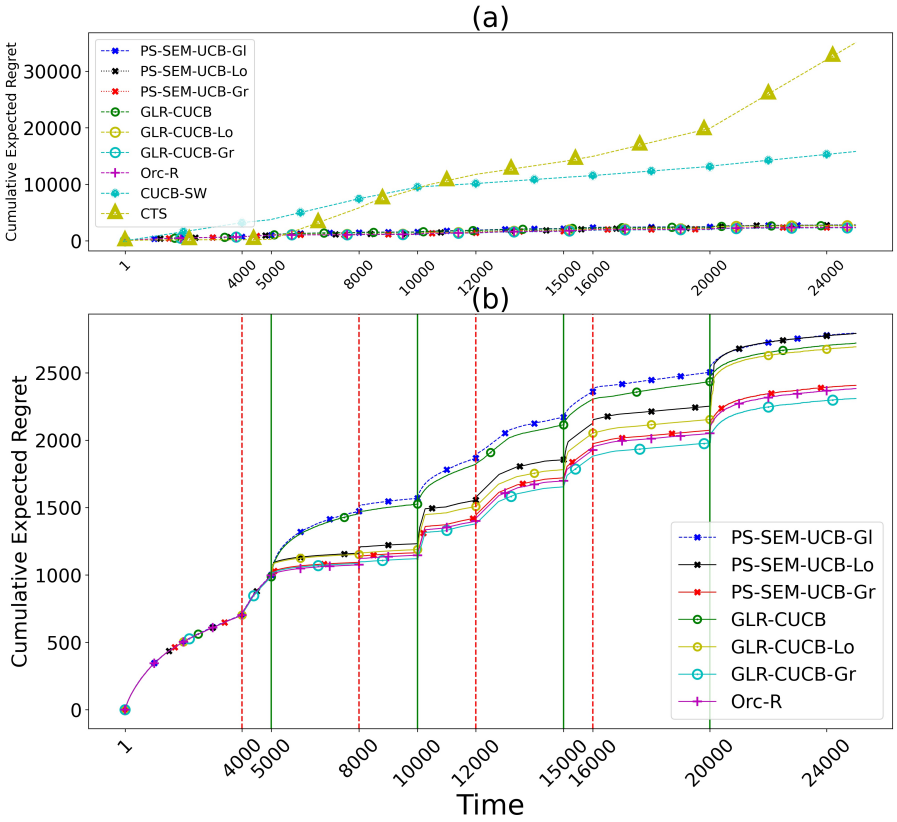


Figure 6.1: Cumulative Expected Regret.

6.5.1 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 $\boldsymbol{W}_t$ 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 $\boldsymbol{b}_t$ follows a multivariate normal distribution with the support in $[0, 1]^{18}$ and a spherical covariance matrix. In supplementary material, Figure 1 visualizes the expected values of base arms' rewards across time, and Figure 2 presents the visualization of optimal super arm across time. As shown in Section 6.2, the reward generation process follows the SEM in eq. (6.3). All distribution-stationary-segments of the environment 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 MSE $= \frac{1}{K^2} \left\| \boldsymbol{W}_t - \hat{\boldsymbol{W}}_t \right\|_{\mathrm{F}}^2$. Figure 6.1-a shows the poor perfor-

mance of *CUCB-SW*, and *CTS*, compared to other algorithms. In Figure 6.1-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 6.1-b, each vertical green solid line represents the time of a distribution-change, and each vertical red dashed line represents a graph-change.

6.5.2 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 Nourani-Koliji *et al.* (2022); nevertheless, we use a 14-day moving average instead of a 7-day moving average. Instead of the L^1-norm in eq. (6.8), we use the Directed Total Variation (DTV) $\sum_{i,j\in\mathcal{K}} \boldsymbol{W}[i,j] \sum_{h=1,\dots,t} [\boldsymbol{y}[i,h] - \boldsymbol{y}[j,h]]^+$ regularizer Sardellitti *et al.* (2017), 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 eq. (6.8). 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 training set. We calculate the prediction error at each time t by $Error(t) = \frac{1}{K|\mathfrak{v}(t)|} \sum_{i\in\mathfrak{v}(t)} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|_1$ where $\mathfrak{v}(t)$ is the validation set at time t with cardinality $|\mathfrak{v}(t)|$. Besides, $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$ are respectively the validation data, and the corresponding predicted value using the estimated graph for day i. Figure 6.2 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 6.2, our algorithm estimates the data for each region efficiently. This helps the agent to find the optimal decision vector.

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

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

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 Wang *et al.* (2022), we divide the country into 4 clusters, using *graph-based clustering* library of *Python*, based on Euclidean distances between regional capitals. In Figure 6.3, 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.

6.6 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.

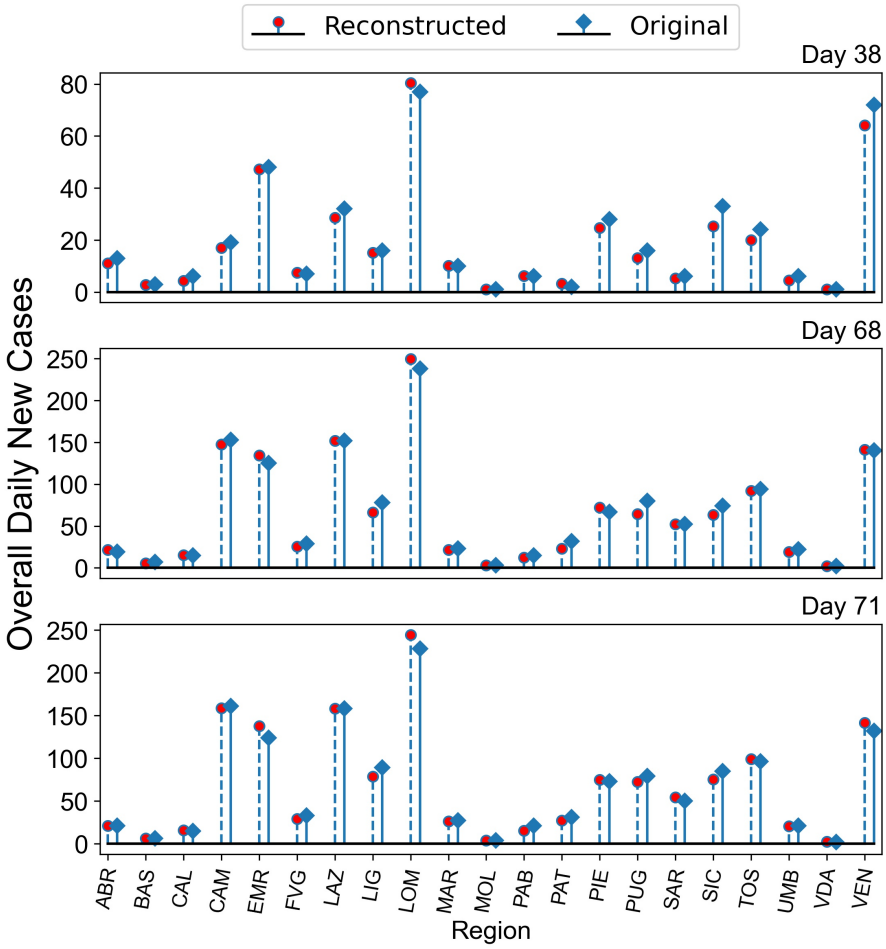


Figure 6.2: Original and reconstructed daily new cases.

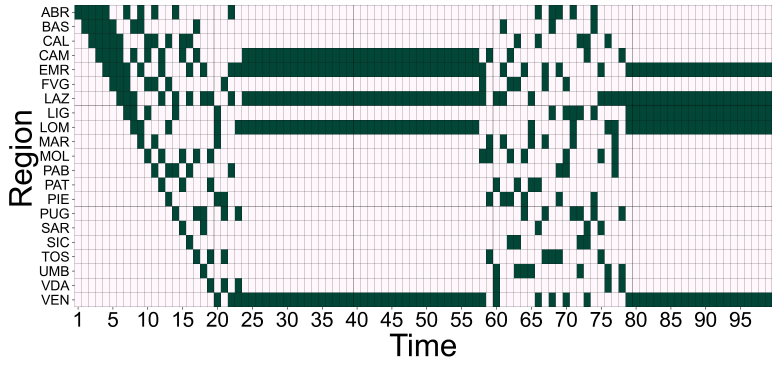


Figure 6.3: Selected regions on each day of the experiment.

Chapter 7

Results and Discussion

7.1 Research Themes

The major themes of this dissertation involve: 1) improving the precision of a reward model by transferring information from related tasks, 2) reducing the need for exploration while circumventing the usual issues of estimation uncertainties, and 3.) enhancing the theoretical regret bounds, with these improvements stemming directly from the aforementioned themes. Under appropriate conditions, we showed the benefits of transferring external knowledge to several MAB settings.

Reward estimates were shown to have been improved in all projects. The works of Chapters 3 to 5 indirectly contributed to this improvement through more precise estimation of feature vectors. On the other hand, Chapter 6, which did not utilize contextual information, illustrated improvements in reward estimations by preserving relevant data during environmental change detection. The idea of minimizing exploration with side information aligns with the notion of increasing confidence in estimation models and is a recurring theme across all chapters. While Chapter 3 seeks to enhance confidence in their estimation using a convex combination of source and target models, Chapter 4 adopted a more nuanced strategy by constraining the confidence ellipsoid and posterior distribution along specific dimensions informed by the meta structure, subsequently tightening the bound on $\left\|\boldsymbol{\theta}^{\star} - \hat{\boldsymbol{\theta}}\right\|_{\boldsymbol{B}}$. In Chapter 5, the agent operates under a greedy policy without explicit exploration, instead offering a bound on the oracle inequality $\left\|\boldsymbol{\Theta}^{\star} - \hat{\boldsymbol{\Theta}}\right\|_2$ for all task parameters. We demonstrated that this bound converges sufficiently fast to zero, which automatically implies an increasing confidence in the reward estimations for every arm. The framework in Chapter 6 lacks context vectors, simplifying the reward model to the mean reward estimate of each base-arm distribution. The agent relies on UCB values for every arm, where repeated selection augments the confidence of an arm's mean reward estimation. By restarting only subsets of the action set in response to environmental changes, the confidence for each individual arm is further increased. In all instances, the decreased exploration requirement directly led to a theoretical

and empirical improvement of the expected regret bound.

In Chapter 3, we showed that the improvement of the regret bound depends on the Euclidean distance between the true feature vectors of source and target and that as long as the source is viable for the learner up to some time κ, the regret grows by $\mathcal{O}\left(U_{\min}\kappa\sqrt{\log(\kappa)}\right)$ and since $U_{\min}$ itself shrinks with $\frac{1}{\sqrt{\kappa}}$, the regret rate up to time κ is $\mathcal{O}\left(\sqrt{\kappa\log(\kappa)}\right)$, an improvement over the LinUCB baseline. Conversely, in Chapter 4, the improvement in the regret is linked to both the problem's dimensionality and the low variance in the meta distribution, the bound itself depends on the estimation of the affine subspace: $\mathcal{O}\left(\sqrt{n}\left(p\log\left(1+\frac{nV^2}{p}\right)+q\log\left(1+\frac{n\sqrt{Y}}{q}\right)\right)\right)$, where Y contains the low variance term and the mean error of the subspace estimation. Chapter 5 highlighted how the network lasso regularization can reduce the regret bound in relation to the total number of tasks, thanks to the cluster structure and the graph properties in our model. We show for graphs that satisfy certain conditions with respect to the underlying cluster structure that the regret is upper bounded by $\tilde{\mathcal{O}}\left(\frac{1}{\phi^2}\sqrt{\bar{T}|\mathcal{V}|}\right)$, with $\bar{T}=\frac{T}{|\mathcal{V}|}$. In Chapter 6, adopting the group restart strategy improves the regret bound because the initialization routines require less exploration time. This leads to greater confidence in arms that do not need restarting and enhances the effectiveness of the change-point detector, as missing a change in the base-arm distributions becomes less probable.

The meta structure exploited in Chapter 4 serves as a generalized source of transfer compared to that in Chapter 3. In the latter case, our model's effectiveness is tied to the Euclidean distance between the feature vectors of the target and source tasks. However, the former adopts a broader approach by employing learned projections to take advantage of low Euclidean distances within subspaces. Similarly, the approach in Chapter 5 takes advantage of the close proximity of the task feature vectors, but only within the same cluster. Between Chapters 4 and 5 we can also draw parallels: In the special case of having fewer clusters than dimensions in the setup of Chapter 5, the setting automatically features a subspace structure for the true feature vectors with the number of dimensions equal to the number of clusters.

A notable feature of the models introduced in Chapters 3 to 5 is their resilience against potential negative transfer. A key point in Chapter 3 is that if certain experts are not suitable for transfer learning in the target domain, the proposed mechanism would discard the expert and return to the classic model, thus mitigating any potential negative transfer in the regret bound with high probability. Similarly, as presented in Chapter 4, if there is no low-dimensional subspace structure in the meta distribution, the algorithm will determine $\hat{\boldsymbol{P}}=\boldsymbol{I}$, which means the classic model without meta learning. In Chapter 5 we transfer information

between tasks by exploiting a cluster structure using graph-induced network lasso regularization. The quality of the graph does not need to be optimal that is, there can be inter cluster edges to some extent without the model failing.

Regarding future directions, it would be beneficial to investigate potential generalizations of the models we presented. The model detailed in Chapter 3 could be adapted to a deep reinforcement learning environment, investigating a convex combination of expert policies and target networks. Instead of assuming a linear subspace as the meta structure in Chapter 4, one might consider a non-linear generalization as manifold, necessitating the application of Gaussian processes for its estimation. The group restart strategy outlined in Chapter 6 depends on preexisting knowledge of groups, so developing an algorithm that identifies groupings of arms during the learning phase would be a more feasible approach.

7.2 Limitations and Open Problems

We discussed the major themes and contributions of this dissertation. However, specific limitations regarding individual chapters, as well as subsequent open problems, require more attention.

Chapter 3 presents a model designed to transfer knowledge from a source task without the need for additional source datasets, relying solely on the condition that the Euclidean distance between the source and target feature vectors is small enough. Although the proposed model improves the regret bound of OFUL with minimal requirements, it still requires prior knowledge of an upper bound $U \geq \|\boldsymbol{\theta}^\star - \boldsymbol{\theta}_S\|_2$, to update the weights, which might be unattainable in practical applications. Furthermore, we only provided theoretical guarantees for LinUCB or OFUL based algorithms and ignored any application to the Bayesian setting. Thus, utilizing the strategy of weighted models in Bayesian based algorithms, such as linear Thompson sampling, remains an open problem.

Although experiments in Chapter 4 can be conducted without any additional knowledge by utilizing the eigenvalues obtained via PCA, the regret bounds of the proposed meta learning model depend on knowing the dimension q and the variance term Var_ρ of the lower dimensional subspace. Subsequently, an open problem is to prove the validity of the algorithm by relying on the PCA estimated eigenvalues instead. First, in our experiments, we have the option to choose the dimensionality of the subspace, which determines the number of eigenvectors to form the projection matrices, based on the largest eigengap found in the ordered set of eigenvalues calculated via PCA. As we apply PCA on the set of estimated feature vectors $\hat{\boldsymbol{\Theta}}$, the obtained eigenvalues vary from the eigenvalues of the true covariance of the

meta distribution. The challenge of estimating a lower bound on the probability that the largest eigengap identified through PCA matches the largest eigengap of the true covariance remains unsolved. And second, for the same reason, we can not simply infer the low variance term Var_ρ: Recall that $\text{Var}_\rho = \sum_{i=1}^{q} \sigma_i$, where $\{\sigma_i\}_{i\in\{1,\dots,q\}}$ are the q smallest eigenvalues of the true covariance. Consequently, it is a challenge to utilize the PCA eigenvalues to establish an upper bound on Var_ρ with high probability.

Another limitation of our model from Chapter 4 concerns the minimum eigenvalue $\lambda_{\min}$ of all regularized estimated gram matrices $\boldsymbol{A}(t)$, where t indexes the respective task. In order for our model to be reliable, we need the minimum eigenvalue to grow sufficiently fast with each round. A sufficiently fast growing eigenvalue ensures that the feature vector estimates $\hat{\boldsymbol{\theta}}(t)$ are close to the true feature vectors of the respective task. In Bastani *et al.* (2021) they show that under a covariate diversity assumption for the contextual linear bandit setting, where each arm is associated with its own feature vector, the minimum eigenvalue grows linearly with high probability under a greedy policy. In Banerjee *et al.* (2023) it was shown that the minimum eigenvalue of any linear bandit algorithm with a regret rate of at most $\sqrt{n}$ increases at least with a rate of $\sqrt{k}$ with the number of rounds k if the set of context vectors is rich enough. On top of that, the work of Oh *et al.* (2021) and Cella *et al.* (2023) as well as our work presented in Chapter 5 show that greedy algorithms can guarantee converging oracle inequalities, if the context vector generation meets certain assumptions. At this point, we want to highlight an arm selection strategy that involves forced exploration with vanishing frequency: Under a mild coverage assumption that the set of available context vectors spans the entire context space, that is, for every unit vector $\boldsymbol{e} \in \mathbb{S}^d$, where $\mathbb{S}^d = \{\boldsymbol{e}_i\}_{i=1}^d$ is the canonical basis set, there is a context vector $\boldsymbol{x}_a$ such that $\boldsymbol{e}^\top \boldsymbol{x}_a > c_0$ for some $c_0 > 0$. We can select actions that maximize only the exploration term: $a = \arg\max_{a'} \sqrt{\boldsymbol{x}_{a'}^\top \boldsymbol{A}^{-1} \boldsymbol{x}_{a'}}$ and subsequently guarantee an increase of the minimum eigenvalue. By forcing this exploration strategy into every round k' with $k' \in \{k' = \lfloor i^l \rfloor : i \in \mathbb{N}\}$ for some $l > 1$, the frequency of forced exploration would vanish over time and the cumulative regret bound would increase by an additive term of the order $n^{\frac{1}{l}}$ that would become negligible. A quantitative analysis that finds a value of l that optimizes the trade-off between the exploration frequency and the benefit within the meta learning algorithm is essential. We can generalize this idea by expressing the set of pure exploration rounds as: $\{k' = \lfloor f(i) \rfloor : i \in \mathbb{N}\}$ with some superlinear function $f : \mathbb{R} \to \mathbb{R}$. The super-linearity of $f(i)$ leads to a sublinear additive term of the order $f^{-1}(n)$ to the regret bound.

The framework in Chapter 5 is based on the assumption of a piecewise-constant signal, implying that any two tasks within the same cluster possess identical feature vectors. To ensure the validity of the regret bounds, it is crucial that this

condition is met without any variations within the clusters. Consequently, a logical direction for future research is to relax this assumption and allow such variations within clusters while maintaining overall intra-cluster smoothness. We expect such a relaxation to affect the network lasso regularization term in eq. (5.4) to include a inter-cluster variation component. Consequently, another challenge would be to consider the possibility of cross-cluster fusions, as large variations in feature vectors within the same cluster can reduce the model's efficiency in distinguishing similar clusters. For that matter, further assumptions about the cluster structure might be necessary.

In Chapter 5, we established an upper bound for the restricted eigenvalue ϕ^2 under our proposed restricted eigenvalue condition and showed that there is an implicit dependency on the dimension of the feature vectors, which in turn influences the regret bound. Specifically, our regret bound in Theorem 5.5.10 must increase at least linearly with the dimension d due to this implicit dependency in ϕ^2. What remains unsolved in our model, and other lines of work that make use of restricted eigenvalue or compatibility conditions, is to find a proper lower bound for the restricted eigenvalue. Although we demonstrated that the regret must grow at least linearly with d, it remains uncertain whether higher orders of d are hidden in the term $\frac{1}{\phi^2}$.

In Chapter 6, a group restart strategy is introduced to facilitate optimal information transfer, specifically concerning base-arm feedback estimations across stationary segments. We demonstrate that with the knowledge of optimal arm groups we minimize both the number of re-initialization rounds after changes are detected and the expected delay of a detected change after it occurs. These factors contribute to enhancing the regret bound. However, as previously noted, the assumption of prior knowledge of these groups is quite strong. Future research that would significantly improve active detection methods for non-stationary settings should focus on developing a clustering algorithm capable of identifying groups of arms that change simultaneously with high probability p_{group}. This would manifest in the immediate regret as an additive term $(1 - p_{\text{group}})\Delta_{\max}$ during each round. The primary challenge involves designing an algorithm that ensures the complementary probability $1 - p_{\text{group}}$ decreases sufficiently fast over time such that it becomes negligible in the cumulative regret bound. It is important to note that the combinatorial bandit setting is particularly suited for applying such a clustering approach, as it provides feedback from multiple actions each round. This works in favor of a change-point detector that can identify multiple changes in a short period, thereby enhancing the efficiency of a clustering approach.

7.3 Summary and Conclusion

This dissertation introduced a variety of algorithms designed to utilize transfer learning across different bandit frameworks. Our approach successfully minimizes exploration and enhances the precision of model estimations. In Chapter 3, we demonstrated the effectiveness of using a convex combination of source and target models to harness the expertise relevant to the current task. We quantified an improvement to the upper regret bound that depends on the euclidean distance between source and target feature vectors and showed that our model is resilient against negative transfer. Meanwhile, Chapter 4 introduced a meta learning algorithm that capitalizes on the similarity between tasks within an unknown subspace. This algorithm learns a projection matrix after a certain amount of initial tasks, which is subsequently applied to future tasks in a modified OFUL algorithm to adjust the axes of the confidence ellipsoid associated with the learned subspace, and serves as a prior in a Thompson sampling algorithm. We expressed the bounds on the expected transfer regrets which depend on the dimensionality of the subspace, its low-variance term as well as the estimation error of the projection matrix. In Chapter 5, we developed a multi-task learning algorithm that facilitates information transfer between tasks utilizing a cluster structure using a graph given to the learner. We exploited the cluster structure implicitly by using the graph in a network lasso regularization. We proposed a novel oracle inequality and showed that it holds with high probability in a bandit setting under the assumption of a satisfied restricted eigenvalue condition for the covariance of the context vector generating distribution. That allowed us to provide regret bounds for the multi-task learning setting and showed which conditions the graph needs to satisfy with respect to the cluster structure. Chapter 6 addressed non-stationary environments by efficiently transferring information across stationary segments employing a group restart strategy for change-point detectors. We motivated a group restart approach by highlighting the improvement, both theoretically and empirically, over local and global restart strategies. Throughout all chapters, we theoretically showed that our regret bounds improve based on the transfer learning related assumptions and substantiated our work with empirical results on both synthetic and real data sets. Our proposed models hold promise for various online learning applications such as recommender systems, online advertising or clinical trials where prior data and knowledge from users, customers or patients are readily available. Furthermore, our research could be of interest for the continual reinforcement learning field that focuses on AI agents that are exposed to a stream of tasks and need to continually adapt. Retaining and transferring previously acquired knowledge, as studied in this dissertation for different frameworks, is promising for the design of future algorithms.

Appendix A

Hypothesis Transfer in Bandits by Weighted Models

Proof of Theorem 3.4.1

We use the lemmas from Abbasi-Yadkori *et al.* (2011) used in their regret analysis, as we use the results for our proof.

Lemma A.0.1. *(Self-normalized bound for vector-valued martingales). Let τ be a stopping time with respect to the filtration $\{F_t\}_{t=0}^{\infty}$ and define $\boldsymbol{S}_t = \boldsymbol{D}^\top(\tau)\boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon}$ as a sub-Gaussian noise vector. Then, for $\delta > 0$, with probability $1-\delta$,*

$$\|\boldsymbol{S}_\tau\|^2_{\boldsymbol{A}^{-1}(\tau)} \leq \log\left(\frac{\det(\boldsymbol{A}(\tau))}{\delta^2\lambda^d}\right) \tag{A.1}$$

Lemma A.0.2. *Suppose $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, ..., \boldsymbol{x}_{a_n} \in \mathbb{R}^d$ and for any $1 \leq k \leq n$, $\|\boldsymbol{x}_{a_k}\| \leq 1$. Let $\boldsymbol{A} = \lambda\boldsymbol{I} + \sum_{k=1}^n \boldsymbol{x}_{a_k}\boldsymbol{x}_{a_k}^\top$ for some $\lambda > 0$. Then,*

$$\det(\boldsymbol{A}) \leq (\lambda + n/d)^d \tag{A.2}$$

Lemma A.0.3. *(Confidence Set Bound). Suppose $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, ..., \boldsymbol{x}_{a_n} \in \mathbb{R}^d$ and for any $1 \leq k \leq n$, $\|\boldsymbol{x}_{a_k}\| \leq 1$. Let $\boldsymbol{A} = \lambda\boldsymbol{I} + \sum_{k=1}^n \boldsymbol{x}_{a_k}\boldsymbol{x}_{a_k}^\top$ for some $\lambda > 0$ and assume $\|\boldsymbol{\theta}^*\| \leq 1$, with $\hat{\boldsymbol{\theta}}_T = \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{y}$. Then, for any $\delta > 0$, with probability of at least 1-δ we have:*

$$\left\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\right\|_{\boldsymbol{A}} \leq \sqrt{d\log\left(\frac{(\lambda + n/d)}{\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} \tag{A.3}$$

Lemma A.0.4. *Suppose $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, ..., \boldsymbol{x}_{a_n} \in \mathbb{R}^d$ is a sequence and for any $1 \leq k \leq \infty$, $\|\boldsymbol{x}_k\| \leq 1$. Let $\boldsymbol{A}(n) = \lambda\boldsymbol{I} + \sum_{k=1}^n \boldsymbol{x}_k\boldsymbol{x}_k^\top$ for some $\lambda > 0$ then*

$$\sum_{k=1}^n \|\boldsymbol{x}_{a_k}\|^2_{A^{-1}(k)} \leq 2d\log\left(1 + \frac{k}{d\lambda}\right) \tag{A.4}$$

Proof of Theorem 3.4.1. First, similar to the confidence set bound in Lemma A.0.3, we require a confidence set bound for $\boldsymbol{\theta}_S$ in the sense of the $\|.\|_{\boldsymbol{A}(k)}$ norm. Since it remains constant, determining its confidence set bound is straightforward. Assuming $\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 = U$, we have

$$\begin{aligned}\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_{\boldsymbol{A}(k)} &= \sqrt{\lambda\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2^2 + \sum_{i=1}^{k}(\boldsymbol{x}_{a_k}^\top(\boldsymbol{\theta}_S - \boldsymbol{\theta}^*))^2} \\ &\leq \sqrt{\lambda\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2^2 + \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2^2 \sum_{i=1}^{k}\|x_{a_k}\|_2^2} \\ &\leq U\sqrt{\lambda + k},\end{aligned}$$

where in the last inequality, the Cauchy-Schwarz inequality was used and the fact that $\|\boldsymbol{x}_{a_k}\| \leq 1$. For simplicity we define the upper confidence set bounds for $\boldsymbol{\theta}_S$ as γ_S and for $\hat{\boldsymbol{\theta}}_T$ as γ_T. Now we determine the maximal time step κ at which $\gamma_S \leq \gamma_T$ is guaranteed. For this we require a lower bound for γ_T:

$$\gamma_T = \sqrt{d\log\left(1 + \frac{k}{\lambda d}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} \geq \sqrt{\frac{2k/\lambda}{2 + k/(\lambda d)} + \log\left(\frac{1}{\delta^2}\right)}, \tag{A.5}$$

where we used $\log(1+x) \geq \frac{2x}{2+x}$ for $x \in (0, \infty)$. With this lower bound we can analytically determine a lower bound κ:

$$\begin{aligned}&U\sqrt{\lambda + \kappa} = \sqrt{\frac{2\kappa/\lambda}{2 + \kappa/(\lambda d)} + \log\left(\frac{1}{\delta^2}\right)} \\ &\iff U^2(\lambda + \kappa) - \frac{2k}{2\lambda + \kappa/d} - \log\left(\frac{1}{\delta^2}\right) = 0 \\ &\iff \kappa^2 + \left(2\lambda d + \lambda - \frac{\log\left(\frac{1}{\delta^2}\right) + 2d}{U^2}\right)\kappa + 2\lambda d\left(\lambda - \frac{\log\left(\frac{1}{\delta^2}\right)}{U^2}\right) = 0\end{aligned}$$

Which is a quadratic inequality with respect to κ and yields the following solution, with the condition $\delta \leq \exp(-2\lambda)$ we retrieve yet another lower bound:

$$
\begin{aligned}
\kappa &= \sqrt{\left(\frac{d+\log\left(\frac{1}{\delta}\right)}{U^2}\right)^2 + \left[\lambda\left(d-\frac{1}{2}\right)\right]^2 + 2\lambda\frac{\log\left(\frac{1}{\delta}\right)(d-\frac{1}{2}) - d(d+\frac{1}{2})}{U^2}} \\
&- \lambda(d+\frac{1}{2}) + \frac{d+\log\left(\frac{1}{\delta}\right)}{U^2} \\
&\geq \sqrt{\left(\frac{d+\log\left(\frac{1}{\delta}\right)}{U^2}\right)^2 + \left[\lambda\left(d+\frac{1}{2}\right)\right]^2 - 2\lambda\frac{\log\left(\frac{1}{\delta}\right)(d+\frac{1}{2}) + d(d+\frac{1}{2})}{U^2}} \\
&- \lambda(d+\frac{1}{2}) + \frac{d+\log\left(\frac{1}{\delta}\right)}{U^2} \\
&= \sqrt{\left(\frac{d+\log\left(\frac{1}{\delta}\right)}{U^2} - \lambda\left(d+\frac{1}{2}\right)\right)^2} - \lambda(d+\frac{1}{2}) + \frac{d+\log\left(\frac{1}{\delta}\right)}{U^2} \\
&= 2\left[d\left(\frac{1}{U^2}-\lambda\right) + \frac{\log\left(\frac{1}{\delta}\right)}{U^2} - \frac{\lambda}{2}\right] \\
&\geq \left\lfloor 2\left[d\left(\frac{1}{U^2}-\lambda\right) + \lambda\left(\frac{2}{U^2}-\frac{1}{2}\right)\right]\right\rfloor
\end{aligned}
$$

Where in the first inequality we used

$$
\begin{aligned}
\lambda^2\left(d+\frac{1}{2}\right)^2 - \lambda^2\left(d-\frac{1}{2}\right)^2 &\leq 2\lambda\frac{\log\left(\frac{1}{\delta}\right)(d-\frac{1}{2})}{U^2} - \left(-2\lambda\frac{\log\left(\frac{1}{\delta}\right)(d+\frac{1}{2})}{U^2}\right) \\
&\iff 2\lambda^2 d \leq \frac{4\lambda\log\left(\frac{1}{\delta}\right)d}{U^2},
\end{aligned}
$$

which holds when $\delta \leq \exp(-\lambda/2)$. Next we we give the upper confidence bound of our model:

$$
|\boldsymbol{x}_{a_k}^\top\hat{\boldsymbol{\theta}} - \boldsymbol{x}_{a_k}^\top\boldsymbol{\theta}^*| \leq \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_{\boldsymbol{A}(k)} ||\boldsymbol{x}_{a_k}||_{\boldsymbol{A}^{-1}(k)} = \Delta(k), \tag{A.6}
$$

where the Cauchy-Schwarz inequality was used. We denote the exploration term of the UCB as $\Delta(k)$. We construct the regret as a sum of immediate regrets $\rho(k)$:

$$
R(n) = \sum_{k=1}^{n}\rho(k), \tag{A.7}
$$

as for the immediate regrets, we define the context vector yielding the highest reward $\boldsymbol{x}_{a^*}$. We then have

$$\begin{aligned}\rho(k) &= \boldsymbol{x}_{a^*}^\top \boldsymbol{\theta}^* - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^* \\ &\leq \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^* \\ &\leq \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) - \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) \\ &= 2\Delta(k).\end{aligned}$$

The first inequality makes use of the UCB principle optimism in the face of uncertainty and the second inequality results from the definition of the confidence set used for the exploration term. The resulting total regret can then be bounded:

$$\begin{aligned}R(n) &= \sum_{k=1}^{n} \rho(k) \\ &\leq 2\sum_{k=1}^{n} \Delta(k) = 2\sum_{k=1}^{n} \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_{\boldsymbol{A}(k)} \|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} \\ &\leq 2\sum_{k=1}^{n} \left[\alpha_S(k) ||\boldsymbol{\theta}_S^* - \boldsymbol{\theta}^*||_{\boldsymbol{A}(k)} + \alpha_T(k) ||\hat{\boldsymbol{\theta}}_T(k) - \boldsymbol{\theta}^*||_{\boldsymbol{A}(k)} \right] ||\boldsymbol{x}_{a_k}||_{\boldsymbol{A}^{-1}(k)} \\ &\leq 2\sum_{k=1}^{n} \alpha_S(k) \left(U\sqrt{\lambda + k} - ||\hat{\boldsymbol{\theta}}_T(k) - \boldsymbol{\theta}^*||_{\boldsymbol{A}(k)} \right) ||\boldsymbol{x}_{a_k}||_{\boldsymbol{A}^{-1}(k)} \\ &+ 2\sum_{k=1}^{n} ||\boldsymbol{\theta}_T(k) - \boldsymbol{\theta}^*||_{\boldsymbol{A}(k)} ||\boldsymbol{x}_{a_k}||_{\boldsymbol{A}^{-1}(k)} \\ &\leq 2U\sqrt{\lambda + \kappa} \sqrt{\kappa \sum_{k=1}^{\kappa} ||\boldsymbol{x}_{a_k}||^2_{\boldsymbol{A}^{-1}(k)}} - R_T(\kappa) + R_T(n) \\ &\leq U\sqrt{8\kappa(\lambda + \kappa) d \log\left(1 + \frac{\kappa}{d\lambda}\right)} - R_T(\kappa) + R_T(n)\end{aligned}$$

While we used Lemma A.0.3 in the fourth inequality, resulting into the classic regret and Lemma A.0.4 in the last step. □

Proof of Theorem 3.4.2

Proof of Theorem 3.4.2. We assume that $\gamma_S > \gamma_T$ from the very beginning:

$$R(n) \leq \sum_{k=1}^{n} \alpha_S(k)(\gamma_S(k) - \gamma_T(k)) \|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} + R_T(n) \tag{A.8}$$

with $R_T(n)$ as the traditional regret bound for LinUCB. We define $\Delta_{\gamma(k)} = \gamma_S(k) - \gamma_T(k)$. We are taking a closer look at the worst case scenario with $\Delta_{\gamma(k)} > 0$ First we show how the weights evolve in the softmax approach:

$$\alpha_S(k) = \frac{1}{1 + Z(k)(\frac{1}{\alpha_S(k-1)} - 1)} = \frac{1}{1 + \prod_{i=1}^{k} Z(i)(\frac{1}{\alpha_S(0)} - 1)}, \tag{A.9}$$

with $Z(k) = \exp\big(\beta\Delta_{\gamma(k)}\big)$, thus in case $\Delta_{\gamma(k)} > 0$ for all k we can further bound the regret as:

$$R(n) \leq \sum_{k=1}^{n} \frac{1}{1 + \prod_{i=1}^{k} Z(i)(\frac{1}{\alpha_S(0)} - 1)} \Delta_{\gamma(k)} \|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} + R_T(n) \tag{A.10}$$

looking at the first sum we know that for large values of $\Delta_{\gamma(k)}$ the sigmoid term decreases rapidly by taking the upper bound: $\frac{1}{1+\prod_{i=1}^{k} Z(i)(\frac{1}{\alpha_S(0)}-1)} \leq \frac{1}{\prod_{i=1}^{k} Z(i)(\frac{1}{\alpha_S(0)}-1)}$ we can minimize this locally by setting $\Delta_{\gamma(k)} = \frac{1}{\beta}$ for the kth summand respectively. thus we can further estimate our upper regret bound such that

$$R(n) \leq \sum_{k=1}^{n} \frac{\exp\Big(-\beta \sum_{i=1}^{k-1} \Delta_{\gamma_i}\Big)}{e\beta(\frac{1}{\alpha_S(0)} - 1)} \|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}(k)} + R_T(n), \tag{A.11}$$

From here we will focus on the negative transfer term only. Since we know that $\Delta_{\gamma(i)}$ grows with each time step, as well as in this case it is supposed to be positive for a bad source scenario, we can further estimate:

$$\sum_{k=1}^{n} \frac{\exp\Big(-\beta \sum_{i=1}^{k-1} \Delta_{\gamma_i}\Big)}{e\beta(\frac{1}{\alpha_S(0)} - 1)} \|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}(k)} \leq \sum_{k=1}^{n} \frac{\exp(-\beta(k-1)\Delta_{\min})}{e\beta(\frac{1}{\alpha_S(0)} - 1)} \tag{A.12}$$

where we used $\|x\|_{A^{-1}(k)} \leq 1$ and defined $\Delta_{\min} = \min_k \Delta_{\gamma(k)}$. With the use of the geometric series we finally obtain:

$$\sum_{k=0}^{n-1} \frac{\exp(-\beta k \Delta_{\min})}{e\beta(\frac{1}{\alpha_S(0)} - 1)} \leq \frac{(1 - \alpha_T(0))}{e\beta\alpha_T(0)(1 - \exp(-\beta\Delta_{\min}))}, \tag{A.13}$$

where we changed the sum indices in eq. (A.13) and applied the geometric series formula. □

A.1 Proof of Theorem 3.5.1

Proof of Theorem 3.5.1. The proof is analogous to Theorem 3.4.1 with the difference that multiple sources are available. Due to the algorithm it always picks the source with the lowest confidence set bound, denoted by: $\|\boldsymbol{\theta}_{S,m} - \boldsymbol{\theta}^*\|_{A(k)} \leq$

$\min_m U_m\sqrt{\lambda+k} = U_{\min}\sqrt{\lambda+k}$. Using this, the rest of the proof follows the same steps as Theorem 3.4.1. □

A.2 Proof of Theorem 3.5.2

Proof of Theorem 3.5.2. We assume $\gamma_{S,j}(k) > \gamma_T(k)$ for all $j \in 1, ..., M$ from the very beginning:

$$R_n \leq \sum_{k=1}^{n}\sum_{j}^{M} \alpha_{S,j}(k)(\gamma_{S,j}(k) - \gamma_T(k))\|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} + R_T \tag{A.14}$$

with R_T as the traditional regret bound for LinUCB. We define $\Delta_j(j) = \gamma_{S,j}(k) - \gamma_T(k)$. We are taking a closer look at the worst case scenario with $\Delta_j(k) > 0$ for all j. First we show how the weights evolve in the softmax approach:

$$\begin{aligned}
\alpha_{S,j}(k) &= \frac{1}{1 + \sum_{i\neq j} \frac{\alpha_{S,i}(k-1)}{\alpha_{S,j}(k-1)} \exp(\beta(\gamma_{S,j}(k) - \gamma_{S,i}(k))) + \exp(\beta\Delta_j(k))\frac{\alpha_T(k-1)}{\alpha_{S,j}(k-1)}} \\
&= \frac{1}{1 + \sum_{i\neq j} \exp\Big(\beta\sum_{l=1}^{k}(\gamma_{S,j}(l) - \gamma_{S,i}(l))\Big) + M\exp\Big(\beta\sum_{l=1}^{k}\Delta_j(l)\Big)(\frac{\alpha_T(0)}{1-\alpha_T(0)})} \\
&\leq \frac{1}{M\exp\Big(\beta\sum_{l=1}^{k}\Delta_j(l)\Big)(\frac{\alpha_T(0)}{1-\alpha_T(0)})},
\end{aligned}$$

were we assumed that each inital source weight is set to $\alpha_{S,j}(0) = \frac{1-\alpha_T(0)}{M}$, thus in case $\Delta_j(k) > 0$ for all j and k we can further bound the regret as:

$$R_n \leq \sum_{k=1}^{n}\sum_{j=1}^{M} \frac{1}{M\exp\Big(\beta\sum_{l=1}^{k}\Delta_j(l)\Big)(\frac{\alpha_T(0)}{1-\alpha_T(0)})}\Delta_j(k)\|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} + R_T \tag{A.15}$$

we know that for large values of Δ_j the respective term decreases rapidly we can minimize these locally by setting $\Delta_j(k) = \frac{1}{\beta}$ for the kth summand respectively for all sources. Thus we can further estimate the negative transfer term such that

$$R_n \leq \sum_{k=1}^{n}\sum_{j=1}^{M} \frac{\exp\Big(-\beta\sum_{l=1}^{k-1}\Delta_j(l)\Big)}{eM\beta\frac{\alpha_T(0)}{1-\alpha_T(0)}}\|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} + R_T. \tag{A.16}$$

From here we will ignore the classic regret term R_T and use it again at the end. Since we know that Δ_j grows with each time step, as well as in this case it

is supposed to be positive for a bad source scenario, we can further estimate the negative transfer term:

$$\sum_{k=1}^{n}\sum_{j=1}^{M}\frac{\exp\Big(-\beta\sum_{l=1}^{k-1}\Delta_j(l)\Big)}{eM\beta\frac{\alpha_T(0)}{1-\alpha_T(0)}}\|\boldsymbol{x}_{a_k}\|_{\boldsymbol{A}^{-1}(k)} \leq \sum_{k=1}^{n}\sum_{j=1}^{M}\frac{\exp(-\beta(k-1)\Delta_{\min,j})}{eM\beta\frac{\alpha_T(0)}{1-\alpha_T(0)}}$$

where we used $\|x\|_{A^{-1}(k)} \leq 1$ and used $\Delta_{\min,j} = \min_k \Delta_j(k)$. With the use of the geometric series we finally obtain:

$$\sum_{k=0}^{n-1}\sum_{j=1}^{M}\frac{\exp(-\beta k\Delta_{\min,j})}{eM\beta\frac{\alpha_T(0)}{1-\alpha_T(0)}} \leq \sum_{j=1}^{M}\frac{(1-\alpha_T(0))}{eM\beta\alpha_T(0)(1-\exp(-\beta\Delta_{\min,j}))}, \tag{A.17}$$

where we changed the sum indices in eq. (A.17) and applied the geometric series formula. □

A.3 Proof of Theorem 3.5.3

The proof of the next Lemma and Theorem is adapted from Abbasi-Yadkori *et al.* (2011).

Lemma A.3.1. *Suppose $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, ..., \boldsymbol{x}_{a_n} \in \mathbb{R}^d$ and for any $1 \leq k \leq n$, $\|\boldsymbol{x}_{a_k}\| \leq 1$. Let $D = \{x_{a_i}\}_{i=1}^{k-1}$, $\boldsymbol{A} = \lambda\boldsymbol{I} + \sum_{k=1}^{n}\boldsymbol{x}_{a_k}\boldsymbol{x}_{a_k}^\top$ for some $\lambda > 0$ and assume $\|\boldsymbol{\theta}^*\| \leq 1$. A source bandit parameter $\boldsymbol{\theta}_S$ is given as well. With the estimation $\hat{\boldsymbol{\theta}}_T = \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{y} - (\boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\theta}_S$, then, for any $\delta > 0$, with probability of at least 1-δ we have:*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_A \leq \sqrt{d\log\left(1+\frac{k}{d\lambda}\right) - 2\log(\delta)} + \sqrt{\lambda}\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 \tag{A.18}$$

Proof. Suppose $\boldsymbol{x}_{a_1}, \boldsymbol{x}_{a_2}, ..., \boldsymbol{x}_{a_n} \in \mathbb{R}^d$ and for any $1 \leq k \leq n$, $\|\boldsymbol{x}_{a_k}\| \leq 1$. Let $D = \{x_{a_i}\}_{i=1}^{k-1}$, $\boldsymbol{A} = \lambda\boldsymbol{I} + \sum_{k=1}^{n}\boldsymbol{x}_{a_k}\boldsymbol{x}_{a_k}^\top$ for some $\lambda > 0$ and assume $\|\boldsymbol{\theta}^*\| \leq 1$. A source bandit parameter $\boldsymbol{\theta}_S$ is given as well. With the estimation $\hat{\boldsymbol{\theta}}_T = \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{y} - (\boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\theta}_S$, then, for any $\delta > 0$, with probability of at least 1-δ we have:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{y} - (\boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{D} - \boldsymbol{I})\boldsymbol{\theta}_S \\
&= \boldsymbol{A}^{-1}\boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}) - \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\theta}_S + \boldsymbol{\theta}_S \\
&= \boldsymbol{\theta}^* - \lambda\boldsymbol{A}^{-1}\boldsymbol{\theta}^* + \boldsymbol{A}^{-1}\boldsymbol{D}^\top\boldsymbol{\epsilon} + \lambda\boldsymbol{A}^{-1}\boldsymbol{\theta}_S
\end{aligned}$$

Next by shifting $\boldsymbol{\theta}^*$ to the left as well as applying the Cauchy-Schwarz inequality after doing using a scalar product with $\boldsymbol{x}$ we get:

$$\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, x \rangle \leq \|x\|_{\boldsymbol{A}^{-1}} (\left\| \boldsymbol{D}^\top \boldsymbol{\epsilon} \right\|_{\boldsymbol{A}^{-1}} + \lambda \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_{\boldsymbol{A}^{-1}}), \tag{A.19}$$

next by using $\|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|^2_{A^{-1}} \leq 1/\lambda \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|^2_2$, Lemma A.0.1 and by plugging in $x = A(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ we get:

$$\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|^2_A \leq \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_A \left(\sqrt{d \log\left(1 + \frac{k}{d\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 \right)$$

thus as confidence set required for our UCB we get:

$$\left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_A \leq \sqrt{d \log\left(1 + \frac{k}{d\lambda}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda} \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2 \tag{A.20}$$

□

Proof of Theorem 3.5.3. We give the upper confidence bound of the biased regularization model:

$$|\boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^*| \leq \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_{\boldsymbol{A}(k)} ||\boldsymbol{x}_{a_k}||_{\boldsymbol{A}(k)^{-1}} = \Delta(k), \tag{A.21}$$

where the Cauchy-Schwarz inequality was used. The next steps, are mostly identical to Theorem 3.4.1. We denote the exploration term of the UCB as $\Delta(k)$. We construct the regret as a sum of immediate regrets $\rho(k)$:

$$R_n = \sum_{k=1}^{n} \rho(k), \tag{A.22}$$

as for the immediate regrets, we define the context vector yielding the highest reward $\boldsymbol{x}_{a^*}$. We then have

$$\begin{aligned}
\rho(k) &= \boldsymbol{x}_{a^*}^\top \boldsymbol{\theta}^* - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^* \\
&\leq \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) - \boldsymbol{x}_{a_k}^\top \boldsymbol{\theta}^* \\
&\leq \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) - \boldsymbol{x}_{a_k}^\top \hat{\boldsymbol{\theta}} + \Delta(k) \\
&= 2\Delta(k).
\end{aligned}$$

The first inequality makes use of the UCB principle optimism in the face of uncertainty and the second inequality results from the definition of the confidence

set used for the exploration term. The resulting total regret can then be bounded:

$$\begin{aligned}
R_n &= \sum_{k=1}^{n} \rho(k) \\
&\leq 2 \sum_{k=1}^{n} \Delta(k) = 2 \sum_{k=1}^{n} \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_{\boldsymbol{A}(k)} \left\| \boldsymbol{x}_{a_k} \right\|_{\boldsymbol{A}(k)^{-1}} \\
&\leq 2 \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_{\boldsymbol{A}_n} \sqrt{n \sum_{k=1}^{n} \left\| \boldsymbol{x}_{a_k} \right\|^2_{A(k)^{-1}}} \\
&\leq \left(\sqrt{d \log\Big(1 + \frac{n}{d\lambda}\Big) + \log\Big(\frac{1}{\delta^2}\Big)} + \sqrt{\lambda} U \right) \sqrt{8nd \log\Big(1 + \frac{n}{d\lambda}\Big)}
\end{aligned}$$

While we used Lemmas A.0.4 and A.3.1 in the last step as well as $U = \left\| \frac{1}{M} \sum_{j=1}^{M} U_j \right\| \geq \|\boldsymbol{\theta}_S - \boldsymbol{\theta}^*\|_2$. □

Appendix B

Meta Learning in Bandits within Shared Affine Subspaces

B.1 Proof of Theorem 4.5.5

In order to prove Theorem 4.5.5 in the main paper we provide proofs of additional Lemmas here or refer to the original works:

Proof of Lemma 4.5.1. Given Lemma 9 of Abbasi-Yadkori *et al.* (2011), we have:

$$\|\boldsymbol{\eta}_k\|^2_{\boldsymbol{B}_k^{-1}} \leq \log\left(\frac{\det(\boldsymbol{B}_k)}{\delta^2 \det\left(\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}}\right)}\right), \tag{B.1}$$

where the term $\det\left(\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}}\right)$ can be further evaluated knowing the eigenvalues of the matrix $\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}}$. With orthogonal projections $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{P}}^{\perp}$ and $\hat{\boldsymbol{P}}^{\perp} = \boldsymbol{I} - \hat{\boldsymbol{P}}$ it holds that for any eigenvector $\boldsymbol{e_P}$ of $\hat{\boldsymbol{P}}$ we have: $\hat{\boldsymbol{P}}^{\perp}\boldsymbol{e_P} = (\boldsymbol{I} - \hat{\boldsymbol{P}})\boldsymbol{e_P} = \boldsymbol{0}$ and vice versa for any eigenvector $\boldsymbol{e_{P^{\perp}}}$ of $\hat{\boldsymbol{P}}^{\perp}$: $\hat{\boldsymbol{P}}\boldsymbol{e_{P^{\perp}}} = \boldsymbol{0}$. Thus any eigenvector of $\hat{\boldsymbol{P}}$ or $\hat{\boldsymbol{P}}^{\perp}$ is also an eigenvector of $\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}}$:

$$\begin{aligned}(\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}})\boldsymbol{e_P} &= (0 + \lambda_2)\boldsymbol{e_P}, \\ (\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}})\boldsymbol{e_{P^{\perp}}} &= (\lambda_1 + 0)\boldsymbol{e_{P^{\perp}}},\end{aligned}$$

with eigenvalues λ_1 and λ_2. Lastly we require the multiplicities of both eigenvalues given by the dimension of nullspaces of the matrices $\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}} - \lambda_1 \boldsymbol{I} = (\lambda_2 - \lambda_1)\hat{\boldsymbol{P}}$ for λ_1 and $\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}} - \lambda_2 \boldsymbol{I} = (\lambda_1 - \lambda_2)\hat{\boldsymbol{P}}^{\perp}$ for λ_2, which are $q = \operatorname{rank}(\hat{\boldsymbol{P}}^{\perp})$ and $p = \operatorname{rank}(\hat{\boldsymbol{P}})$ respectively. Thus we get:

$$\det\left(\lambda_1 \hat{\boldsymbol{P}}^{\perp} + \lambda_2 \hat{\boldsymbol{P}}\right) = \lambda_1^q \lambda_2^p, \tag{B.2}$$

finalizing our proof. □

Proof of Lemma 4.5.2. Let λ_i' be the singular values of $\boldsymbol{D}^\top \boldsymbol{D}$ and $\|\boldsymbol{x}_a\| \leq 1$ then we have:

$$\begin{aligned}
\log\left(\frac{\det(\boldsymbol{B})}{\det\left(\lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}\right)}\right) &\leq \sum_{i=1}^{p} \log\left(1 + \frac{\lambda_i'}{\lambda_2}\right) + \sum_{i=p+1}^{d} \log\left(1 + \frac{\lambda_i'}{\lambda_1}\right) \\
&\leq p \log\left(1 + \frac{1}{p\lambda_2} \sum_{i=1}^{p} \lambda_i'\right) + q \log\left(1 + \frac{1}{q\lambda_1} \sum_{i=p+1}^{d} \lambda_i'\right) \\
&\leq p \log\left(1 + \frac{k}{p\lambda_2}\right) + q \log\left(1 + \frac{k}{q\lambda_1}\right)
\end{aligned}$$

where we applied the Jensen inequality in the second inequality and bounded the trace by $k\|\boldsymbol{x}_{a_k}\|^2 \leq k$ in the last inequality. □

Proof of Lemma 4.5.3. We leave out the subscript k during the proof for readability purposes. Our estimation of $\boldsymbol{\theta}^*$ for the projected LinUCB algorithm yields:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{B}^{-1}\left(\boldsymbol{D}^\top \boldsymbol{y} + \lambda_1 \boldsymbol{w}\right), \tag{B.3}$$

thus we can write:

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_{\boldsymbol{B}} &= \left\|\boldsymbol{B}^{-1}\left(\boldsymbol{D}^\top \boldsymbol{y} + \lambda_1 \boldsymbol{w}\right) - \boldsymbol{\theta}^*\right\|_{\boldsymbol{B}} \\
&= \left\|\boldsymbol{B}^{-1}\left(\boldsymbol{D}^\top (\boldsymbol{D}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}) + \lambda_1 \boldsymbol{w}\right) - \boldsymbol{\theta}^*\right\|_{\boldsymbol{B}} \\
&= \left\|\boldsymbol{B}^{-1}\left(\boldsymbol{D}^\top \boldsymbol{\epsilon} + \lambda_1 \boldsymbol{w}\right) - \boldsymbol{B}^{-1}\left(\lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}\right)\boldsymbol{\theta}^*\right\|_{\boldsymbol{B}} \\
&= \left\|\boldsymbol{D}^\top \boldsymbol{\epsilon} + \lambda_1 \boldsymbol{w} - \left(\lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}\right)\boldsymbol{\theta}^*\right\|_{\boldsymbol{B}^{-1}} \\
&\leq \left\|\boldsymbol{D}^\top \boldsymbol{\epsilon}\right\|_{\boldsymbol{B}^{-1}} + \lambda_1 \left\|\hat{\boldsymbol{P}}^\perp\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\|_{\boldsymbol{B}^{-1}} + \lambda_2 \left\|\hat{\boldsymbol{P}}\boldsymbol{\theta}^*\right\|_{\boldsymbol{B}^{-1}} \\
&\leq \sqrt{2\log\left(\frac{\sqrt{\det(\boldsymbol{B})}}{\delta\sqrt{\det\left(\lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}\right)}}\right)} + \frac{\lambda_1}{\lambda_{\min}(\boldsymbol{B})}\left\|\hat{\boldsymbol{P}}^\perp\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)\right\| + \frac{\lambda_2}{\lambda_{\min}(\boldsymbol{B})}\left\|\hat{\boldsymbol{P}}\boldsymbol{\theta}^*\right\| \\
&\leq \sqrt{2\log\left(\frac{\sqrt{\det(\boldsymbol{B})}}{\delta\sqrt{\det\left(\lambda_1 \hat{\boldsymbol{P}}^\perp + \lambda_2 \hat{\boldsymbol{P}}\right)}}\right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \\
&\leq \sqrt{p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right) + \log\left(\frac{1}{\delta^2}\right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W,
\end{aligned}$$

where we used Lemma 4.5.1 in the second inequality. Here, $\lambda_{\min}()$ is a function returning the minimal eigenvalue of a given matrix. □

For the upper bound on the projection based error term in Lemma 4.5.4, need to make some definitions: We denote $\boldsymbol{\mu}$ as the true mean of the distribution of tasks ρ, $\bar{\boldsymbol{\theta}}^* = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}^*(i)$ as the mean estimated by the true task parameters and $\bar{\boldsymbol{\theta}} = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}(i)$ as the mean estimated by the $L2-$regularized ridge estimators. We define $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_d$ as ordered eigenvalues of the true covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\Delta} = \boldsymbol{\theta}^* - \boldsymbol{\mu}$ as random variable with mean zero and $\boldsymbol{\xi} = \boldsymbol{\mu} - \bar{\boldsymbol{\theta}}$ as difference between the estimated and true mean, furthermore we define the covariance matrices $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^* = \frac{1}{t}\sum_{i=1}^{t}(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*)(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*)^\top, \hat{\boldsymbol{\Sigma}} = \frac{1}{t}\sum_{i=1}^{t}(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}})^\top$ as the true covariance matrix, the covariance estimated by $\boldsymbol{\theta}^*(i)$ and the covariance estimated by $\hat{\boldsymbol{\theta}}(i)$. We also define vertical concatenations $\boldsymbol{U} = [\boldsymbol{u}_j^\top]_{j\in\{1,...,p\}}^\top$, $\boldsymbol{U}^* = [\boldsymbol{u}_j^{*\top}]_{j\in\{1,...,p\}}^\top$ and $\hat{\boldsymbol{U}} = [\hat{\boldsymbol{u}}_j^\top]_{j\in\{1,...,p\}}^\top$, with $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_p\}$, $\{\boldsymbol{u}_1^*, ..., \boldsymbol{u}_p^*\}$, $\{\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_p\}$ being the eigenvectors corresponding to the p largest eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^*$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Similarly we define $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{U}^\top, \boldsymbol{P}^* = \boldsymbol{U}^*\boldsymbol{U}^{*\top}, \hat{\boldsymbol{P}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top$ as the true projection, the projection estimated by the true task parameters $\boldsymbol{\theta}^*(i)$ and the projection estimated by $\boldsymbol{\theta}(i)$ respectively. For the following parts we need to define the matrix norms: We denote the matrix norm $\|\cdot\|$ as the spectral norm and $\|\cdot\|_F$ as the Frobenius norm. We also require some auxiliary Lemmas:

Lemma B.1.1 (Smale and Zhou (2007)). *Let $\boldsymbol{\theta}^*(1), ..., \boldsymbol{\theta}(t)^* \in \mathbb{R}^d$ be vector valued random variables sampled from a distribution ρ with true mean $\boldsymbol{\mu}$ and $\|\boldsymbol{\theta}^*(i)\| \leq V$, $\forall i \in \{1, ..., t\}$. Then the following holds with probability $1 - \delta$:*

$$\left\|\bar{\boldsymbol{\theta}}^* - \boldsymbol{\mu}\right\| \leq \frac{2\log\left(\frac{2}{\delta}\right)V}{t} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)\mathrm{Var}_{\max}}{t}},$$

with $\mathrm{Var}_{\max} = \mathbb{E}\left[\|\boldsymbol{\Delta}\|^2\right] = \mathrm{tr}(\boldsymbol{\Sigma})$ as the total variance of distribution ρ.

Lemma B.1.2 (Corollary 5.50 of Vershynin (2012)). *Consider a sub-Gaussian distribution in $\mathbb{R}^d$ with true covariance $\boldsymbol{\Sigma}$ and the covariance $\boldsymbol{\Sigma}^*$ estimated from t samples as it was defined above. Let $\delta \in (0, 1)$, then we have with probability $1 - \delta$:*

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^*\| \leq \sqrt{C\frac{\log(2/\delta)}{t}},$$

with C as an absolute constant.

Lemma B.1.3 (Theorem 2 in Yu *et al.* (2015)). *Let $\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d\times d}$ be two symmetric matrices with eigenvalues $\sigma_1 \geq ... \geq \sigma_d$ and $\hat{\sigma}_1 \geq ... \geq \hat{\sigma}_d$ respectively. Fix $1 \leq p \leq d$ and let $\boldsymbol{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_p]$ and $\hat{\boldsymbol{U}} = [\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_p]$ with eigenvectors $\boldsymbol{u}_i$ and $\hat{\boldsymbol{u}}_i$ of matrices $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Assume that the eigengap satisfies $\Delta_\sigma = \sigma_p - \sigma_{p+1} > 0$, then there exists an orthogonal matrix $\boldsymbol{O}$ such that the the following holds:*

$$\left\|\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{O}\right\|_F \leq \frac{\sqrt{8}\min\left(\sqrt{p}\left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\right\|, \left\|\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}\right\|_F\right)}{\Delta_\sigma},$$

Proof of Lemma 4.5.4. For the proof we will use the triangular inequality to express the bound in terms of the true variance along the orthogonal subspace, the projected mean estimation error and the projection estimation error. For the mean estimation error we apply an additional triangular inequality in order to estimate it with respect to the true mean estimation error $\left\|\boldsymbol{P}^\perp(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^*)\right\|$ and the error $\left\|\bar{\boldsymbol{\theta}}^*-\bar{\boldsymbol{\theta}}\right\|$, with the former being a simple concentration bound and the latter being estimated from the oracle inequality for $\boldsymbol{\theta}$. We intend to express the projection error with respect to the estimation error on the covariance matrix. Bounding the term $\left\|\boldsymbol{P}-\hat{\boldsymbol{P}}\right\|$ requires the Davis-Kahan Theorem. Thus we begin the proof:

$$\begin{aligned}
\left\|\hat{\boldsymbol{P}}^\perp(\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}})\right\| &\leq \left\|\hat{\boldsymbol{P}}^\perp-\boldsymbol{P}^\perp\right\|\left\|\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}}\right\| + \left\|\boldsymbol{P}^\perp(\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}})\right\| \\
&\leq \left\|\hat{\boldsymbol{P}}^\perp-\boldsymbol{P}^\perp\right\|\left\|\boldsymbol{\theta}^*-\bar{\boldsymbol{\theta}}\right\| + \left\|\boldsymbol{P}^\perp\boldsymbol{\Delta}\right\| + \left\|\boldsymbol{P}^\perp\boldsymbol{\xi}\right\| \\
&\leq 2V\left\|\hat{\boldsymbol{P}}-\boldsymbol{P}\right\| + \left\|\boldsymbol{P}^\perp\boldsymbol{\Delta}\right\| + \left\|\boldsymbol{P}^\perp\boldsymbol{\xi}\right\|,
\end{aligned}$$

where we used $\boldsymbol{P}_i^\perp - \boldsymbol{P}_j^\perp = \boldsymbol{P}_i - \boldsymbol{P}_j$ for all projection matrices $\boldsymbol{P}_i, \boldsymbol{P}_j$ in the last inequality. We deliver an upper bound on all of these terms separately: The second term being straight forward with $\boldsymbol{P}^\perp$ as the true orthogonal projection:

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{P}^\perp\boldsymbol{\Delta}\right\|^2\right] = \mathrm{Var}_\rho, \tag{B.4}$$

with Var_ρ denoting the low variance of distribution ρ along the orthogonal subspace. This holds simply due to our problem setting.

The last term yields the mean estimation error of tasks along the orthogonal subspace which was similarly discussed in Cella *et al.* (2020):

$$\left\|\boldsymbol{P}^\perp(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}})\right\| \leq \left\|\boldsymbol{P}^\perp(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^*)\right\| + \left\|(\bar{\boldsymbol{\theta}}^*-\bar{\boldsymbol{\theta}})\right\| \tag{B.5}$$

The first term can simply be bounded by a concentration inequality which was also discussed in Lemma 3 of Cella *et al.* (2020) by using Lemma B.1.1 we state with probability of $1-\delta$ that the following holds:

$$\left\|\boldsymbol{P}^\perp(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^*)\right\| \leq \frac{2\log\left(\frac{2}{\delta}\right)}{t} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)\mathrm{Var}_\rho}{t}}.$$

By choosing $\delta = 1/t$ and taking the expectation value with respect to the task distribution, we have:

$$\mathbb{E}_{\boldsymbol{\theta}^*\sim\rho}\left[\left\|\boldsymbol{P}^{\perp}\left(\boldsymbol{\mu}-\bar{\boldsymbol{\theta}}^*\right)\right\|\right] = \mathcal{O}\left(\frac{2\log(2t)}{t}+\sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}\right).$$

We will denote $\epsilon_{\boldsymbol{\mu}} \coloneqq \frac{2\log(2t)}{t}+\sqrt{\frac{2\log(2t)\mathrm{Var}_\rho}{t}}$ for the rest of the proof As for the second term in eq. (B.5), we assume that all previously learnt tasks were running for at least n rounds and use the subscript $i \in \{1,...,t\}$ to refer to a given task:

$$\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}^*-\bar{\boldsymbol{\theta}}\right\| &\leq \max_i \|\boldsymbol{\theta}(i)^*-\boldsymbol{\theta}(i)\| && \text{(B.6)}\\
&\leq \max_i \frac{\|\boldsymbol{\theta}^*(i)-\boldsymbol{\theta}(i)\|_{\boldsymbol{A}_n(i)}}{\sqrt{\lambda_{\min}(\boldsymbol{A}_n(i))}} && \text{(B.7)}\\
&\leq \frac{1}{\sqrt{\log(n)}}\left(\sqrt{d\log\Big(1+\frac{n}{d\lambda}\Big)+\log\left(\frac{1}{\delta^2}\right)}+\sqrt{\lambda}V\right), && \text{(B.8)}
\end{aligned}$$

where we used a linear regression result $\lambda_{\min}(\boldsymbol{A}_n) \geq \log(n)$ from Lai and Wei (1982). For the most general case, we will keep $\lambda_{\min} = \min_i \lambda_{\min}(\boldsymbol{A}_n(i))$ Choosing $\delta = 1/n$, $\lambda = \frac{1}{nV^2}$ and taking the expectation with respect to the arm selection process yields:

$$\mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}^*-\bar{\boldsymbol{\theta}}\right\|\right] \leq \mathcal{O}\left(\frac{1}{\lambda_{\min}}\sqrt{d\log\left(1+\frac{n^2V^2}{d}\right)+2}+\sqrt{\frac{1}{n}}\right). \qquad \text{(B.9)}$$

We denote $\beta_d \coloneqq \frac{1}{\lambda_{\min}}\sqrt{d\log\big(1+\frac{n^2V^2}{d}\big)+2}+\sqrt{\frac{1}{n}}$ We note that this upper bound is independent from the task distribution.

What is left is to upper bound the term $\left\|\boldsymbol{P}-\hat{\boldsymbol{P}}\right\|$:

$$\begin{aligned}
\left\|\boldsymbol{P}-\hat{\boldsymbol{P}}\right\| &= \left\|\boldsymbol{U}\boldsymbol{U}^\top-\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top\right\|\\
&= \left\|\boldsymbol{U}\boldsymbol{U}^\top-\hat{\boldsymbol{U}}\boldsymbol{O}\boldsymbol{O}^\top\hat{\boldsymbol{U}}^\top\right\|\\
&= \left\|\boldsymbol{U}\boldsymbol{U}^\top+\hat{\boldsymbol{U}}\boldsymbol{O}\boldsymbol{U}^\top-\hat{\boldsymbol{U}}\boldsymbol{O}\boldsymbol{U}^\top-\hat{\boldsymbol{U}}\boldsymbol{O}\boldsymbol{O}^\top\hat{\boldsymbol{U}}^\top\right\|\\
&= \left\|\hat{\boldsymbol{U}}\boldsymbol{O}(\boldsymbol{U}^\top-\boldsymbol{O}^\top\hat{\boldsymbol{U}}^\top)+(\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{O})\boldsymbol{U}^\top\right\|\\
&\leq \left\|\hat{\boldsymbol{U}}\boldsymbol{O}(\boldsymbol{U}^\top-\boldsymbol{O}^\top\hat{\boldsymbol{U}}^\top)\right\|+\left\|(\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{O})\boldsymbol{U}^\top\right\|\\
&\leq 2\left\|\boldsymbol{U}-\hat{\boldsymbol{U}}\boldsymbol{O}\right\|_F,
\end{aligned}$$

where we used Cauchy-Schwarz in the last inequality and the fact that $\boldsymbol{O}$ is a

orthogonal matrix and $\boldsymbol{U}^\top \boldsymbol{U} = \hat{\boldsymbol{U}}^\top \hat{\boldsymbol{U}} = \boldsymbol{I}$. Now we are able to apply Lemma B.1.3:

$$\left\| \boldsymbol{P} - \hat{\boldsymbol{P}} \right\| \leq \frac{\sqrt{32} \min\left(\sqrt{p} \left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\|, \left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\|_F \right)}{\Delta_\sigma} \tag{B.10}$$

Using the triangular inequality we bound the term $\left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\|$:

$$\left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \right\| \leq \| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^* \| + \left\| \boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}} \right\|. \tag{B.11}$$

The first term in eq. (B.11) is a simple concentration inequality for covariance matrices. Using Lemma B.1.2 we have with probability $1 - \delta$:

$$\| \boldsymbol{\Sigma}^* - \boldsymbol{\Sigma} \| \leq \sqrt{\frac{C \log\left(\frac{2}{\delta}\right)}{t}},$$

with an absolute constant C. Setting $\delta = 1/t$ and taking the expectation value yields:

$$\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho} \left[\| \boldsymbol{\Sigma}^* - \boldsymbol{\Sigma} \| \right] = \mathcal{O}\left(\sqrt{\frac{C \log(2t)}{t}} \right).$$

We denote $\epsilon_{\boldsymbol{\Sigma}} \coloneqq \sqrt{\frac{C \log(2t)}{t}}$. Finally we need to bound $\left\| \boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}} \right\|$. We denote $\boldsymbol{\Theta}^* = \left[\left(\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^* \right)^\top \right]_{\{i \in 1, \dots, t\}}$ and $\hat{\boldsymbol{\Theta}} = \left[\left(\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}} \right)^\top \right]_{i \in \{1, \dots, t\}}$, with vertically concatenated vectors, such that we have:

$$\begin{aligned}
\left\| \boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}} \right\| &\leq \left\| \boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}} \right\|_F \\
&= \frac{1}{t} \left\| \boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}}^\top \hat{\boldsymbol{\Theta}} \right\|_F \\
&= \frac{1}{t} \left\| \boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^* - \boldsymbol{\Theta}^{*\top} \hat{\boldsymbol{\Theta}} + \boldsymbol{\Theta}^{*\top} \hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}^\top \hat{\boldsymbol{\Theta}} \right\|_F \\
&= \frac{1}{t} \left\| \boldsymbol{\Theta}^{*\top} \left(\boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}} \right) + \left(\boldsymbol{\Theta}^{*\top} - \hat{\boldsymbol{\Theta}}^\top \right) \hat{\boldsymbol{\Theta}} \right\|_F \\
&\leq \frac{1}{t} \left(\| \boldsymbol{\Theta}^* \|_F + \left\| \hat{\boldsymbol{\Theta}} \right\|_F \right) \left\| \boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}} \right\|_F
\end{aligned}$$

We can further bound this while also taking the expectation, using:

$$\mathbb{E}\left[\left\| \boldsymbol{\Theta}^* - \hat{\boldsymbol{\Theta}} \right\|_F \right] \leq \mathbb{E}\left[\sqrt{t \beta_d^2} + \sqrt{\sum_{i=1}^{t} \| \boldsymbol{\theta}^*(i) - \boldsymbol{\theta}(i) \|^2} \right] \leq 2\sqrt{t} \beta_d.$$

where we used the result of eq. (B.9). The same estimation can be done for the term $\|\boldsymbol{\Theta}^*\|_F + \left\|\hat{\boldsymbol{\Theta}}\right\|_F$:

$$\|\boldsymbol{\Theta}^*\|_F + \left\|\hat{\boldsymbol{\Theta}}\right\|_F \leq \sqrt{t}\left(\max_i \left\|\boldsymbol{\theta}^*(i) - \bar{\boldsymbol{\theta}}^*\right\| + \max_i \left\|\boldsymbol{\theta}(i) - \bar{\boldsymbol{\theta}}\right\|\right) \leq 4\sqrt{t}V$$

Thus we conclude:

$$\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}\right\|\right] \leq 8V\beta_d$$

Inserting the results into eq. (B.10) gives:

$$\mathbb{E}\left[\left\|\boldsymbol{P} - \hat{\boldsymbol{P}}\right\|\right] \leq \sqrt{32p}\frac{8V\beta_d + \epsilon_{\boldsymbol{\Sigma}}}{\Delta_\sigma} \tag{B.12}$$

After estimation of every term of our original expression we can summarize it by taking the expectation and applying Jensen's inequality:

$$\mathbb{E}_{\theta^*\sim\rho}\left[\mathbb{E}\left[\left\|\hat{\boldsymbol{P}}^{\perp}(\boldsymbol{\theta}^* - \bar{\boldsymbol{\theta}})\right\|\right]\right] = \mathcal{O}\left(\sqrt{\mathrm{Var}_\rho + \beta_d^2\left(1 + 64\sqrt{2p}\frac{V^2}{\Delta_\sigma}\right)^2 + \epsilon_{\boldsymbol{\mu}}^2 + \frac{128p\epsilon_{\boldsymbol{\Sigma}}^2 V^2}{\Delta_\sigma^2}}\right) \tag{B.13}$$

□

Lemma B.1.4. *(Abbasi-Yadkori* et al.*, 2011, Lemma 11) Let $\boldsymbol{x}_{a_k}$ be a sequence in $\mathbb{R}^d$ with $\|\boldsymbol{x}_{a_k}\| \leq 1$ and $\boldsymbol{B}$ defined as usual. Then we have:*

$$\sum_{k=1}^{n} \left\|\boldsymbol{x}_{a_{k-1}}\right\|^2_{\boldsymbol{B}_{k-1}^{-1}} \leq 2S_{k-1}^{\lambda_1,\lambda_2}$$

Proof of Theorem 4.5.5. First we denote $\xi \coloneqq \|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}}\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_{\boldsymbol{B}}$ as exploration

term. Then we continue by estimating the peudo regret $R(n)$:

$$
\begin{aligned}
R(n) &= \sum_{k=1}^{n} \left(\boldsymbol{x}_{a_{k-1}^*} - \boldsymbol{x}_{a_{k-1}} \right)^\top \boldsymbol{\theta}^* \\
&\leq \sum_{k=1}^{n} \boldsymbol{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} - \boldsymbol{x}_{a_{k-1}}^\top \boldsymbol{\theta}^* \\
&\leq \sum_{k=1}^{n} \boldsymbol{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} - \boldsymbol{x}_{a_{k-1}}^\top \hat{\boldsymbol{\theta}}_{k-1} + \xi_{k-1} \\
&= \sum_{k=1}^{n} 2\xi_{k-1} \\
&\leq \sum_{k=1}^{n} \left(\sqrt{p \log\left(\frac{1}{\delta} + \frac{k-1}{p\lambda_2 \delta} \right) + q \log\left(\frac{1}{\delta} + \frac{k-1}{q\lambda_1 \delta} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \left\| \boldsymbol{x}_{a_{k-1}} \right\|_{\boldsymbol{B}_{k-1}^{-1}} \\
&\leq \left(\sqrt{p \log\left(\frac{1}{\delta} + \frac{n}{p\lambda_2 \delta} \right) + q \log\left(\frac{1}{\delta} + \frac{n}{q\lambda_1 \delta} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \\
&\quad \sqrt{n \sum_{k=1}^{n} \left\| \boldsymbol{x}_{a_{k-1}} \right\|_{\boldsymbol{B}_{k-1}^{-1}}^2} \\
&\leq \left(\sqrt{p \log\left(1 + \frac{n}{p\lambda_2} \right) + q \log\left(1 + \frac{n}{q\lambda_1} \right) + \log\left(\frac{1}{\delta^2} \right)} + \sqrt{\lambda_2} V + \frac{\lambda_1}{\sqrt{\lambda_2}} W \right) \\
&\quad \sqrt{2n \left(p \log\left(1 + \frac{n}{p\lambda_2} \right) + q \log\left(1 + \frac{n}{q\lambda_1} \right) \right)}
\end{aligned}
$$

The first and second inequality make use of the OFUL principle and the definition of the UCB function. We used Lemma 4.5.3 in the third inequality and Lemma B.1.4 in the last inequality. This regret holds with probability $1 - \delta$.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho} \left[\mathbb{E} \left[R(n) \right] \right] \leq & \left(\sqrt{p \log\left(1 + \frac{nV^2}{p} \right) + q \log\left(1 + \frac{n\sqrt{Y}}{q} \right) + \log\left(\frac{1}{\delta^2} \right)} + 1 + V \right) \\
& \sqrt{2n \left(p \log\left(1 + \frac{nV^2}{p} \right) + q \log\left(1 + \frac{n\sqrt{Y}}{q} \right) \right)}
\end{aligned}
$$

We obtain the final results by setting $\delta = 1/n$, take the expectation value, followed by an additional expectation value with respect to the task distribution:

$\mathbb{E}_{\boldsymbol{\theta}^* \sim \rho}[\mathbb{E}[R(n)]]$, setting $\lambda_1 = \frac{1}{\sqrt{Y}}$, $\lambda_2 = \frac{1}{V^2}$ and application of Jensen's inequality. □

B.2 Proof of Theorem 4.6.4

The following proofs are adapted from Agrawal and Goyal (2013) and are required to finish the proof on the regret bound. Before proceeding, we define the concept of a saturated arm, which is basically a measurement of the required exploration for any arm.

Definition B.2.1. *We call an arm a saturated if* $g_n \|\boldsymbol{x}_a\|_{\boldsymbol{B}^{-1}} < l_n \|\boldsymbol{x}_{a^*}\|_{\boldsymbol{B}^{-1}}$ *and unsaturated otherwise, with* $g_n = \sqrt{2d + 6\log(n)}v + l_n$*. The set of saturated arms at round* k *is denoted as* $\mathcal{C}_k$*.*

We will also utilize the following Lemma from Hsu *et al.* (2012), which is a special case of the inequality in Hanson and Wright (1971):

Lemma B.2.2 (Proposition 1.1 in Hsu *et al.* (2012))**.** *Let* $\boldsymbol{x} \in \mathbb{R}^d$ *be a* d*-dimensional standard normal variable and* $\boldsymbol{C} \in \mathbb{R}^{d \times d}$ *a matrix. Then we have for all* $t > 0$*:*

$$\Pr\left(\|\boldsymbol{C}\boldsymbol{x}\|^2 > \operatorname{tr}\big(\boldsymbol{C}^\top \boldsymbol{C}\big) + 2\sqrt{\operatorname{tr}\Big(\big(\boldsymbol{C}^\top \boldsymbol{C}\big)^2\Big)t} + 2\big\|\boldsymbol{C}^\top \boldsymbol{C}\big\|t\right) \leq e^{-t}$$

Proof of Lemma 4.6.3. The probability of event E_r is determined using Lemma 4.5.3: we have with probability $1 - \delta$:

$$\begin{aligned}
|\boldsymbol{x}_a^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right)| &\leq \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}} \left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{B}_k} \\
&\leq \|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}} \left(\sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{1}{\delta^2}\right)} + \frac{\lambda_2}{\sqrt{\lambda_1}} W + \sqrt{\lambda_1} V\right),
\end{aligned}$$

by substituting $\delta \to \frac{\delta}{n^2}$ and further upper bounding $S_k^{\lambda_1,\lambda_2}$ we get:

$$\begin{aligned}
\sqrt{S_k^{\lambda_1,\lambda_2} + \log\left(\frac{n^2}{\delta^2}\right)} &= \sqrt{p\log\left(1 + \frac{k}{p\lambda_2}\right) + q\log\left(1 + \frac{k}{q\lambda_1}\right) + \log\left(\frac{n^2}{\delta^2}\right)} \\
&\leq \sqrt{p\log\left(n\left(\frac{n}{\delta}\right)^{2/d}\right) + q\log\left(n\left(\frac{n}{\delta}\right)^{2/d}\right)} \\
&\leq \sqrt{\log\left(\frac{1}{\delta}\right)(d+2)\log(n)},
\end{aligned}$$

and therefore we have:

$$|\boldsymbol{x}_a^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right)| \leq \left(\sqrt{\log\left(\frac{1}{\delta}\right)(d+2)\log(n)} + \frac{\lambda_2}{\sqrt{\lambda_1}}W + \sqrt{\lambda_1}V\right)\|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}}$$
$$\leq \sqrt{2\log\left(\frac{1}{\delta}\right)(d+2)\log(n) + 2K^2}\|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}},$$

with $K = \frac{\lambda_2}{\sqrt{\lambda_1}}W + \sqrt{\lambda_1}V$. Since we substituted $\delta \to \frac{\delta}{n^2}$, this event has a probability of at least $1 - \frac{\delta}{n^2}$.
For proof of the bound on the probability of event E_θ we have for all $a \in \mathcal{A}_k$:

$$\left|\boldsymbol{x}_a^\top \left(\hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k\right)\right| = \left|\boldsymbol{x}_a^\top \boldsymbol{B}_k^{-\frac{1}{2}} \boldsymbol{B}_k^{\frac{1}{2}} \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right|$$
$$\leq v\sqrt{\boldsymbol{x}_a^\top \boldsymbol{B}_k^{-1}\boldsymbol{x}_a}\left\|\frac{1}{v}\boldsymbol{B}_k^{\frac{1}{2}} \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right\|$$

.

By definition, the term $\frac{1}{v}\boldsymbol{B}_k^{\frac{1}{2}} \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)$ is d-dimensional standard normal variable, such that we can apply Lemma B.2.2, where we set $\boldsymbol{C} = \boldsymbol{I}$ and $t = 2\log(n)$:

$$\Pr\left(\left\|\frac{1}{v}\boldsymbol{B}_k^{\frac{1}{2}} \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)\right\| > \sqrt{d + \sqrt{8d\log(n)} + 4\log(n)}\right) \leq \frac{1}{n^2}.$$

Thus the following inequality holds with probability of at least $1 - \frac{1}{n^2}$ for all $a \in \mathcal{A}_k$:

$$\left|\boldsymbol{x}_a^\top \left(\hat{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\theta}}_k\right)\right| \leq v\|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}}\sqrt{d + \sqrt{8d\log(n)} + 4\log(n)}$$
$$\leq v\|\boldsymbol{x}_a\|_{\boldsymbol{B}_k^{-1}}\sqrt{2d + 6\log(n)},$$

where we used the inequality of arithmetic and geometric means in the last step. □

Lemma B.2.3. *For any filtration $\mathcal{F}_{k-1}$ such that E_r is true, we have:*

$$\Pr(\boldsymbol{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k > \boldsymbol{x}_{a_k^*}^\top \boldsymbol{\theta}^* + l_n\left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}) \geq c_n$$

and:

$$\Pr(a_k \in \mathcal{C}_k|\mathcal{F}_{k-1}) \leq \frac{1}{c_n}\Pr(a_k \notin \mathcal{C}_k|\mathcal{F}_{k-1}) + \frac{1}{c_n n^2},$$

with $c_n = \frac{1}{4e\sqrt{\pi n^\alpha}}$.

Proof. Assuming the event E_r holds and $\boldsymbol{x}_{a^*}^\top \tilde{\boldsymbol{\theta}}$ is a Gaussian random variable with mean $\boldsymbol{x}_{a^*}^\top \hat{\boldsymbol{\theta}}$ and variance $v\|x_{a^*}\|_{\boldsymbol{B}^{-1}}$, we can apply the anti-concentration inequality such that:

$$\begin{aligned}
&\Pr(\boldsymbol{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k \geq \boldsymbol{x}_{a_k^*}^\top \hat{\boldsymbol{\theta}}_k + l_n \left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}} | \mathcal{F}_{k-1}) \\
&= \Pr\left(\frac{\boldsymbol{x}_{a_k^*}^\top \left(\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k\right)}{v\left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}} \geq \frac{\boldsymbol{x}_{a_k^*}^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right) + l_n \left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}}{v\left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}} \middle| \mathcal{F}_{k-1} \right) \geq \frac{1}{4\sqrt{\pi}} \exp\bigl(-Z^2\bigr),
\end{aligned}$$

with

$$\begin{aligned}
Z &= \frac{\boldsymbol{x}_{a_k^*}^\top \left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right) + l_n \left\|x_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}}{v\left\|x_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}} \\
&\leq \frac{2l_n \left\|x_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}}{v\left\|x_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}}} \\
&\leq \frac{2\sqrt{2\log\bigl(\frac{1}{\delta}\bigr)(d+2)\log(n) + 2K^2}}{4\sqrt{\log\bigl(\frac{1}{\delta}\bigr)\frac{d+2}{\alpha}}} \\
&\leq \sqrt{\frac{\alpha}{2}\log(n) + \frac{\alpha K^2}{8\log\bigl(\frac{1}{\delta}\bigr)(d+2)}} \\
&\leq \sqrt{\frac{\alpha}{2}\log(n) + 1}.
\end{aligned}$$

Thus we have

$$\Pr(\boldsymbol{x}_{a_k^*}^\top \tilde{\boldsymbol{\theta}}_k \geq \boldsymbol{x}_{a_k^*}^\top \hat{\boldsymbol{\theta}}_k + l_n \left\|\boldsymbol{x}_{a_k^*}\right\|_{\boldsymbol{B}_k^{-1}} | \mathcal{F}_{k-1}) \geq \frac{1}{4e\sqrt{\pi n^\alpha}}$$

The proof of the second inequality is provided in Lemma 3 of Agrawal and Goyal (2013).

□

Lemma B.2.4. *Let* $\text{regret}_k = (\boldsymbol{x}_{a^*}^\top - \boldsymbol{x}_{a_k}^\top)\boldsymbol{\theta}^*$ *be defined as the instantaneous regret at round* k *and* $\text{regret}'_k = \text{regret}_k I(E_r)$.

Define

$$X_k = \text{regret}'_{\text{k}} - \frac{g_n}{c_n} I\left(a_k \notin \mathcal{C}\right) \|x_{a^*}\|_{\boldsymbol{B}_k^{-1}} - \frac{2g_n}{c_n n^2} - \frac{2g_n^2}{l_n} \|x_{a_k}\|_{\boldsymbol{B}_k^{-1}}$$

and

$$Y_k = \sum_{t=1}^{k} X_k,$$

then $(Y_k; k = 1, ..., n)$ *is a super-martingale process with respect to filtration* $\mathcal{F}_{k-1}$.

Proof. The proof is provided in Lemma 4 of Agrawal and Goyal (2013). □

Proof of Theorem 4.6.4. Each value in X_k is bounded by $\frac{2g_n^2}{c_n l_n}$ which implies a bounded difference on the super-martingale Y_k with $|Y_k - Y_{k-1}| \leq \frac{8g_n^2}{c_n l_n}$, allowing us to apply Azuma-Hoeffding's inequality during the proof. Thus we have with probability $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\sum_{k=1}^{n} \text{regret}'_k &\leq \sum_{k=1}^{n} \left(\frac{g_n}{c_n} I(a_k \notin \mathcal{C}_k \|x_{a^*}\|_{\boldsymbol{B}_k^{-1}}) \right) + \frac{2g_n}{c_n n^2} + \frac{2g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\boldsymbol{B}_k^{-1}} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \sum_{k=1}^{n} \left(\frac{g_n^2}{c_n l_n} I(a_k \notin \mathcal{C}_k \|x_{a_k}\|_{\boldsymbol{B}_k^{-1}}) \right) + \frac{2g_n}{c_n n^2} + \frac{2g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\boldsymbol{B}_k^{-1}} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \frac{3g_n^2}{c_n l_n} \sum_{k=1}^{n} \|x_{a_k}\|_{\boldsymbol{B}_k^{-1}} + \frac{2g_n}{c_n n^2} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&\leq \frac{3g_n^2}{c_n l_n} \sqrt{2n S_n^{\lambda_1, \lambda_2}} + \frac{2g_n}{c_n n^2} + \frac{8g_n^2}{c_n l_n} \sqrt{2n \log\left(\frac{2}{\delta}\right)} \\
&= \frac{g_n^2}{c_n l_n} \left(\sqrt{18n S_n^{\lambda_1, \lambda_2}} + \sqrt{128n \log\left(\frac{2}{\delta}\right)} \right) + \frac{2g_n}{c_n n^2} \\
&= \left(\frac{l_n + 2\sqrt{2d + 6\log(n)} v + (2d + 6\log(n)) v^2 / l_n}{c_n} \right) \\
&\times \left(\sqrt{18n S_n^{\lambda_1, \lambda_2}} + \sqrt{128n \log\left(\frac{2}{\delta}\right)} \right) + \frac{2g_n}{c_n n^2} \\
&= \mathcal{O}\Bigg(\left(2\sqrt{\frac{2d^2 \log\left(\frac{1}{\delta}\right) + 6d \log\left(\frac{1}{\delta}\right) \log(n)}{\alpha}} + \frac{2d^{\frac{3}{2}} \sqrt{\log\left(\frac{1}{\delta}\right)}}{\alpha \sqrt{\log(n)}} + \frac{6\sqrt{d \log\left(\frac{1}{\delta}\right) \log(n)}}{\alpha} \right) \\
&\times \sqrt{n^{1+\alpha} S_n^{\lambda_1, \lambda_2}} \Bigg)
\end{aligned}$$

We used our definition for saturated arms in the second inequality and Lemma B.1.4 in the fourth inequality. Now similar as done in Theorem 4.5.5 we set $\delta = \frac{1}{n}$ and take the expectation value. Additionally we set $\alpha = \frac{1}{\log(n)}$:

$$\mathbb{E}\left[\sum_{k=1}^{n} \text{regret}'_k\right] = \mathcal{O}\left(\left(d^{\frac{3}{2}}\log(n) + \sqrt{d}\log(n)^2\right)\sqrt{nS_n^{\lambda_1,\lambda_2}}\right)$$

Inserting $\lambda_2 = 1/V^2$ and $\lambda_1 = \frac{1}{\sqrt{Y}}$, while taking the second expectation value with respect to the task distribution and applying Jensen's inequality, gives the final result:

$$\mathcal{R}(n) = \mathcal{O}\left(\left(d^{\frac{3}{2}}\log(n) + \sqrt{d}\log(n)^2\right)\sqrt{n\left(p\log\left(1+\frac{nV^2}{p}\right) + q\log\left(1+\frac{n\sqrt{Y}}{q}\right)\right)}\right)$$

□

Appendix C

Cluster Agnostic Network Lasso Bandits

C.1 Some helper results

Proposition C.1.1 (Bounds on norms of matrix products)**.** *Let* $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ *and* $\boldsymbol{N} \in \mathbb{R}^{n\times p}$*. Then*

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

Proof.

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

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

Second inequality We have

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

Third inequality We have

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

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

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

Fourth inequality We have

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

□

Proposition C.1.2 (Decomposition of a signal over a graph)**.** *For any $\mathcal{C} \in \mathcal{P}$*

- *Let $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ be a graph signal. Let us denote by $\boldsymbol{Z}_{\mathcal{C}}$ the signal obtained from $\boldsymbol{Z}$ by setting rows of vertices outside of $\mathcal{C}$ to zeros, and let $\boldsymbol{Z}_{|\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}| \times d}$ be the signal obtained from $\boldsymbol{Z}_{\mathcal{C}}$ by removing the rows of vertices outside of $\mathcal{C}$. Also, let $\boldsymbol{B}_{|\mathcal{C}} \in \mathbb{R}^{|\mathcal{E}_{\mathcal{C}}| \times |\mathcal{C}|}$ be the matrix obtained by taking $\boldsymbol{B}_{\mathcal{C}}$, and removing rows of edges that link $\mathcal{C}$ to its outside, and the resulting null columns. It is clear that*
$$\boldsymbol{B}_{\mathcal{C}} \boldsymbol{Z} = \boldsymbol{B}_{\mathcal{C}} \boldsymbol{Z}_{\mathcal{C}} = \boldsymbol{B}_{|\mathcal{C}} \boldsymbol{Z}_{|\mathcal{C}} \tag{C.1}$$
- *Let $\boldsymbol{Q}_{\mathcal{C}} \coloneqq \boldsymbol{B}_{\mathcal{C}}^\dagger \boldsymbol{B}_{\mathcal{C}}$. Then*
$$\boldsymbol{I}_{|\mathcal{V}|} = \sum_{\mathcal{C} \in \mathcal{P}} \boldsymbol{J}_{\mathcal{C}} + \boldsymbol{Q}_{\mathcal{C}} \tag{C.2}$$
$$\boldsymbol{Q}_{\partial \mathcal{P}^c} \coloneqq \boldsymbol{B}_{\partial \mathcal{P}^c}^\dagger \boldsymbol{B}_{\partial \mathcal{P}^c} = \sum_{\mathcal{C} \in \mathcal{P}} \boldsymbol{Q}_{\mathcal{C}} \tag{C.3}$$
where $\boldsymbol{J}_{\mathcal{C}} = \frac{\mathbf{1}_{\mathcal{C}} \mathbf{1}_{\mathcal{C}}^\top}{|\mathcal{C}|}$, $\boldsymbol{Q}_{\mathcal{C}} = \boldsymbol{B}_{\mathcal{C}}^\dagger \boldsymbol{B}_{\mathcal{C}} \qquad \forall \mathcal{C} \in \mathcal{P}$ and $\boldsymbol{Q}_{\partial \mathcal{P}^c} \coloneqq \boldsymbol{B}_{\partial \mathcal{P}^c}^\dagger \boldsymbol{B}_{\partial \mathcal{P}^c}$.

While $\sum_{\mathcal{C} \in \mathcal{P}} \boldsymbol{J}_{\mathcal{C}}$ projects each entry of a graph signal onto the mean vector value of its respective cluster, its residual $\boldsymbol{Q}_{\partial \mathcal{P}^c}$ 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 $\boldsymbol{B}_{|\mathcal{C}}^\dagger$ the pseudo-inverse of $\boldsymbol{B}_{|\mathcal{C}}$ it is a well-known linear algebra result that the matrix $Q_{|\mathcal{C}} \coloneqq \boldsymbol{B}_{|\mathcal{C}}^\dagger \boldsymbol{B}_{|\mathcal{C}}$ is the projector onto the null space of $\boldsymbol{B}_{|\mathcal{C}}$. Since $\mathcal{C}$ is connected, the null space of $\boldsymbol{B}_{|\mathcal{C}}$ 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 null space

is $\boldsymbol{J}_{|\mathcal{C}|} \coloneqq \frac{\mathbf{1}_{|\mathcal{C}|}\mathbf{1}_{|\mathcal{C}|}}{|\mathcal{C}|}$, we deduce that

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

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

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

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

Proposition C.1.3 (On the minimum topological centrality index of a graph vertex)**.** *Let $\mathcal{G}$ be a connected graph with incidence matrix $\boldsymbol{B}$ and vertex set size N, and let $\boldsymbol{L} \coloneqq \boldsymbol{B}^{\top}\boldsymbol{B}$. Let $c(\mathcal{G})$ denote the minimum value of inverses of diagonal element of $\boldsymbol{L}^{\dagger}$, called its minimum topological centrality index. Also let $a(\mathcal{G})$ be its algebraic connectivity, defined as the minimum non null eigenvalue of $\boldsymbol{L}$. Then*

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

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

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

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 $\boldsymbol{L}_{\text{comp}}$ be the Laplacian of complete graph built on the vertices of $\mathcal{G}$. Then we have $\boldsymbol{L}_{comp} = N(\boldsymbol{I}_N - \boldsymbol{J}_N)$, where J is the square matrix of dimension N having $1/N$ as entries. From Fontan and Altafini (2021, Lemma 4), we have

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

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

On the other hand, $\boldsymbol{L} \preccurlyeq \boldsymbol{L}_{\text{comp}}$ Hence, by Fontan and Altafini (2021, lemma 4) we have for any $u \neq 0$

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

This implies that the maximum diagonal entry of $\boldsymbol{L}^{\dagger}$ is at least equal to that of $\boldsymbol{L}_{\text{comp}}^{\dagger}$, *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 $\boldsymbol{B} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ corresponding to an arbitrary orientation under the following form

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

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 $\boldsymbol{E} \coloneqq \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ as the error signal and its rows by $\{\boldsymbol{\epsilon}_m\}_{m=1}^{|\mathcal{V}|}$.

While $\sum_{k=1}^{C} \boldsymbol{J}_{\mathcal{C}}$ projects each entry of a graph signal onto the mean vector value of its respective cluster, its residual $\boldsymbol{Q}_{\partial\mathcal{P}^c}$ 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 $\boldsymbol{K} \in \mathbb{R}^{|\mathcal{V}|\times d}$ as the matrix of vertically concatenated row vectors $\boldsymbol{\eta}_m^{\top}\boldsymbol{X}_m$.

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

Table C.1: Notation table for time dependent quantities.

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
$\boldsymbol{\epsilon}_m \in \mathbb{R}^d$	estimation error for user/bandit $m : \hat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m$
$\boldsymbol{E} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d}$	vertical concatenation of row vectors $\boldsymbol{\epsilon}_m$
$\boldsymbol{\eta}_m \in \mathbb{R}^{\lvert\mathcal{T}_m(t)\rvert}$	vector of sub-Gaussian noise of user m
$\boldsymbol{x}(t) \in \mathbb{R}^d$	context vector received at time t
$m(t) \in \mathbb{N}$	user at time t
$\boldsymbol{X}_m \in \mathbb{R}^{\lvert\mathcal{T}_m(t)\rvert \times d}$	data matrix of user m
$\boldsymbol{X} \in \mathbb{R}^{t \times d}$	data matrix of context vectors of all users
$\boldsymbol{A}_m \in \mathbb{R}^{d \times d}$	$\boldsymbol{X}_m^\top \boldsymbol{X}_m$ (potentially associated to time t)
$\boldsymbol{A}_\mathcal{V} \in \mathbb{R}^{d\lvert\mathcal{V}\rvert \times d\lvert\mathcal{V}\rvert}$	$\mathrm{diag}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_m)$
$\boldsymbol{K} \in \mathbb{R}^{\lvert\mathcal{V}\rvert \times d}$	matrix of vertically concatenated row vectors $\boldsymbol{\eta}_m^\top \boldsymbol{X}_m$

N.B. Except for the results concerning the regret bound, we consider the case $\kappa \geq 0$ rather than $\kappa \geq 1$ in our proofs.

C.2.2 Oracle inequality

In this section, we present all intermediary theoretical results leading to Theorem C.2.7 stating the oracle inequality. To reduce clutter, we omit the dependence on t of several quantities. For instance, we write α and $\hat{\boldsymbol{\Theta}}$ instead of $\alpha(t)$ and $\hat{\boldsymbol{\Theta}}(t)$.

Definition C.2.1 (Restricted Eigenvalue (RE) condition and norm, generalization of Definition 5.5.3)**.** *Let $\{\boldsymbol{M}_i\}_{i=1}^{\lvert\mathcal{V}\rvert} \subset \mathbb{R}^{d\times d}$ be a set of positive semi-definite matrices. We say that the matrix $\boldsymbol{M}_\mathcal{V} \coloneqq \mathrm{diag}(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_{\lvert\mathcal{V}\rvert})$ verifies the restricted eigenvalue condition with constants $\kappa \geq 0$ and $\phi > 0$ if*

$$\phi^2 \|\boldsymbol{Z}\|_{\mathrm{RE}}^2 \leq \sum_{i\in\mathcal{V}} \|\boldsymbol{z}_i\|_{\boldsymbol{M}_i}^2 \quad \forall \boldsymbol{Z} \in \mathcal{S} \text{ with rows } \{\boldsymbol{z}_i\}_{i\in\mathcal{V}},$$

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

$$\mathcal{S} \coloneqq \{\boldsymbol{Z} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times d}; a_1\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right) \|\boldsymbol{Z}\|_{\partial\mathcal{P}^c} \leq a_2\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right) \left\|\overline{\boldsymbol{Z}}_\mathcal{P}\right\|_F + (1-\kappa)^+ \|\boldsymbol{Z}\|_{\partial\mathcal{P}}\},$$

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

and the RE semi-norm is defined by $\|\boldsymbol{Z}\|_{\mathrm{RE}} \coloneqq \left\|\overline{\boldsymbol{Z}}_\mathcal{P}\right\|_F \vee (1-\kappa)^+ \left\|\boldsymbol{B}_{\partial\mathcal{P}}^\dagger \boldsymbol{B}_{\partial\mathcal{P}} \boldsymbol{Z}\right\|$ We

Table C.2: Notation table for time independent quantities.

Notation	Meaning
$\mathcal{V}$	set of graph vertices
$\mathcal{E}$	set of graph edges
$\boldsymbol{B}_I \in \mathbb{R}^{\lvert\mathcal{E}\rvert\times\lvert\mathcal{V}\rvert}, I \subseteq \mathcal{E}$	Graph incidence Matrix obtained by setting rows of edges outside I to zeros
$\boldsymbol{B}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{E}\rvert\times\lvert\mathcal{V}\rvert}$	cf. Definition 5.5.1
$\boldsymbol{L} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	$\boldsymbol{B}^\top\boldsymbol{B}$
$\boldsymbol{\theta}_m \in \mathbb{R}^d$	true preference vector of user/bandit m
$\boldsymbol{\Theta} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times d}$	matrix of true vertically concatenated row preferences 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 in $\mathcal{C}$ restricted to the graph in $\mathcal{C}$
$w(\partial\mathcal{P})$	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
$\lVert\cdot\rVert_{\boldsymbol{A}}$	Semi-norm defined by PSD matrix $\boldsymbol{A}$: $\lVert\boldsymbol{x}\rVert_{\boldsymbol{A}}^2 \coloneqq \boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$
$\lVert\cdot\rVert_F$	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 signal over edges of I
$\boldsymbol{A}^\dagger$	Moore-Penrose pseudo-inverse of matrix $\boldsymbol{A}$
vec	vectorization operator consisting in concatenating the columns vertically
$\otimes$	Kronecker product
$\mathbf{1}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$	Vector with entries corresponding to vertices in $\mathcal{C}$ equal to 1 and 0 elsewhere
$\boldsymbol{J}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	equal to $\frac{\mathbf{1}_{\mathcal{C}}\mathbf{1}_{\mathcal{C}}^\top}{\lvert\mathcal{C}\rvert}$
$\boldsymbol{Q}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}$	equal to $\boldsymbol{B}_{\mathcal{C}}^\dagger\boldsymbol{B}_{\mathcal{C}}$
$\boldsymbol{Q}_I \in \mathbb{R}^{\lvert\mathcal{V}\rvert\times\lvert\mathcal{V}\rvert}, I \subseteq \mathcal{E}$	equal to $\boldsymbol{B}_I^\dagger\boldsymbol{B}_I$
$\boldsymbol{e}_k$	elementary vectors of dimension depending on the context
σ	Sub-Gaussianity constant / variance proxy

have the structure dependent unknown constants ψ and κ, for which we assume they guarantee $a_1\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right) > 0$.

Proof of Proposition 5.5.6. Let $\boldsymbol{Z} = \mathbf{1}_{\mathcal{C}}\boldsymbol{v}^\top$ be a constant per cluster signal, with $\mathbf{1}_{\mathcal{C}} \in \mathbb{R}^{\lvert\mathcal{V}\rvert}$ as indicator vector with the ith entry equal to 1 if $i \in \mathcal{C}$ and 0 otherwise.

Then $\boldsymbol{Z}$ is contained in any cone $\mathcal{S}$ defined in the RE condition and we have:

$$\begin{aligned}\left\|\overline{\boldsymbol{Z}_{\mathcal{C}}}\right\|_F^2 = \|\boldsymbol{Z}\|_F^2 &= \left\|\mathbf{1}_{\mathcal{C}}\boldsymbol{v}^\top \boldsymbol{v}\mathbf{1}_{\mathcal{C}}^\top\right\| \\ &= \mathbf{1}_{\mathcal{C}}^\top \mathbf{1}_{\mathcal{C}} \|\boldsymbol{v}\|^2 \\ &= |\mathcal{C}| \|\boldsymbol{v}\|^2\end{aligned}$$

For the right hand side of the RE condition we have:

$$\begin{aligned}\sum_{i\in\mathcal{V}} \|\boldsymbol{z}_i\|_{\boldsymbol{M}_i}^2 &= \operatorname{vec}(\boldsymbol{Z}^\top)^\top \boldsymbol{M} \operatorname{vec}(\boldsymbol{Z}^\top) \\ &= \operatorname{vec}(\mathbf{1}_{\mathcal{C}}\boldsymbol{v}^\top)^\top \boldsymbol{M} \operatorname{vec}(\mathbf{1}_{\mathcal{C}}\boldsymbol{v}^\top) \\ &= (\mathbf{1}_{\mathcal{C}} \otimes \boldsymbol{v})^\top \left(\sum_{i\in\mathcal{V}} \boldsymbol{e}_i\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i\right) (\mathbf{1}_{\mathcal{C}} \otimes \boldsymbol{v}) \\ &= \left(\mathbf{1}_{\mathcal{C}}^\top \otimes \boldsymbol{v}^\top\right) \left(\sum_{i\in\mathcal{V}} \boldsymbol{e}_i\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i\right) (\mathbf{1}_{\mathcal{C}} \otimes \boldsymbol{v}) \\ &= \sum_{i\in\mathcal{V}} \mathbf{1}_{\mathcal{C}}^\top \boldsymbol{e}_i\boldsymbol{e}_i^\top \mathbf{1}_{\mathcal{C}} \otimes \boldsymbol{v}^\top \boldsymbol{M}_i \boldsymbol{v} \\ &= \sum_{i\in\mathcal{C}} \boldsymbol{v}^\top \boldsymbol{M}_i \boldsymbol{v} = \boldsymbol{v}^\top \left(\sum_{i\in\mathcal{C}} \boldsymbol{M}_i\right) \boldsymbol{v}\end{aligned}$$

Plugging the results into the RE condition, we get:

$$\begin{aligned}&\implies \quad \phi^2 |\mathcal{C}| \|\boldsymbol{v}\|^2 \leq \boldsymbol{v}^\top \left(\sum_{i\in\mathcal{C}} \boldsymbol{M}_i\right) \boldsymbol{v} \\ &\implies \quad \phi^2 \leq \frac{\boldsymbol{v}^\top \left(\sum_{i\in\mathcal{C}} \boldsymbol{M}_i\right) \boldsymbol{v}}{\|\boldsymbol{v}\|^2 |\mathcal{C}|} \\ &\implies \quad \phi \leq \sqrt{\lambda_{\min}\left(\frac{\sum_{i\in\mathcal{C}} \boldsymbol{M}_i}{|\mathcal{C}|}\right)}\end{aligned}$$

□

Lemma C.2.2 (A first deterministic inequality)**.** *Let t be a time step. We have*

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

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

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

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

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

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

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

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

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

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

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

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) to (C.10), we obtain the result of the statement. □

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

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

Lemma C.2.3 (Bounding the error restricted to the boundary)**.** *The total variation of $\boldsymbol{E}$ restricted to the boundary verifies*

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

Proof. The proof relies on a decomposition of the $\|\boldsymbol{E}\|_{\partial\mathcal{P}}$ term from Proposition C.1.2. We have

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

where $\overline{\boldsymbol{E}}_{\mathcal{P}}$ 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 $\boldsymbol{\epsilon}_{\mathcal{C}}$ the value of any row of $\overline{\boldsymbol{E}}_{\mathcal{P}}$ belonging to cluster $\mathcal{C}$, which is equal to the mean of errors $\boldsymbol{E}$ over that cluster. Also, we denote by $(\overline{\boldsymbol{E}}_{\mathcal{P}})_{\partial\mathcal{P}}$ the signal obtained from $\overline{\boldsymbol{E}}_{\mathcal{P}}$ 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:

$$\begin{aligned}
\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_{\partial\mathcal{P}} &= \left\|\boldsymbol{B}_{\partial\mathcal{P}}\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_{2,1} \\
&= \left\|\boldsymbol{B}_{\partial\mathcal{P}}(\overline{\boldsymbol{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\|_{2,1} \\
&\leq \|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\left\|(\overline{\boldsymbol{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\| \quad \text{(by Proposition C.1.1)} \\
&\leq \|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\left\|(\overline{\boldsymbol{E}}_{\mathcal{P}})_{\partial\mathcal{P}}\right\|_{F} \\
&= \|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}|\partial_v\mathcal{C}|\|\boldsymbol{\epsilon}_{\mathcal{C}}\|^2} \\
&= \|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}\frac{|\partial_v\mathcal{C}|}{|\mathcal{C}|}|\mathcal{C}|\|\boldsymbol{\epsilon}_{\mathcal{C}}\|^2} \\
&\leq \|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\sqrt{\sum_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|\|\boldsymbol{\epsilon}_{\mathcal{C}}\|^2} \\
&= \sqrt{2}w(\partial\mathcal{P})\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_{F} \qquad \text{(C.13)}
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\left\|\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}\boldsymbol{B}_{\partial\mathcal{P}^c}\boldsymbol{E}\right\|_{\partial\mathcal{P}} &= \left\|\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}\boldsymbol{B}_{\partial\mathcal{P}^c}\boldsymbol{E}\right\|_{2,1} \\
&\leq \left\|\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}\right\|_{\infty,1}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}\right\|_{F}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|(\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c})^{\top}\boldsymbol{B}^{\top}_{\partial\mathcal{P}}\right\|_{F}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \\
&\leq \left\|\boldsymbol{B}^{\top}_{\partial\mathcal{P}}\right\|_{2,1}\sqrt{\left\|\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}(\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c})^{\top}\right\|_{\infty,\infty}}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \quad \text{(by Proposition C.1.1)} \\
&\leq \frac{\left\|\boldsymbol{B}^{\top}_{\partial\mathcal{P}}\right\|_{1,1}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c}. \\
&= 2\frac{w(\partial\mathcal{P})}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c}. \qquad \text{(C.14)}
\end{aligned}$$

The result is obtained by combining Equations (C.12) to (C.14). □

Theorem C.2.4 (Theorem 2.1 of Hsu *et al.* (2012))**.** *At time step t, let $\boldsymbol{A} \in \mathbb{R}^{b\times t}$ where $b \in \mathbb{N}^*$, and let $\boldsymbol{v} \in \mathbb{R}^t$ be a random vector such that for some $\sigma \geq 0$, we have*

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

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

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

Lemma C.2.5 (Empirical process bound)**.** *Let $\boldsymbol{X}_m \in \mathbb{R}^{|\mathcal{T}_m| \times d}$ denotes the matrix of collected context vectors for task $m \in \mathcal{V}$, then, given collected context matrices $\{\boldsymbol{X}_m\}_{m\in\mathcal{V}}$, for any $\delta \in (0,1)$ we have with probability of at least $1-\delta$:*

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

where

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

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

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

where $\boldsymbol{X}_\mathcal{V} \coloneqq \operatorname{diag}(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{|\mathcal{V}|}) \in \mathbb{R}^{t\times d|\mathcal{V}|}$.

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

$$\begin{aligned}
\mathbb{E}\left[P(t)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\{u_t\eta_t\}P(t-1)|\mathcal{F}_{t-1}\right]\right] \quad \text{(by the law of total expectation)} \\
&= \mathbb{E}\left[P(t-1)\mathbb{E}\left[\exp(u_t\eta_t)|\mathcal{F}_{t-1}\right]\right] \quad \text{(because } \{\eta_s\}_{s=1}^{t-1} \text{ are } \mathcal{F}_{t-1} \text{ measurable.)} \\
&\leq \exp\left(\frac{1}{2}\sigma^2 u_t^2\right) \mathbb{E}\left[P(t-1)\right] \quad \text{(by the conditional sub-Gaussianity assumption)} \\
&\leq \prod_{s=1}^t \exp\left(\frac{1}{2}\sigma^2 u_s^2\right) \quad \text{(by induction)} \\
&= \exp\left(\frac{1}{2}\sigma^2 \|\boldsymbol{u}\|^2\right).
\end{aligned} \tag{C.17}$$

From Equations (C.16) and (C.17), we can apply Theorem C.2.4 to matrix $\boldsymbol{X}_\mathcal{V}$ and

random vector $\boldsymbol{\eta}$, which implies that with a probability at least $1-\delta$, we have

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

where we used the equalities $\|\boldsymbol{X}_{\mathcal{V}}\|_F = \sum_{m\in\mathcal{V}}\operatorname{Tr}(\boldsymbol{A}_m)$, $\|\boldsymbol{X}_{\mathcal{V}}\|^2 = \max_{m\in\mathcal{V}}\|\boldsymbol{A}_m\|$ and $\left\|\boldsymbol{X}_{\mathcal{V}}\boldsymbol{X}_{\mathcal{V}}^\top\right\|_F^2 = \left\|\boldsymbol{X}_{\mathcal{V}}^\top\boldsymbol{X}_{\mathcal{V}}\right\|_F^2 = \sum_{m\in\mathcal{V}}\|\boldsymbol{A}_m\|_F^2$. To arrive the the statement of the theorem, we use the fact that the context vectors have Euclidean norms of at most 1.

□

Proposition C.2.6 (Probabilistic inequality)**.** *With a probability at least* $1-\delta$*, we have*

$$\frac{1}{2t\alpha}\sum_{m\in\mathcal{V}}\|\boldsymbol{X}_m\boldsymbol{\epsilon}_m\|^2 + a_1(\mathcal{G},\boldsymbol{\Theta},\alpha_0)\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \leq a_2(\mathcal{G},\boldsymbol{\Theta},\alpha_0)\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_F + (1-\kappa)\|\boldsymbol{E}\|_{\partial\mathcal{P}}, \tag{C.18}$$

where $0 \leq \kappa < \frac{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}{2w(\partial\mathcal{P})}$.

Proof. The proof is a combination of the results of Lemmas C.2.2, C.2.3 and C.2.5. We have

$$\begin{aligned}
&\frac{1}{2t\alpha_\delta}\sum_{m\in\mathcal{V}}\|\boldsymbol{X}_m\boldsymbol{\epsilon}_m\|^2 + \|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \\
&\leq \frac{1}{t\alpha_\delta}\langle\boldsymbol{K},\boldsymbol{E}\rangle + \|\boldsymbol{E}\|_{\partial\mathcal{P}} \quad \text{(by Lemma C.2.2)} \\
&\leq \frac{1}{\alpha_0}\|\boldsymbol{E}\|_F + \kappa\|\boldsymbol{E}\|_{\partial\mathcal{P}} + (1-\kappa)\|\boldsymbol{E}\|_{\partial\mathcal{P}} \quad \text{(by Lemma C.2.5)} \\
&\leq \frac{\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_F}{\alpha_0} + \frac{\|\boldsymbol{E}\|_{\partial\mathcal{P}^c}}{\alpha_0\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}} + \kappa w(\partial\mathcal{P})\left(\sqrt{2}\max_{\mathcal{C}\in\mathcal{P}}\sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_F + 2\frac{\|\boldsymbol{E}\|_{\partial\mathcal{P}^c}}{\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}}\right) \\
&+ (1-\kappa)\|\boldsymbol{E}\|_{\partial\mathcal{P}},
\end{aligned}$$

where the last line is an application of Lemma C.2.3. Grouping the terms by the type of norm applied to $\boldsymbol{E}$ finishes the proof. □

Theorem C.2.7 (Oracle inequality, generalization of Theorem 5.5.8)**.** *Assume that the RE assumption holds for the empirical multi-task Gram matrix with constants* $\kappa \in \left[1, \frac{1}{2w(\partial\mathcal{P})}\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}\right)$*,* $\psi \in \left(0, \frac{1}{w(\partial\mathcal{P})}\min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})} - 2\right)$ *and* $\phi > 0$*. Suppose*

that $\max_{m\in\mathcal{V}}|\mathcal{T}_m(t)| \le bt$ *for some* $b>0$ *and* $\alpha_0 \ge \frac{1}{\psi w(\partial\mathcal{P})}$. *Then, with a probability at least* $1-\delta(t)$, *we have*

$$\left\|\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}(t)\right\|_F \le 2\frac{\sigma\alpha_0}{\phi^2\sqrt{t}} f(\mathcal{G},\boldsymbol{\Theta},\alpha_0)\sqrt{1+2b\sqrt{|\mathcal{V}|\log\frac{1}{\delta(t)}}+2b\log\frac{1}{\delta(t)}},$$

where

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

Proof. Using the previously established results, we obtain

$$\begin{aligned}
&\frac{1}{2t}\sum_{m\in\mathcal{V}}\|\boldsymbol{X}_m\boldsymbol{\epsilon}_m\|^2+\alpha\|\boldsymbol{E}\|_{\partial\mathcal{P}^c}\\
\le&\alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\boldsymbol{E}_{\mathcal{P}}\|_F+\alpha_\delta(1-\kappa)^+\|\boldsymbol{E}\|_{\partial\mathcal{P}} \quad \text{(by Proposition C.2.6)}\\
=&\alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\boldsymbol{E}_{\mathcal{P}}\|_F+\alpha_\delta(1-\kappa)^+\left\|\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{B}_{\partial\mathcal{P}}^\dagger\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{E}\right\|_{2,1} \quad \text{(by properties of the pseudo-inverse)}\\
\le&\alpha_\delta a_2(\boldsymbol{\Theta},\mathcal{G})\|\boldsymbol{E}_{\mathcal{P}}\|_F+\alpha_\delta\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\mathbb{1}_{\le 1}(\kappa)(1-\kappa)^+\left\|\boldsymbol{B}_{\partial\mathcal{P}}^\dagger\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{E}\right\| \quad \text{(by Proposition C.1.1)}\\
\le&\alpha_\delta(a_2(\boldsymbol{\Theta},\mathcal{G})+\mathbb{1}_{\le 1}(\kappa)\sqrt{2}w(\partial\mathcal{P}))\|\boldsymbol{E}\|_{\mathrm{RE}} \quad \text{(by definition of the } \|\|_{\mathrm{RE}} \text{ norm)}\\
\le&\alpha\frac{a_2(\boldsymbol{\Theta},\mathcal{G})+\mathbb{1}_{\le 1}(\kappa)\sqrt{2}w(\partial\mathcal{P})}{\phi\sqrt{t}}\sqrt{\sum_{m\in\mathcal{V}}\|\boldsymbol{\epsilon}_m\|^2_{\boldsymbol{A}_m}} \quad \text{(using the RE assumption)}\\
\le&\frac{\beta\alpha_\delta^2(a_2(\boldsymbol{\Theta},\mathcal{G})+\mathbb{1}_{\le 1}(\kappa)\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1})^2}{2\phi^2}+\frac{1}{2\beta t}\sum_{m\in\mathcal{V}}\|\boldsymbol{X}_m\boldsymbol{\epsilon}_m\|^2, \qquad \text{(C.19)}
\end{aligned}$$

where the last inequality holds for any $\beta>0$, and is a consequence of the property that $uv \le \dfrac{u^2+v^2}{2}$ for any $u,v\in\mathbb{R}$. In the second to last inequality we used the RE assumption, here it is important to mention that the assumption does not hold for any choice of α_0. In the definition of $\mathcal{S}$ i.e. the set matrices for which the RE condition holds, we have $\alpha_0 = \frac{1}{\psi w(\partial\mathcal{P})}$. We can also observe that this set is non increasing for the inclusion operator i.e. the RE condition would become weaker, for increasing α_0. Thus for any $\alpha_0 \ge \frac{1}{\psi w(\partial\mathcal{P})}$ the respective set of the RE assumption is contained in $\mathcal{S}$ and any matrix contained in the smaller set is automatically contained in $\mathcal{S}$, allowing us to use the RE condition in the proof due to our lower bound on $\alpha_0 \ge \frac{1}{\psi w(\partial\mathcal{P})}$.

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

$$\begin{aligned}\|\boldsymbol{Q}_{\partial\mathcal{P}^c}\boldsymbol{E}\|_F &= \left\|\boldsymbol{B}^{\dagger}_{\partial\mathcal{P}^c}\boldsymbol{B}_{\partial\mathcal{P}^c}\boldsymbol{E}\right\|_F \\ &\leq \sqrt{\left\|\boldsymbol{L}^{\dagger}_{\partial\mathcal{P}^c}\right\|_{\infty,\infty}}\|\boldsymbol{E}\|_{\partial\mathcal{P}^c} \\ &\leq \frac{2\alpha_\delta(a_2(\boldsymbol{\Theta},\mathcal{G},\alpha_0)+\mathbb{1}_{\leq 1}(\kappa)\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1})^2}{\phi^2 a_1(\boldsymbol{\Theta},\mathcal{G},\alpha_0)\min\limits_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}} \quad \text{(Equation (C.19) with } \beta=1\text{).}\end{aligned} \tag{C.20}$$

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

$$\begin{aligned}\left\|\overline{\boldsymbol{E}}_{\mathcal{P}}\right\|_F^2 &\leq \frac{1}{t\phi^2}\sum_{m\in\mathcal{V}}\|\boldsymbol{X}_m\boldsymbol{\epsilon}_m\|^2 \quad \text{(by RE assumption on empirical multi-task Gram matrix)} \\ &\leq \frac{4\alpha_\delta^2(a_2(\boldsymbol{\Theta},\mathcal{G},\alpha_0)+\mathbb{1}_{\leq 1}(\kappa)\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1})^2}{\phi^4} \quad \text{(by Equation (C.19) with } \beta=2\text{).}\end{aligned} \tag{C.21}$$

The result is then obtained by combining Equations (C.20) and (C.21) along with using the fact that $\boldsymbol{E}=\overline{\boldsymbol{E}}_{\mathcal{P}}+\boldsymbol{Q}_{\partial\mathcal{P}^c}\boldsymbol{E}$ and the expressions of $a_1(\boldsymbol{\Theta},\mathcal{G},\alpha_0)$ and $a_2(\boldsymbol{\Theta},\mathcal{G},\alpha_0)$, and bounding $\alpha_\delta(t)$ as follows:

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

□

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

From the adapted to the empirical multi-task Gram matrix

Lemma C.2.8 (Bounding a quadratic form using projections)**.** *Let* $\boldsymbol{M}_1, \cdots, \boldsymbol{M}_p \in \mathbb{R}^{d\times d}$ *be symmetric matrices, and let* $\boldsymbol{J} \coloneqq \frac{1}{p}\mathbf{1}\mathbf{1}^\top$*, and* $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{J}$*. Then, for any* $\boldsymbol{Z} \in \mathbb{R}^{p\times d}$ *with rows* $\{\boldsymbol{z}_i\}_{i=1}^p$*, we have:*

$$\left|\sum_{i=1}^p \boldsymbol{z}_i^\top \boldsymbol{M}_i \boldsymbol{z}_i\right| \leq \frac{1}{p}\left\|\sum_{i=1}^p \boldsymbol{M}_i\right\| \|\boldsymbol{Z}\|_{\boldsymbol{J}}^2 + 2\sqrt{\left\|\frac{1}{p}\sum_{i=1}^p \boldsymbol{M}_i^2\right\|}\|\boldsymbol{Z}\|_{\boldsymbol{Q}}\|\boldsymbol{Z}\|_{\boldsymbol{J}} + \max_{1\leq i\leq p}\|\boldsymbol{M}_i\|\|\boldsymbol{Z}\|_{\boldsymbol{Q}}^2$$

Proof. We have

$$\begin{aligned}\left|\sum_{i=1}^p \boldsymbol{z}_i^\top \boldsymbol{M}_i \boldsymbol{z}_i\right| &= \left|\sum_{i=1}^p \bar{\boldsymbol{z}}^\top \boldsymbol{M}_i \bar{\boldsymbol{z}} + 2\sum_{i=1}^p (\boldsymbol{z}_i - \bar{\boldsymbol{z}})^\top \boldsymbol{M}_i \bar{\boldsymbol{z}} + \sum_{i=1}^p (\boldsymbol{z}_i - \bar{\boldsymbol{z}})^\top \boldsymbol{M}_i (\boldsymbol{z}_i - \bar{\boldsymbol{z}})\right| \\ &\leq \left|\bar{\boldsymbol{z}}^\top \sum_{i=1}^p \boldsymbol{M}_i \bar{\boldsymbol{z}}\right| + 2\left|\sum_{i=1}^p \boldsymbol{e}_i^\top \boldsymbol{Q}\boldsymbol{Z}\boldsymbol{M}_i\bar{\boldsymbol{z}}\right| + \left|\sum_{i=1}^p \boldsymbol{e}_i^\top \boldsymbol{Q}\boldsymbol{Z}\boldsymbol{M}_i\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i\right|\end{aligned} \tag{C.22}$$

where we used the fact that $\boldsymbol{z}_i - \bar{\boldsymbol{z}} = \boldsymbol{Z}^\top\boldsymbol{e}_i - \boldsymbol{Z}^\top\boldsymbol{J}\boldsymbol{e}_i = \boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i$.

Let us now examine every term on the right-hand side of Equation (C.22). For the first term, we have

$$\left|\bar{\boldsymbol{z}}^\top \sum_{i=1}^p \boldsymbol{M}_i \bar{\boldsymbol{z}}\right| \leq \left\|\sum_{i=1}^p \boldsymbol{M}_i\right\| \|\bar{\boldsymbol{z}}\|^2 = \left\|\frac{1}{p}\sum_{i=1}^p \boldsymbol{M}_i\right\| \|\boldsymbol{Z}\|_{\boldsymbol{J}}^2. \tag{C.23}$$

For the second term, we have

$$\begin{aligned}
\left|\sum_{i=1}^{p} \boldsymbol{e}_i^\top \boldsymbol{Q}\boldsymbol{Z}\boldsymbol{M}_i\bar{\boldsymbol{z}}\right| &\leq \left\|\sum_{i=1}^{p} \boldsymbol{M}_i\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i\right\|\|\bar{\boldsymbol{z}}\| \\
&= \left\|\sum_{i=1}^{p} (\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i)\operatorname{vec}(\boldsymbol{Z}^\top\boldsymbol{Q})\right\|\|\bar{\boldsymbol{z}}\| \\
&\leq \left\|\sum_{i=1}^{p} (\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i)\right\|\left\|\operatorname{vec}(\boldsymbol{Z}^\top\boldsymbol{Q})\right\|\|\bar{\boldsymbol{z}}\| \\
&= \left\|\sum_{i=1}^{p} (\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i)\right\|\|\boldsymbol{Q}\boldsymbol{Z}\|_F\|\bar{\boldsymbol{z}}\| \\
&= \sqrt{\left\|(\sum_{i=1}^{p}(\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i))^\top \sum_{i=1}^{p}(\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i)\right\|}\|\boldsymbol{Q}\boldsymbol{Z}\|_F\|\bar{\boldsymbol{z}}\| \\
&= \sqrt{\left\|\sum_{i=1}^{p}\sum_{j=1}^{p}(\boldsymbol{e}_i^\top \otimes \boldsymbol{M}_i))(\boldsymbol{e}_j \otimes \boldsymbol{M}_j)\right\|}\|\boldsymbol{Q}\boldsymbol{Z}\|_F\|\bar{\boldsymbol{z}}\| \\
&= \sqrt{\left\|\sum_{i=1}^{p}\sum_{j=1}^{p}(\boldsymbol{e}_i^\top\boldsymbol{e}_j \otimes \boldsymbol{M}_i\boldsymbol{M}_j)\right\|}\|\boldsymbol{Q}\boldsymbol{Z}\|_F\|\bar{\boldsymbol{z}}\| \\
&= \sqrt{\left\|\sum_{i=1}^{p}\boldsymbol{M}_i^2\right\|}\|\boldsymbol{Q}\boldsymbol{Z}\|_F\|\bar{\boldsymbol{z}}\|. \qquad \text{(C.24)}
\end{aligned}$$

Finally, for the last term, we have

$$\begin{aligned}
\left|\sum_{i=1}^{p} \boldsymbol{e}_i^\top \boldsymbol{Q}\boldsymbol{Z}\boldsymbol{M}_i\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i\right| &\leq \sum_{i=1}^{p}\|\boldsymbol{M}_i\|\left\|\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i\right\|^2 \\
&\leq \max_{1\leq i\leq p}\|\boldsymbol{M}_i\|\sum_{i=1}^{p}\left\|\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{e}_i\right\|^2 \\
&= \max_{1\leq i\leq p}\|\boldsymbol{M}_i\|\|\boldsymbol{Q}\boldsymbol{Z}\|_F^2. \qquad \text{(C.25)}
\end{aligned}$$

Combining Equations (C.23) to (C.25) yields the result. □

We also define an operator norm that is induced by the $\|\|_{\mathrm{RE}}$ introduced in Definition C.2.1.

Definition C.2.9 ((RE,$\mathcal{S}$)-induced operator norm)**.** *Let $\{\boldsymbol{M}_m\}_{m\in\mathcal{V}} \subseteq \mathbb{R}^{d\times d}$ be symmetric matrices associated to the graph nodes $\mathcal{V}$, and let $\boldsymbol{M}_{\mathcal{V}} \coloneqq \operatorname{diag}\left(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_{|\mathcal{V}|}\right) \in$*

$\mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$. *For any cluster* $\mathcal{C} \in \mathcal{P}$, *let the cluster mean and mean of squares associated to those matrices be given by*

$$\overline{\boldsymbol{M}}_{\mathcal{C}} \coloneqq \frac{1}{|\mathcal{C}|}\sum_{m\in\mathcal{C}} \boldsymbol{M}_m, \qquad \overline{\boldsymbol{M}^2}_{\mathcal{C}} \coloneqq \frac{1}{|\mathcal{C}|}\sum_{m\in\mathcal{C}} \boldsymbol{M}_m^2.$$

The RE-induced operator norm of $\boldsymbol{M}_{\mathcal{V}}$ *is defined as*

$$\|\boldsymbol{M}\|_{\mathrm{RE},\mathcal{S}} \coloneqq \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}}_{\mathcal{C}}\right\| \vee \sqrt{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} \vee \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{m\in\mathcal{V}} \|\boldsymbol{M}_m\|. \tag{C.26}$$

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. For the sake of readability we remove the arguments of the constants: $a_1 = a_1\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right)$, $a_2 = a_2\left(\mathcal{G}, \boldsymbol{\Theta}, \frac{1}{\psi w(\partial\mathcal{P})}\right)$,

Proposition C.2.10 (Restricted spectral norm)**.** *Let* $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}|\times d}$ *verifying*

$$a_1\|\boldsymbol{Z}\|_{\partial\mathcal{P}^c} \leq a_2\left\|\overline{\boldsymbol{Z}}_{\mathcal{P}}\right\|_F + (1-\kappa)^+\|\boldsymbol{Z}\|_{\partial\mathcal{P}}$$

Let $\{\boldsymbol{M}_m\}_{m\in\mathcal{V}} \subseteq \mathbb{R}^{d\times d}$ *be symmetric matrices associated to the graph nodes* $\mathcal{V}$, *and let* $\boldsymbol{M}_{\mathcal{V}} \coloneqq \mathrm{diag}(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_{|\mathcal{V}|}) \in \mathbb{R}^{d|\mathcal{V}|\times d|\mathcal{V}|}$. *Then we have:*

$$\left|\sum_{m\in\mathcal{V}} \boldsymbol{z}_m^\top \boldsymbol{M}_m \boldsymbol{z}_m\right| \leq \|\boldsymbol{M}\|_{\mathrm{RE},\mathcal{S}}^2 \left(1 + \frac{a_2 + (1-\kappa)^+\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}}{a_1}\right)^2 \|\boldsymbol{Z}\|_{\mathrm{RE}}^2. \tag{C.27}$$

Proof. For any cluster $\mathcal{C}$, we denote by $\boldsymbol{B}_{\mathcal{C}}$ the incidence matrix obtained by setting the rows of $\boldsymbol{B}$ outside the edges linking nodes in $\mathcal{C}$ to null vectors. The latter's null space 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 $\boldsymbol{Q}_{\mathcal{C}} \coloneqq \boldsymbol{B}_{\mathcal{C}}^\dagger \boldsymbol{B}_{\mathcal{C}}$.

On the one hand, for any signal $\boldsymbol{Z} \in \mathbb{R}^{|\mathcal{V}| \times d}$ we have

$$\begin{aligned}\|\boldsymbol{Z}\|_{\partial\mathcal{P}^c} &= \sum_{\mathcal{C}\in\mathcal{P}} \|\boldsymbol{B}_{\mathcal{C}}\boldsymbol{Z}\|_{2,1} \\ &\geq \sum_{\mathcal{C}\in\mathcal{P}} \frac{\left\|\boldsymbol{B}_{\mathcal{C}}^{\dagger}\boldsymbol{B}_{\mathcal{C}}\boldsymbol{Z}\right\|_F}{\sqrt{\left\|\boldsymbol{L}_{\mathcal{C}}^{\dagger}\right\|_{\infty,\infty}}} \\ &\geq \min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} \sum_{\mathcal{C}\in\mathcal{P}} \|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}}\end{aligned}$$

Hence, by the proposition's assumptions, $\boldsymbol{Z}$ verifies

$$\begin{aligned}\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} a_1 \sum_{\mathcal{C}\in\mathcal{P}} \|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}} &\leq (a_2\left\|\overline{\boldsymbol{Z}}_{\mathcal{P}}\right\|_F + (1-\kappa)\|\boldsymbol{Z}\|_{\partial\mathcal{P}}) \\ &\leq a_2\left\|\overline{\boldsymbol{Z}}_{\mathcal{P}}\right\|_F + (1-\kappa)^+\|\boldsymbol{B}_{\partial\mathcal{P}}\|_{2,1}\left\|\boldsymbol{B}_{\partial\mathcal{P}}^{\dagger}\boldsymbol{B}_{\partial\mathcal{P}}\boldsymbol{Z}\right\| \\ &\leq (a_2 + (1-\kappa)^+\|\boldsymbol{B}\|_{2,1})\|\boldsymbol{Z}\|_{\mathrm{RE}}\end{aligned}$$

From Lemma C.2.8, we have

$$\begin{aligned}&\left|\sum_{m\in\mathcal{V}} \boldsymbol{z}_m^{\top}\boldsymbol{M}_m\boldsymbol{z}_m\right| \\ &\leq \sum_{\mathcal{C}\in\mathcal{P}} \left|\sum_{m\in\mathcal{C}} \boldsymbol{z}_m^{\top}\boldsymbol{M}_m\boldsymbol{z}_m\right| \\ &\leq \sum_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}}_{\mathcal{C}}\right\|\|\boldsymbol{Z}\|_{\boldsymbol{J}_{\mathcal{C}}}^2 + 2\sum_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|}\|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}}\|\boldsymbol{Z}\|_{\boldsymbol{J}_{\mathcal{C}}} + \sum_{\mathcal{C}\in\mathcal{P}} \max_{m\in\mathcal{C}} \|\boldsymbol{M}_m\|\|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}}^2,\end{aligned} \tag{C.28}$$

where we used Equation (C.1).

This allows us to bound every term in Equation (C.28). For the second term on

the right-hand side, we have

$$\begin{aligned}
&\sum_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} \|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}} \|\boldsymbol{Z}\|_{\boldsymbol{J}_{\mathcal{C}}} \\
\leq &\max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} \left\|\overline{\boldsymbol{Z}}_{\mathcal{P}}\right\|_F \sqrt{\sum_{\mathcal{C}\in\mathcal{P}} \|\boldsymbol{Z}\|^2_{\boldsymbol{Q}_{\mathcal{C}}}} \\
\leq &\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-\frac{1}{2}}}{a_1} \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} (a_2 + (1-\kappa)^+ \|\boldsymbol{B}\|_{2,1}) \|\boldsymbol{Z}\|^2_{\mathrm{RE}}
\end{aligned} \tag{C.29}$$

As for the third term, we have

$$\begin{aligned}
\sum_{\mathcal{C}\in\mathcal{P}} \max_{m\in\mathcal{C}} \|\boldsymbol{M}_m\| \|\boldsymbol{Z}\|^2_{\boldsymbol{Q}_{\mathcal{C}}} &\leq \max_{m\in\mathcal{V}} \|\boldsymbol{M}_m\| \left(\sum_{\mathcal{C}\in\mathcal{P}} \|\boldsymbol{Z}\|_{\boldsymbol{Q}_{\mathcal{C}}}\right)^2 \\
&\leq \max_{m\in\mathcal{V}} \|\boldsymbol{M}_m\| \frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1}}{a_1^2} (a_2 + (1-\kappa)^+ \|\boldsymbol{B}\|_{2,1})^2 \|\boldsymbol{Z}\|^2_{\mathrm{RE}}
\end{aligned} \tag{C.30}$$

Consequently, denoting $v = \dfrac{a_2 + (1-\kappa)^+ \|\boldsymbol{B}\|_{2,1}}{a_1}$, and combining Equations (C.28) to (C.30), we obtain

$$\begin{aligned}
&\left|\sum_{m\in\mathcal{V}} \boldsymbol{z}_m^\top \boldsymbol{M}_m \boldsymbol{z}_m\right| \\
&\left(\max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}}_{\mathcal{C}}\right\| + 2v \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} + v^2 \max_{i\in\mathcal{V}} \|\boldsymbol{M}_i\|\right) \|\boldsymbol{Z}\|^2_{\mathrm{RE}} \\
\leq &\left(\max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}}_{\mathcal{C}}\right\|) \vee \sqrt{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{M}^2}_{\mathcal{C}}\right\|} \vee \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{i\in\mathcal{V}} \|\boldsymbol{M}_i\|\right) (1+v)^2 \|\boldsymbol{Z}\|^2_{\mathrm{RE}},
\end{aligned}$$

which finishes the proof. □

Proposition C.2.11 (Inheritance of a RE condition from a close matrix)**.** *Assume that the matrix $\boldsymbol{V}_{\mathcal{V}}$ verifies the RE condition with constant $\phi > 0$, and that $\left\|\dfrac{\boldsymbol{A}_{\mathcal{V}}}{t} - \boldsymbol{V}_{\mathcal{V}}\right\|_{\mathrm{op,RE}} \leq \gamma\phi^2$ for some $\gamma \in \left(0, \left(1 + \frac{a_2 + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1}\right)^{-2}\right)$. Then $\dfrac{\boldsymbol{A}_{\mathcal{V}}}{t}$ verifies the RE condition with constant*

$$\hat{\phi} = \phi \sqrt{1 - \gamma \left(1 + \frac{a_2 + (1-\kappa)^+ \sqrt{2} w(\partial \mathcal{P})}{a_1}\right)^2} \tag{C.31}$$

Proof. From Proposition C.2.6, we know that

$$\begin{aligned}
\frac{1}{t}\epsilon_{\mathcal{V}}^\top \boldsymbol{A}_{\mathcal{V}} \epsilon_{\mathcal{V}} &= \frac{1}{|\mathcal{V}|}\epsilon_{\mathcal{V}}^\top \boldsymbol{V}_{\mathcal{V}} \epsilon_{\mathcal{V}} + \epsilon_{\mathcal{V}}^\top \boldsymbol{\Delta}_{\mathcal{V}} \epsilon_{\mathcal{V}} \\
&\geq \frac{1}{|\mathcal{V}|}\epsilon_{\mathcal{V}}^\top \boldsymbol{V}_{\mathcal{V}} \epsilon_{\mathcal{V}} - \left|\epsilon_{\mathcal{V}}^\top \boldsymbol{\Delta}_{\mathcal{V}} \epsilon_{\mathcal{V}}\right| \\
&\geq \left(\phi^2 - \max_{m \in \mathcal{V}} \|\boldsymbol{\Delta}_{\mathcal{V}}\|_{\text{op,RE}} \left(1 + \frac{a_2 + (1-\kappa)^+ \|\boldsymbol{B}_{\partial \mathcal{P}}\|_{2,1}}{a_1}\right)^2\right) \|\boldsymbol{E}\|_{\text{RE}}^2 \\
&\geq \left(\phi^2 - \gamma\phi^2 \left(1 + \frac{a_2 + (1-\kappa)^+ \|\boldsymbol{B}_{\partial \mathcal{P}}\|_{2,1}}{a_1}\right)^2\right) \|\boldsymbol{E}\|_{\text{RE}}^2
\end{aligned}$$

where the third inequality is an applicaiton of Proposition C.2.10. □

Theorem C.2.12 (Matrix Freedman Inequality, Tropp (2011))**.** *Consider a matrix martingale $\{\boldsymbol{M}(t)\}_{t\geq 1}$ with dimension $d_1 \times d_2$. Let $\{\boldsymbol{N}(t)\}_{t\geq 1}$ be the associated difference sequence. Assume that for some $A > 0$, we have $\|\boldsymbol{N}(t)\| \leq A \quad \forall t \geq 1$ almost surely. Define for any $t \geq 1$:*

$$\begin{aligned}
\boldsymbol{W}_{col}(t) &\coloneqq \sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{N}(\tau)\boldsymbol{N}(\tau)^\top | \mathcal{F}_{\tau-1}\right] \\
\boldsymbol{W}_{row}(t) &\coloneqq \sum_{\tau=1}^{t} \mathbb{E}\left[\boldsymbol{N}(\tau)^\top \boldsymbol{N}(\tau) | \mathcal{F}_{\tau-1}\right].
\end{aligned}$$

Then, for any $u, v > 0$,

$$\mathbb{P}\left[\exists t \geq 1; \|\boldsymbol{M}(t)\| \geq u \text{ and } \|\boldsymbol{W}_{col}\|(t) \vee \|\boldsymbol{W}_{row}(t)\| \leq v\right] \leq (d_1 + d_2) \exp\left(-\frac{3u^2}{6v + 2Au}\right)$$

Corollary C.2.13. *Let $\{\boldsymbol{N}(\tau)\}_{\tau=1}^t$ 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 $\{\boldsymbol{M}(n)\}_{\tau=1}^N$ of random matrices defined by*

$$\boldsymbol{M}(n) = \sum_{i=1}^{n} \boldsymbol{N}(t_i) - \mathbb{E}\left[\boldsymbol{N}(t_i) | \mathcal{F}_{t_i - 1}\right] \tag{C.32}$$

Then $\{\boldsymbol{M}(n)\}_{n=1}^N$ is a martingale adapted to the filtration $\{\mathcal{F}_{t_n}\}_{n=1}^N$.

Moreover,if $\|\boldsymbol{N}(\tau)\| \leq b \quad \forall \tau \in [t]$ *for some* $b > 0$*, then we have*

$$\mathbb{P}\left[\|\boldsymbol{M}(N)\| \geq u\right] \leq (d_1 + d_2) \exp\left(-\frac{3u^2}{6Nb^2 + 2\sqrt{2}bu}\right). \tag{C.33}$$

Proof. We denote $\mathbb{E}\left[\cdot|\mathcal{F}_s\right]$ as $\mathbb{E}_s\left[\cdot\right]$ for any $s \in \mathbb{N}$. Also, let $\boldsymbol{C}(s) \coloneqq \mathbb{E}_{s-1}\left[\boldsymbol{N}(s)\right]$, which is $\mathcal{F}_{s-1}$-measurable by construction. We have for any $n \in [N]$,

$$\mathbb{E}_{t_{n-1}}\left[\boldsymbol{C}(t_n)\right] = \mathbb{E}_{t_{n-1}}\left[\mathbb{E}_{t_n-1}\left[\boldsymbol{N}(t_n)\right]\right] = \mathbb{E}_{t_{n-1}}\left[\boldsymbol{N}(t_n)\right] \tag{C.34}$$

$$\implies \mathbb{E}_{t_{n-1}}\left[\boldsymbol{N}(t_n) - \boldsymbol{C}(t_n)\right] = 0 \tag{C.35}$$

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 $\tau \geq 1$

$$\|\boldsymbol{N}(\tau) - \boldsymbol{C}(\tau)\|^2 = \left\|(\boldsymbol{N}(\tau) - \boldsymbol{C}(\tau))^2\right\| \tag{C.36}$$

$$\leq \operatorname{Tr}\big((\boldsymbol{N}(\tau) - \boldsymbol{C}(\tau))^2\big) \tag{C.37}$$

$$= \operatorname{Tr}\big((\boldsymbol{N}(\tau) - \boldsymbol{C}(\tau))^2\big) \tag{C.38}$$

$$= \|\boldsymbol{N}(\tau)\|_F^2 - 2\operatorname{Tr}(\boldsymbol{C}(\tau)\boldsymbol{N}(\tau)) + \operatorname{Tr}\big(\boldsymbol{C}(\tau)^2\big) \tag{C.39}$$

$$\leq \|\boldsymbol{N}(\tau)\|_F^2 + \operatorname{Tr}\big(\boldsymbol{C}(\tau)^2\big) \leq 2b^2 \tag{C.40}$$

Hence $\boldsymbol{N}(\tau) - \boldsymbol{C}(\tau)$ is integrable for any $\tau \geq 1$. This shows that $\boldsymbol{M}(n)$ 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 C.2.12. 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 $\boldsymbol{W}_{\text{col}}$ and $\boldsymbol{W}_{\text{row}}$ from Theorem C.2.12. Notice that the two matrices are equal since the difference sequence matrices $\boldsymbol{N}(t_s)$ are symmetric.

Hence, for any $n \in [N]$, we have

$$\|\boldsymbol{W}_{\mathrm{col}}(N)\| \vee \|\boldsymbol{W}_{\mathrm{row}}(N)\| \leq \mathrm{Tr}(\boldsymbol{W}_{\mathrm{col}}(N)) \vee \mathrm{Tr}(\boldsymbol{W}_{\mathrm{row}}(N)) \tag{C.41}$$

$$= \mathrm{Tr}\left(\sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[(\boldsymbol{N}(t_n) - \boldsymbol{C}(t_n))^2\right]\right) \tag{C.42}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\|\boldsymbol{N}(t_n)\|_F^2\right] - \mathbb{E}_{t_n-1}\left[2\,\mathrm{Tr}(\boldsymbol{C}(t_n)\boldsymbol{N}(t_n))\right] \tag{C.43}$$

$$+ \mathrm{Tr}\big(\boldsymbol{C}(t_n)^2\big) \tag{C.44}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\|\boldsymbol{N}(t_n)\|_F^2\right] - \mathrm{Tr}\big(\boldsymbol{C}(t_n)^2\big) \tag{C.45}$$

$$\leq \sum_{n=1}^{N} \mathbb{E}_{t_n-1}\left[\|\boldsymbol{N}(t_n)\|_F^2\right] \leq Nb^2. \tag{C.46}$$

By Theorem C.2.12, we have for any $u > 0$

$$2d\exp\left(-\frac{3u^2}{6Nb^2 + 2\sqrt{2}bu}\right) \geq \mathbb{P}\left[\exists n \geq 1; \|\boldsymbol{M}(n)\| \geq u \text{ and } \|\boldsymbol{W}_{\mathrm{col}}(n)\| \leq Nb^2\right] \tag{C.47}$$

$$\geq \mathbb{P}\left[\|\boldsymbol{M}(N)\| \geq u \text{ and } \|\boldsymbol{W}_{\mathrm{col}}(N)\| \leq Nb^2\right] \tag{C.48}$$

$$= \mathbb{P}\left[\|\boldsymbol{M}(N)\| \geq u\right] \tag{C.49}$$

where the last line holds because we showed that the inequality $\|\boldsymbol{W}_{\mathrm{col}}(N)\| \leq Nb^2$ holds almost surely. □

Proposition C.2.14 (Concentration of the empirical multi-task Gram matrix around the adapted one)**.** *Let $t \geq 1$, $b > 0$. Then we have:*

$$\mathbb{P}\left[\left\|\frac{\boldsymbol{A}_{\mathcal{V}}(t)}{t} - \boldsymbol{V}_{\mathcal{V}}\right\|_{\mathrm{op,RE}} > \gamma \middle| \max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \leq bt\right] \leq d(2|\mathcal{P}|e^{-A_1 t} + (|\mathcal{V}|+|\mathcal{P}|)e^{-A_2 t} + 2|\mathcal{V}|e^{-A_3 t}),$$

where

$$A_1 \coloneqq \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} |\mathcal{C}| t}{6b + 2\sqrt{2}\gamma}$$
$$A_2 \coloneqq \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) t}{6b + 2\sqrt{2}\gamma \sqrt{\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}} |\mathcal{C}|}}}$$
$$A_3 \coloneqq \frac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2 t}{6b + 2\sqrt{2}\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}$$

Proof. For $\gamma > 0$, let us define

$$\boldsymbol{\Delta}_m \coloneqq \frac{\boldsymbol{A}_{\mathcal{V}}}{t} - \boldsymbol{V}_{\mathcal{V}} \quad \text{and } G_{\text{Gram},\gamma} \coloneqq \left\{ \frac{1}{t} \|\boldsymbol{\Delta}_{\mathcal{V}}\|_{\text{RE},\mathcal{S}} \leq \gamma \right\},$$

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 C.2.9. We can express the complementary of this event as the disjunction of a finite number of events as follows:

$$G^c_{\text{Gram},\gamma} \tag{C.50}$$
$$= \left\{ \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{\Delta}}_{\mathcal{C}}\right\| \vee \sqrt{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{\mathcal{C}\in\mathcal{P}} \left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}\right\|} \vee \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^{-1} \max_{m\in\mathcal{V}} \|\boldsymbol{\Delta}_m\| > t\gamma \right\} \tag{C.51}$$
$$= \bigcup_{\mathcal{C}\in\mathcal{P}} \left\{ \left\|\overline{\boldsymbol{\Delta}}_{\mathcal{C}}\right\| > t\gamma \right\} \cup \bigcup_{\mathcal{C}\in\mathcal{P}} \left\{ \left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}\right\| > t^2\gamma^2 \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) \right\} \cup \bigcup_{m\in\mathcal{V}} \left\{ \|\boldsymbol{\Delta}_m\| > t\gamma \min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C}) \right\} \tag{C.52}$$

The first and third event can be bounded by considering the sequence $\boldsymbol{x}\boldsymbol{x}^\top(\tau)$ adapted to the filtration $\{\mathcal{F}_\tau\}$, verifying $\left\|\boldsymbol{x}\boldsymbol{x}^\top(\tau)\right\| \leq$.

Bounding the probability of the first event Let $\mathcal{C} \in \mathcal{P}$ be a cluster. By definition, we have

$$|\mathcal{C}|\overline{\boldsymbol{\Delta}}_{\mathcal{C}}(t) = \sum_{m\in\mathcal{C}} \sum_{\tau\in\mathcal{T}_m(t)} \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]$$
$$= \sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]$$

We will apply Corollary C.2.13 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

$$\begin{aligned}
\mathbb{P}\left[\left\|\overline{\boldsymbol{\Delta}}_{\mathcal{C}}(t)\right\| > \gamma t \middle| \max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \leq bt\right] &\leq 2d\exp\left(\frac{-3\gamma^2|\mathcal{C}|^2t^2}{6\sum_{m\in\mathcal{C}}|\mathcal{T}_m(t)| + 2\sqrt{2}\gamma|\mathcal{C}|t}\right) \\
&\leq 2d\exp\left(\frac{-3\gamma^2|\mathcal{C}|^2t^2}{6|\mathcal{C}|bt + 2\sqrt{2}\gamma|\mathcal{C}|t}\right) \\
&= 2d\exp\left(\frac{-3\gamma^2|\mathcal{C}|t}{6b + 2\sqrt{2}\gamma}\right) \\
&\leq 2d\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|t}{6b + 2\sqrt{2}\gamma}\right) \qquad \text{(C.53)}
\end{aligned}$$

Bounding the probability of the third event Let $m \in \mathcal{V}$ be a task index. We apply Corollary C.2.13 for the sequence of time steps in $\mathcal{T}_m(t)$. We have

$$\boldsymbol{\Delta}_m(t) = \sum_{\tau\in\mathcal{T}_m(t)} \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]$$

is a martingale sequence, hence

$$\begin{aligned}
\mathbb{P}\left[\|\boldsymbol{\Delta}_m(t)\| > \gamma\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t \middle| \max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \leq bt\right] &\leq 2d\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2t^2}{6|\mathcal{T}_m(t)| + 2\sqrt{2}\gamma\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}\right) \\
&\leq 2d\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2t^2}{6bt + 2\sqrt{2}\gamma\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}\right) \\
&= 2d\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})^2t}{6b + 2\sqrt{2}\gamma\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}\right). \qquad \text{(C.54)}
\end{aligned}$$

Bounding the probability of the second event Let $\mathcal{C} \in \mathcal{P}$ be a cluster, and let us denote $\boldsymbol{e}_m$ the m^{th} canonical vector of $\mathbb{R}^{|\mathcal{C}|}$. We have

$$\begin{aligned}
\left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}(t)\right\| &= \frac{1}{|\mathcal{C}|}\left\|\sum_{m\in\mathcal{C}}\left(\sum_{\tau\in\mathcal{T}_m(t)} \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)^2\right\| \\
&= \frac{1}{|\mathcal{C}|}\left\|\sum_{m\in\mathcal{C}} \boldsymbol{e}_m^\top \otimes \left(\sum_{\tau\in\mathcal{T}_m(t)} \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)\right\|^2 \\
&= \frac{1}{|\mathcal{C}|}\left\|\sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \boldsymbol{e}_{m(\tau)}^\top \otimes \left(\boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right)\right\|^2 \\
&= \frac{1}{|\mathcal{C}|}\left\|\sum_{\tau\in\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)} \boldsymbol{e}_{m(\tau)}^\top \otimes \boldsymbol{x}\boldsymbol{x}(\tau) - \mathbb{E}\left[\boldsymbol{e}_{m(\tau)} \otimes \boldsymbol{x}\boldsymbol{x}(\tau)|\mathcal{F}_{\tau-1}\right]\right\|^2,
\end{aligned}$$

where the last equality holds since $m(\tau)$ is measurable w.r.t. $\mathcal{F}_{\tau-1}$. We will apply the Corollary C.2.13 to the set of time steps $\bigcup_{m\in\mathcal{C}}\mathcal{T}_m(t)$ and the adapted sequence $\boldsymbol{e}_{m(\tau)}^\top \otimes \boldsymbol{x}\boldsymbol{x}(\tau)$ of matrices in $\mathbb{R}^{d\times d|\mathcal{C}|}$. Hence we have

$$\begin{aligned}
&\mathbb{P}\left[\sqrt{\left\|\overline{\boldsymbol{\Delta}^2}_{\mathcal{C}}(t)\right\|} > \gamma t \min_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}\Big|\max_{m\in\mathcal{V}}|\mathcal{T}_m(t)| \le bt\right] \\
&\le d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t^2}{6\sum_{m\in\mathcal{C}}|\mathcal{T}_m(t)| + 2\sqrt{2}\gamma\sqrt{|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}}\right) \\
&\le d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6|\mathcal{C}|b + 2\sqrt{2}\gamma\sqrt{|\mathcal{C}|\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}}\right) \\
&= d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6b + 2\sqrt{2}\gamma\sqrt{\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{|\mathcal{C}|}}}\right) \\
&\le d(1+|\mathcal{C}|)\exp\left(\frac{-3\gamma^2\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})t}{6b + 2\sqrt{2}\gamma\sqrt{\frac{\min_{\mathcal{C}\in\mathcal{P}} c_{\mathcal{G}}(\mathcal{C})}{\min_{\mathcal{C}\in\mathcal{P}}|\mathcal{C}|}}}\right) \qquad \text{(C.55)}
\end{aligned}$$

Union bound We conclude the result of the statement via a union bound using Equation (C.52). □

Proposition C.2.15 (Concentration of the empirical multi-task Gram matrix around the adapted one, simplified)**.** *Let $t \geq 1$, $b > 0$. Assume that $\max_{m \in \mathcal{V}} |\mathcal{T}_m(t)| \leq bt$. Then we have:*

$$\mathbb{P}\left[\left\|\frac{\boldsymbol{A}_{\mathcal{V}}}{t} - \boldsymbol{V}_{\mathcal{V}}\right\|_{\mathrm{op,RE}} > \gamma\right] \leq 6d|\mathcal{V}| \exp\left(\frac{-3\gamma^2(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}_{\mathcal{G}}(\mathcal{C})^2)t}{6b + 2\sqrt{2}\gamma}\right),$$

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 C.2.14, we deduce the result. □

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.* (2021). In particular, we use their Lemma 10.

Theorem C.2.16 (Lemma 10 of Oh *et al.* (2021))**.** *Under Assumption 5.3.2 on the context generating distribution, let $t \geq 1$. We have for any $\boldsymbol{\theta} \in \mathbb{R}^d$:*

$$\sum_{\boldsymbol{x}\in\mathcal{A}(t)} \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top \mathbb{1}\left\{\boldsymbol{x} \in \underset{\tilde{\boldsymbol{x}}\in\mathcal{A}(t)}{\arg\max} \langle \boldsymbol{\theta}, \tilde{\boldsymbol{x}}\rangle\right\}\right] \succcurlyeq \frac{1}{2\nu\omega}\overline{\boldsymbol{\Sigma}} \tag{C.56}$$

Proposition C.2.17 (RE condition from the true to the adapted Gram matrix)**.** *Under Assumption 5.3.2, for any $t \geq 1$, the adapted Gram matrix $\boldsymbol{V}_{\mathcal{V}}(t)$ verifies*

the compatibility condition with constants κ *and* $\frac{\phi}{\sqrt{2\nu\omega}}$.

Proof. For $t \geq 1$, we have

$$\mathbb{E}\left[\boldsymbol{x}(t)\boldsymbol{x}(t)^\top | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{\boldsymbol{x}\in\mathcal{A}(t)} \boldsymbol{x}(t)\boldsymbol{x}(t)^\top | \mathcal{F}_{t-1}\right] \tag{C.57}$$

Let $m \in \mathcal{V}$. We have

$$\begin{aligned}
\boldsymbol{V}_m(t) &= \frac{1}{t}\sum_{\tau\in\mathcal{T}_m(t)} \mathbb{E}\left[\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^\top | \mathcal{F}_{\tau-1}\right] \\
&= \frac{1}{t}\sum_{\tau\in\mathcal{T}_m(t)} \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^\top | \boldsymbol{\theta}_m(\tau-1), \mathcal{F}_{\tau-1}\right] | \mathcal{F}_{\tau-1}\right] \quad \text{(law of total expectation)} \\
&= \frac{1}{t}\sum_{\tau\in\mathcal{T}_m(t)} \mathbb{E}\left[\boldsymbol{x}(\tau)\boldsymbol{x}(\tau)^\top | \boldsymbol{\theta}_m(\tau-1)\right] \quad (\boldsymbol{x}(\tau) \text{ is fully determined by } \boldsymbol{\theta}_m(\tau-1)) \\
&= \frac{1}{t}\sum_{\tau\in\mathcal{T}_m(t)} \mathbb{E}\left[\sum_{\boldsymbol{x}\in\mathcal{A}(\tau)} \boldsymbol{x}\boldsymbol{x}^\top \mathbb{1}\left\{\boldsymbol{x} \in \underset{\tilde{\boldsymbol{x}}\in\mathcal{A}(t)}{\arg\max} \langle \boldsymbol{\theta}, \tilde{\boldsymbol{x}}\rangle\right\} | \boldsymbol{\theta}_m(\tau-1)\right] \\
&\succcurlyeq \frac{1}{2\nu\omega}\overline{\boldsymbol{\Sigma}} \quad \text{(by Theorem C.2.16).}
\end{aligned} \tag{C.58}$$

Now, let $\boldsymbol{Z} \in \mathcal{S}$, where $\mathcal{S}$ is defined with constant κ of Assumption 5.5.4. Then

$$\begin{aligned}
\sum_{m\in\mathcal{V}} \|\boldsymbol{z}\|_{\boldsymbol{V}_m(t)} &\geq \frac{1}{2\nu\omega}\sum_{m\in\mathcal{V}} \|\boldsymbol{z}_m\|_{\overline{\boldsymbol{\Sigma}}} \quad \text{by Equation (C.58)} \\
&\geq \frac{\phi^2}{2\nu\omega}\|\boldsymbol{Z}\|_{\mathrm{RE}}^2 \quad \text{(by Assumption 5.5.4),}
\end{aligned}$$

which finishes the proof. □

Theorem C.2.18 (RE condition holding for the empirical multi-task Gram matrix, generalization of Theorem 5.5.9)**.** *Under assumptions 5.3.2 and 5.5.4, let* $t \geq 1$*, and let* κ, ϕ *be the constants from Assumption 5.5.4. Assume that* $\max_{m\in\mathcal{V}} |\mathcal{T}_m(t)| \leq bt$*. Then, for any* $\gamma \in \left(0, \left(1 + \frac{a_2 + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1}\right)^{-2}\right)$*, the empirical multi-task Gram matrix verifies the RE condition with constants* κ *and* $\hat{\phi}$*, with*

$$\hat{\phi} = \tilde{\phi}\sqrt{1 - \gamma\left(1 + \frac{a_2 + (1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1}\right)^2}, \tag{C.59}$$

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

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 C.2.17. Then:

$$1-6d|\mathcal{V}|\exp\left(\frac{-3\gamma^2\tilde{\phi}^4(\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C})\wedge\tilde{c}_{\mathcal{G}}(\mathcal{C})^2)t}{6b+2\sqrt{2}\gamma\tilde{\phi}^2}\right) \tag{C.60}$$

$$\leq\mathbb{P}\left[\left\|\frac{\boldsymbol{A}_{\mathcal{V}}}{t}-\boldsymbol{V}_{\mathcal{V}}\right\|_{\text{op,RE}}\leq\gamma\tilde{\phi}^2\right]\quad\text{(by Proposition C.2.15)} \tag{C.61}$$

$$\leq\mathbb{P}\left[\frac{\boldsymbol{A}_{\mathcal{V}}}{t}\text{ satisfies the RE condition with constant }\kappa\text{ and }\hat{\phi}\right]\quad\text{(by Proposition C.2.11),} \tag{C.62}$$

where $\hat{\phi}=\tilde{\phi}\sqrt{1-\gamma\left(1+\frac{a_2+(1-\kappa)^+\sqrt{2}w(\partial\mathcal{P})}{a_1}\right)^2}$. □

C.2.4 Regret bound

Lemma C.2.19 (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* $1-\delta$*:*

$$\max_{m\in\mathcal{V}}\frac{|\mathcal{T}_m(t)|}{t}\leq\frac{1}{|\mathcal{V}|}+2\sqrt{\frac{1}{t|\mathcal{V}|}\log\frac{|\mathcal{V}|}{\delta}}+\frac{4}{3t}\log\frac{|\mathcal{V}|}{\delta}. \tag{C.63}$$

Proof. We have $|\mathcal{T}_m(t)|\coloneqq\sum_{\tau=1}^{t}[m(\tau)=m]$, where $\forall t,\forall m\in\mathcal{V},\mathbb{P}[m(t)=m]=\frac{1}{|\mathcal{V}|}$, 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 $w\geq 0$,

$$\mathbb{P}\left[\frac{|\mathcal{T}_m(t)|}{t}\geq\frac{1}{|\mathcal{V}|}+w\right]\leq\exp\left(-\frac{tw^2}{2(1-\frac{1}{|\mathcal{V}|})(\frac{1}{|\mathcal{V}|}+\frac{w}{3})}\right)\leq\exp\left(-\frac{tw^2}{2(\frac{1}{|\mathcal{V}|}+\frac{w}{3})}\right)$$

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

$$
\begin{aligned}
& \mathbb{P}\left[\frac{|\mathcal{T}_m(t)|}{t} \geq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}} + \frac{4}{3t}\log\frac{1}{\delta}\right] \leq \delta \\
\Longrightarrow\ & \mathbb{P}\left[\max_{m\in\mathcal{V}}\frac{|\mathcal{T}_m(t)|}{t} \geq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{\frac{1}{|\mathcal{V}|}\log\frac{1}{\delta}}{t}} + \frac{4\log\frac{1}{\delta}}{3t}\right] \leq |\mathcal{V}|\delta
\end{aligned}
$$

The result is obtained by adjusting the value of δ. □

Theorem C.2.20 (Regret bound, generalization of Theorem 5.5.10)**.** *Let the mean horizon per node be $\overline{T} = \frac{T}{|\mathcal{V}|}$. Under assumptions5.3.1 to 5.3.3 and 5.5.4 and $\kappa > 0$, the expected regret of the Network Lasso Bandit algorithm is upper bounded as follows:*

$$
\begin{aligned}
\mathcal{R}(\overline{T}) = \mathcal{O}\Bigg(& \frac{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log\big(\overline{T}|\mathcal{V}|\big)} + \sqrt[4]{|\mathcal{V}|\log\big(\overline{T}|\mathcal{V}|\big)}\right) \\
& + \frac{1}{A}\log(d|\mathcal{V}|) + \sqrt{|\mathcal{V}|}\Bigg),
\end{aligned}
$$

with $A = \dfrac{3\gamma^2 \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} + 2\sqrt{2}\gamma}$.

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:

$$G_{\text{pro}}(t) \coloneqq \left\{ \frac{1}{t} \|\boldsymbol{K}\|_F \leq \frac{\alpha(t)}{\alpha_0} \right\} \tag{C.64}$$

- $G_{\text{sel}}(t)$ is the event that the number of selections of all tasks is bounded by its expected value up to a small constant $\rho(t)$

$$G_{\text{sel}}(t) \coloneqq \left\{ \max_{m \in \mathcal{V}} \frac{|\mathcal{T}_m(t)|}{t} \leq \frac{1}{|\mathcal{V}|} + \frac{\rho(t)}{t} \right\} \tag{C.65}$$

- $G_{\text{RE}}(t)$ is the event that the empirical multi-task Gram matrix $\frac{1}{t}\boldsymbol{A}_{\mathcal{V}}(t)$ satisfies the RE condition.

$$G_{\text{RE}}(t) \coloneqq \left\{ \frac{1}{t}\boldsymbol{A}_{\mathcal{V}}(t) \text{ verifies the RE condition with constants } \kappa, \hat{\phi} \right\} \tag{C.66}$$

Event $G_{\text{pro}}(t)$ is the most straightforward to cover since our bound on the empirical process given in Lemma C.2.5 holds with a probability of at least $1 - \delta(t)$, thus:

$$\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{\text{sel}}(t)\right] \leq \delta(t), \tag{C.67}$$

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 $G_{\text{sel}}(t)$ can be determined using Bernstein's inequality:

From Lemma C.2.19 we can select $\rho(t) = 2\sqrt{\frac{t}{|\mathcal{V}|} \log \frac{|\mathcal{V}|}{\delta_{\text{sel}}(t)}} + \frac{4}{3} \log \frac{|\mathcal{V}|}{\delta_{\text{sel}}(t)}$ as well as $\mathbb{P}\left[G_{\text{sel}}(t)^c\right] \leq \delta_{\text{sel}}(t)$.

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:

$$\begin{aligned}
\mathbb{E}\left[r(t)\right] &\leq \mathbb{E}\left[r(t) | G_{\text{sel}}(t)\right] + 2\mathbb{P}\left[G_{\text{sel}}(t)^c\right] \\
&\leq \mathbb{E}\left[r(t) | G_{\text{pro}}(t) \cap G_{\text{RE}}(t) \cap G_{\text{sel}}(t)\right] \\
&\quad + 2\left(\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{\text{sel}}(t)\right] + \mathbb{P}\left[G_{\text{RE}}(t)^c | G_{\text{sel}}(t)\right] + \mathbb{P}\left[G_{\text{sel}}(t)^c\right]\right),
\end{aligned} \tag{C.68}$$

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.68) 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:

$$\sum_{t=1}^{T} \mathbb{E}\left[r(t) | G_{\text{pro}}(t) \cap G_{\text{RE}}(t) \cap G_{\text{sel}}(t)\right] \leq \sum_{t=1}^{T} \left\| \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(t) \right\|_F$$

Taking the result of our oracle inequality in Theorem C.2.7, we point out that only $\alpha(t)$ is time dependent such that the rest of the terms can be pulled outside the sum:

$$\begin{aligned}
\sum_{t=1}^{T} \left\| \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(t) \right\|_F &\leq \sum_{t=1}^{T} 2\frac{\alpha_0 \sigma}{\hat{\phi}^2 \sqrt{t}} f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \sqrt{1 + 2b\sqrt{|\mathcal{V}| \log \frac{1}{\delta(t)}} + 2b \log \frac{1}{\delta(t)}} \\
&= \frac{2\alpha_0 \sigma}{\hat{\phi}^2} f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \sum_{t=1}^{T} \sqrt{\frac{1}{t} + \frac{2b}{t}\sqrt{2|\mathcal{V}| \log(t)} + \frac{4b}{t} \log(t)} \\
&\leq \frac{2\alpha_0 \sigma}{\hat{\phi}^2} f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \int_0^T \frac{1}{\sqrt{t}} + \sqrt{\frac{2b}{t}\left(\sqrt{2|\mathcal{V}| \log(T)} + 2\log(T)\right)}\, dt \\
&\leq \frac{2\alpha_0 \sigma}{\hat{\phi}^2} f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0) \\
&\times \left(2\sqrt{T} + \left(\frac{\sqrt{8T}}{|\mathcal{V}|} + 4\sqrt[4]{\frac{32 \log(|\mathcal{V}|T)T}{|\mathcal{V}|}} + \sqrt{\frac{16}{3} \log(|\mathcal{V}|T) \log(T)}\right)\right. \\
&\times \left.\left(\sqrt[4]{2|\mathcal{V}| \log(T)} + \sqrt{2\log(T)}\right)\right) \\
&= \mathcal{O}\left(\frac{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log\left(\overline{T}|\mathcal{V}|\right)} + \sqrt[4]{|\mathcal{V}| \log\left(\overline{T}|\mathcal{V}|\right)}\right)\right),
\end{aligned}$$

where

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

We upper bounded the sum with an integral i.e. $\sum_{t=1}^{T} g(t) \leq \int_0^T g(t)dt$ for monotonically decreasing functions $g(t)$ in the last inequality. Also b is the bound on the concentration of the fraction of observation per task provided by Lemma C.2.19.

For $t_0 = \sqrt{|\mathcal{V}|}$ we find by inserting the result to Lemma C.2.19 for all $t > t_0$:

$$\begin{aligned}\frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{1}{t|\mathcal{V}|}\log\frac{|\mathcal{V}|}{\delta}} + \frac{4}{3t}\log\frac{|\mathcal{V}|}{\delta} &\leq \frac{1}{|\mathcal{V}|} + 2\sqrt{\frac{2\log\left(|\mathcal{V}|\sqrt{|\mathcal{V}|}\right)}{\sqrt{|\mathcal{V}|}|\mathcal{V}|} + \frac{8\log\left(|\mathcal{V}|\sqrt{|\mathcal{V}|}\right)}{3\sqrt{|\mathcal{V}|}}} \\ &= \frac{1}{|\mathcal{V}|} + \frac{2}{\sqrt{|\mathcal{V}|}}\left[\sqrt{\frac{3}{\sqrt{|\mathcal{V}|}}\log(|\mathcal{V}|)} + 2\log(|\mathcal{V}|)\right] \\ &= \mathcal{O}\left(\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}}\right) = b.\end{aligned}$$

Finally we bound the sum over the instantaneous regret term for the bad events:

$$\sum_{t=1}^{T} 2\left(\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{\text{sel}}(t)\right] + \mathbb{P}\left[G_{\text{RE}}(t)^c | G_{\text{sel}}(t)\right] + \mathbb{P}\left[G_{\text{sel}}(t)^c\right]\right)$$

By construction, we have $\max(\mathbb{P}\left[G_{\text{pro}}(t)^c | G_{\text{sel}}(t)\right], \mathbb{P}\left[G_{\text{sel}}(t)^c\right]) \leq \delta(t) = \frac{1}{t^2}$. Hence,

$$\sum_{t=1}^{T} \mathbb{P}\left[G_{\text{pro}}(t)^c | G_{\text{sel}}(t)\right] + \mathbb{P}\left[G_{\text{sel}}(t)^c\right] \leq 2\sum_{t=1}^{T}\frac{1}{t^2} \leq 2\left(1 + \int_1^T \frac{dt}{t^2}\right) \leq 4 \tag{C.69}$$

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 $t_0 \geq 1$

$$\begin{aligned}\sum_{t=t_0}^{T} \mathbb{P}\left[G_{\text{RE}}(t)^c | G_{\text{sel}}(t)\right] &\leq 6d|\mathcal{V}| \sum_{t=t_0}^{T} \exp(-At) \quad \text{(by Theorem C.2.18)} \\ &\leq 6d|\mathcal{V}| \frac{e^{-At_0}}{1-e^{-A}} \leq 6d|\mathcal{V}| e^{-At_0}\left(1 + \frac{1}{A}\right) \\ &\leq 6d|\mathcal{V}| e^{-At_0}\left(1 + \frac{1}{A}\right)\end{aligned}$$

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 $\sum_{t=t_0}^{T} \mathbb{P}\left[G_{\text{RE}}(t)^c | G_{\text{sel}}(t)\right] \leq u$. 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{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)\right) \\
&\leq \left\lceil\sqrt{|\mathcal{V}|}\right\rceil + \left\lceil\frac{1}{A}\log\left(\frac{6d|\mathcal{V}|(1+\frac{1}{A})}{u}\right)\right\rceil + 2u + 8 \\
&+ \mathcal{O}\left(\frac{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\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{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\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{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)\right) \\
&= \mathcal{O}\Bigg(\sqrt{|\mathcal{V}|} + \frac{1}{A}\log(d|\mathcal{V}|) \\
&+ \frac{\alpha_0 f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\sqrt{\overline{T}}}{\hat{\phi}^2}\left(\sqrt{|\mathcal{V}|} + \sqrt{\log(\overline{T}|\mathcal{V}|)} + \sqrt[4]{|\mathcal{V}|\log(\overline{T}|\mathcal{V}|)}\right)\Bigg) \\
&= \mathcal{O}\Bigg(\sqrt{|\mathcal{V}|} + \frac{1}{A}\log(d|\mathcal{V}|) \\
&+ \frac{\alpha_0 \nu\omega f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)\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)\Bigg),
\end{aligned}
$$

where we set $u = \frac{1}{2A}$ in the third inequality.

□

Proof of Corollary 5.5.11. Assuming $\frac{w(\partial\mathcal{P})(\psi+2\kappa)}{\min\limits_{\mathcal{C}\in\mathcal{P}}\sqrt{c_{\mathcal{G}}(\mathcal{C})}} \leq \Omega$, with some positive constant $\Omega < 1$ and setting $\alpha_0 = \frac{1}{\psi w(\partial\mathcal{P})}$ then the term $f(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0)$ can be bounded as:

$$
\begin{aligned}
f\left(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0 = \frac{1}{\psi w(\partial\mathcal{P})}\right) &= a_2 \left(\frac{a_2}{a_1 \min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}} + 1\right) \\
&= \left(w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)\right)\left(\frac{w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - w(\partial\mathcal{P})(\psi + 2\kappa)} + 1\right) \\
&= \mathcal{O}\left(\frac{w(\partial\mathcal{P})^2 \max\limits_{\mathcal{C}\in\mathcal{P}} \iota_{\mathcal{G}}(\mathcal{C})}{(1-\Omega)\min\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})}}\right).
\end{aligned}
$$

Ω acts as a threshold for the quality of any graph the satisfy this bound. Similarly we can find a bound on the term $\frac{1}{A}$:

$$
\begin{aligned}
\frac{1}{A} &= \frac{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} + \frac{1}{\sqrt{2}}\left(1 + \frac{a_2}{a_1}\right)^{-2}}{\frac{3}{4}\left(1 + \frac{a_2}{a_1}\right)^{-4} \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))} \\
&\leq \frac{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} + \frac{1}{\sqrt{2}}\left(1 + \frac{w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{1-\Omega}\right)^{-2}}{\frac{3}{4}\left(1 + \frac{w(\partial\mathcal{P})(\psi + \sqrt{2}\kappa) \max\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{1-\Omega}\right)^{-4} \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))} \\
&= \mathcal{O}\left(\frac{\left(1 + \frac{w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max\limits_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{1-\Omega}\right)^2}{\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}\right) \\
&= \mathcal{O}\left(\frac{w(\partial\mathcal{P})^2 \max\limits_{\mathcal{C}\in\mathcal{P}} \iota_{\mathcal{G}}(\mathcal{C})}{(1-\Omega)^2 \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}\right)
\end{aligned}
$$

Inserting the terms into the regret bound yields the final result.

□

Proof of Corollary 5.5.12. With $\alpha_0 = \frac{1}{\psi w(\partial\mathcal{P})}$ we have for $\alpha_0 f\left(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0\right)$ first:

$$\begin{aligned}
\alpha_0 f\left(\mathcal{G}, \boldsymbol{\Theta}, \alpha_0\right) &= \frac{1}{\psi w(\partial\mathcal{P})}\left(w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)\right) \\
&\quad \times \left(\frac{w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - w(\partial\mathcal{P})(\psi + 2\kappa)} + 1\right) \\
&= \frac{\left(\psi + \sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{\psi}\left(\frac{w(\partial\mathcal{P})\left(\psi + \sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}\right)}{\min_{\mathcal{C}\in\mathcal{P}} \sqrt{c_{\mathcal{G}}(\mathcal{C})} - w(\partial\mathcal{P})(\psi + 2\kappa)} + 1\right) \\
&= 1 + \frac{\sqrt{2}\kappa \max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})}}{\psi} = 1.
\end{aligned}$$

Where we used $\max_{\mathcal{C}\in\mathcal{P}} \sqrt{\iota_{\mathcal{G}}(\mathcal{C})} = 0$ in the last equality for the zero boundary case. For $\frac{1}{A}$ we have:

$$\frac{1}{A} = \frac{6\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} + \frac{1}{\sqrt{2}}\left(1 + \frac{a_2}{a_1}\right)^{-2}}{\frac{3}{4}\left(1 + \frac{a_2}{a_1}\right)^{-4} \min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))} = \frac{8\frac{\log(|\mathcal{V}|)}{\sqrt{|\mathcal{V}|}} + \frac{\sqrt{2^3}}{3}}{\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C}))}.$$

Inserting the terms into the regret bound yields the final result.

□

Proof of Corollary 5.5.13. In the case of fully connected clusters, we know from Proposition C.1.3 that the topologival centrality index of any cluster is given by:

$$c_{\mathcal{G}}(\mathcal{C}) = \frac{|\mathcal{C}|^2}{|\mathcal{C}| - 1} > |\mathcal{C}|$$

This allows us to lower bound the following term:

$$\min_{\mathcal{C}\in\mathcal{P}}(\tilde{c}_{\mathcal{G}}(\mathcal{C}) \wedge \tilde{c}^2_{\mathcal{G}}(\mathcal{C})) > |\mathcal{C}|$$

Inserting both lower bounds into our result of Corollary 5.5.11 yields the final result.

□

C.3 Additional experimental details

C.3.1 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.3.2 Solving the Network Lasso problem

We implement the Primal-Dual algorithm proposed in Jung (2020) 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.3.3 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 and construct the Laplacian $\boldsymbol{L}$ of the resulting graph $\mathcal{G}$. We measure the minimum topological centrality index $\min_{1\leq i\in 200}(\boldsymbol{L}_{\mathcal{C}}^{\dagger})_{ii}^{-1}$, and the algebraic connectivity, i.e. the minimum non-null eigenvalue of $\boldsymbol{L}$.

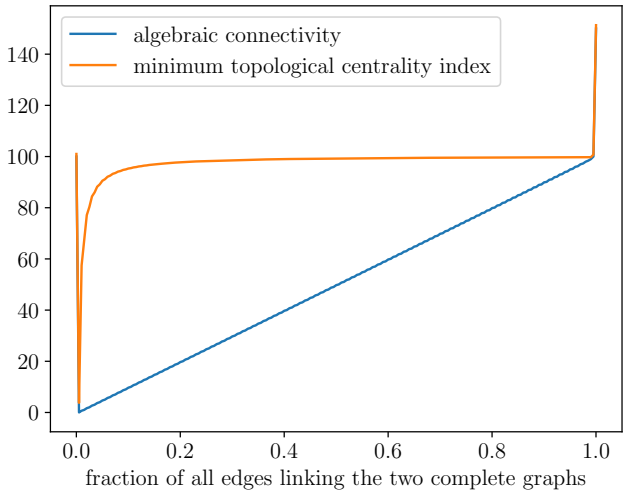


Figure C.1: Minimum Topological centrality index vs Algebraic Connectivity, for a graph formed by connecting two fully connected initial graphs $\mathcal{G}_1, \mathcal{G}_2$ with size 100 each.

Appendix D

Piecewise-Stationary Combinatorial Semi-Bandit with Causally Related Rewards

D.1 Proof of Theorem 6.4.2

The theoretical analysis relies on the provided regret upper bound for the CSB problem with causally related rewards in the stationary environment Nourani-Koliji *et al.* (2022). In addition, we used theoretical analysis of regret Zhou *et al.* (2020) 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 D.1.1. *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 Zhou *et al.* (2020). For the proof of Theorem 6.4.2 we explicitly need the false alarm probability in the stationary scenario;

Lemma D.1.2 (Lemma 4.5 in Zhou *et al.* (2020))**.** *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 D.1.3 (Lemma 12 in Besson and Kaufmann (2019))**.** *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 D.1.4. (Azuma (1967)) *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 $D \geq 0$,*

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

And finally we require the upper regret bound of the stationary case in Nourani-Koliji *et al.* (2022):

Proof of Lemma 6.4.1 improved version of Theorem 1 in Nourani-Koliji et al. *(2022).* The proof follows mostly the work of Nourani-Koliji *et al.* (2022). The *index set* of a decision vector is defined as $\boldsymbol{x} \in \mathcal{X}$ by $\mathcal{I}(\boldsymbol{x}) = \{k \mid \boldsymbol{x}[k] \neq 0, \forall k \in [K]\}$ and the confidence bound of base arm k at time t is defined as $\boldsymbol{C}_t[k] = \sqrt{\frac{(m+1)\ln t}{n_{k,t}}}$. At each time t, we store the empirical average of instantaneous rewards $\hat{\mu}_{k,t}$ and the calculated confidence bounds $\boldsymbol{C}_t[k]$ of all base arms $k \in \{1, ..., K\}$ in vectors $\hat{\boldsymbol{\mu}}_t$ and $\boldsymbol{C}_t$, respectively. We have $\boldsymbol{U}_t = \hat{\boldsymbol{\mu}}_t + \boldsymbol{C}_t$. In order to make the proof more readable, we use the equivalent formulation: $\mathbf{1}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1}\text{diag}(\boldsymbol{U}_{t-1})\boldsymbol{x}_t = \mathbf{1}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_{t-1})^{-1}\text{diag}(\boldsymbol{x}_t)\boldsymbol{U}_{t-1}$. At each time t, we define the *selection index* for a decision vector $\boldsymbol{x} \in \mathcal{X}$ as $I_t(\boldsymbol{x}) = \mathbf{1}^\top(\mathbf{I} - \hat{\boldsymbol{W}}_{t-1})^{-1}\text{diag}(\boldsymbol{x})\boldsymbol{U}_{t-1}$. 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 $\boldsymbol{x} \in \mathcal{X}$, the counter $\mathcal{T}_{\boldsymbol{x}}(t)$ is used to represent the total number of times the decision vector $\boldsymbol{x}$ is selected up to time t. Finally, for each base arm $k \in [K]$, we define a counter $\mathfrak{T}_k(t)$ which is updated as follows. At each time t after the initialization phase that a suboptimal decision vector $\boldsymbol{x}_t$ is selected, there is at least one base arm $k \in [K]$ such that $k = \underset{k \in \mathcal{I}(\boldsymbol{x}_t)}{\text{argmin}}\, n_{k,t}$. In this case, if the base arm k is unique, we increment $\mathfrak{T}_k(t)$ 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

$$\begin{aligned}\mathcal{R}(T) &= T\psi(\boldsymbol{x}^*) - \sum_{t=1}^{T} \psi(\boldsymbol{x}_t) \\ &= \sum_{\boldsymbol{x}:\psi(\boldsymbol{x})<\psi(\boldsymbol{x}^*)} \Delta(\boldsymbol{x})\mathbb{E}[\mathcal{T}_{\boldsymbol{x}}(T)], \end{aligned} \tag{D.2}$$

with $\psi(\boldsymbol{x})$ as reward for superarm $\boldsymbol{x}$. 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 $\mathfrak{T}_k(t)$ 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 Gai *et al.* (2012)

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

which implies that

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

Therefore, we observe that

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

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

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

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

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

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

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

Based on the definition of $\mathfrak{T}_k(t)$, we have $\mathfrak{T}_k(t) \leq n_{k,t}$, $\forall k \in [K]$. Therefore, when $\mathfrak{T}_k(t) \geq l$, the following holds Gai *et al.* (2012).

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

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

kth element in $\mathcal{I}(\boldsymbol{x}_{t+1})$. Hence, we have

$$\begin{aligned}\mathfrak{T}_k(T) \leq l + \sum_{t=K}^{T} \mathbb{1}\Bigg\{ & \min_{0<n_{v_1},\ldots,n_{v_{|\mathcal{I}(\boldsymbol{x}^*)|}}\leq t} \sum_{j=1}^{|\mathcal{I}(\boldsymbol{x}^*)|} \boldsymbol{v}_{t+1}^\top[v_j](\hat{\mu}_{v_j,t} + \boldsymbol{C}_t[v_j]) \leq \\ & \max_{l\leq n_{u_1},\ldots,n_{u_{|\mathcal{I}(\boldsymbol{x}_{t+1})|}}\leq t} \sum_{j=1}^{|\mathcal{I}(\boldsymbol{x}_{t+1})|} \boldsymbol{u}_{t+1}^\top[u_j](\hat{\mu}_{u_j,t} + \boldsymbol{C}_t[u_j]) \Bigg\} \\ \leq l + \sum_{t=1}^{\infty} \sum_{n_{v_1}=1}^{t} \cdots \sum_{n_{v_{|\mathcal{I}(\boldsymbol{x}^*)|}}=1}^{t} & \sum_{n_{u_1}=l}^{t} \cdots \sum_{n_{u_{|\mathcal{I}(\boldsymbol{x}_{t+1})|}}=l}^{t} \\ \mathbb{1}\Bigg\{ & \sum_{j=1}^{|\mathcal{I}(\boldsymbol{x}^*)|} \boldsymbol{v}_{t+1}^\top[v_j](\hat{\mu}_{v_j,t} + \boldsymbol{C}_t[v_j]) \\ & \leq \sum_{j=1}^{|\mathcal{I}(\boldsymbol{x}_{t+1})|} \boldsymbol{u}_{t+1}^\top[u_j](\hat{\mu}_{u_j,t} + \boldsymbol{C}_t[u_j]) \Bigg\}. \end{aligned} \tag{D.9}$$

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

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

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

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

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

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

First, we consider eq. (D.11) by finding the upper bound for

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

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

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

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

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

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

Therefore, we have

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

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

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

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

$$\vdots$$

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

we conclude for any arm that:

$$\begin{aligned}&\mathbb{P}\Big[\boldsymbol{y}_t^\top[v_k](\hat{\mu}_{v_k,t}+\boldsymbol{C}_t[v_k])\leq \boldsymbol{y}_t^\top[v_k]\mu_{v_k,t}\Big]\\ &\quad\overset{(a)}{=}\mathbb{P}\Big[n_{v_k,t}(\hat{\mu}_{v_k,t}+\boldsymbol{C}_t[v_k])\leq n_{v_k,t}\mu_{v_k,t}\Big]\\ &\quad\overset{(b)}{\leq} e^{-(2/n_{v_k,t})n_{v_k,t}^2\boldsymbol{C}_t[v_k]^2}\\ &\quad\overset{(c)}{=} e^{-2(m+1)\ln t}\\ &\quad= t^{-2(m+1)},\end{aligned} \tag{D.21}$$

where (a) holds since $\boldsymbol{y}_t^\top[v_k]\geq 0$, $\forall k$, (b) follows from Lemma D.1.4, and (c) results from the definition of $\boldsymbol{C}_t$. Hence, for Event $\mathcal{I}$, we conclude that

$$\mathbb{P}\left[\mathcal{I}\right]\leq |\mathcal{I}(\boldsymbol{x}^*)|t^{-2(m+1)}\leq mt^{-2(m+1)}. \tag{D.22}$$

Now, we consider Event $\mathcal{II}$. Based on Theorem 1 in Bazerque *et al.* (2013), we know that we can identify the adjacency matrix $\boldsymbol{W}$ uniquely by K samples gathered during the initialization period of our proposed algorithm. This means that with probability 1, after the time point $t_{\text{init}} = K < \infty$, $\hat{\boldsymbol{W}}_t = \boldsymbol{W}$ holds for all $t > t_{\text{init}}$. Therefore, for $t > K$, Event $\mathcal{II}$ holds with probability 1.

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

$$\mathbb{P}\left[\mathbf{1}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_t)^{-1}\text{diag}(\boldsymbol{x}^*)(\hat{\boldsymbol{\mu}}_t + \boldsymbol{C}_t) \leq \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t\right] \leq mt^{-2(m+1)} \tag{D.23}$$

For eq. (D.12), we have similar results as follows.

$$\mathbb{P}\Big[\mathbf{1}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_t)^{-1}\text{diag}(\boldsymbol{x}_{t+1})(\hat{\boldsymbol{\mu}}_t - \boldsymbol{C}_t) \geq \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t\Big] \leq mt^{-2(m+1)}. \tag{D.24}$$

Finally, for eq. (D.13) we have

$$\begin{aligned}
&\mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\qquad - 2\mathbf{1}^\top(\boldsymbol{I} - \hat{\boldsymbol{W}}_t)^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{C}_t \\
&\overset{(a)}{=} \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\qquad - 2\sum_{j \in \mathcal{I}(\boldsymbol{x}_{t+1})} \omega_t^\top[j]\boldsymbol{C}_t[j] \\
&\overset{(b)}{=} \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\qquad - 2\sum_{j \in \mathcal{I}(\boldsymbol{x}_{t+1})} \omega_t^\top[j]\sqrt{\frac{(m+1)\ln t}{n_{j,t}}} \\
&\overset{(c)}{\geq} \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\qquad - 2mw_{\max}\sqrt{\frac{(m+1)\ln T}{l}} \\
&\overset{(d)}{\geq} \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\quad - \Delta_{\min} \\
&\overset{(e)}{\geq} \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}^*)\boldsymbol{\mu}_t - \mathbf{1}^\top(\boldsymbol{I} - \boldsymbol{W})^{-1}\text{diag}(\boldsymbol{x}_{t+1})\boldsymbol{\mu}_t \\
&\quad - \Delta(\boldsymbol{x}_{t+1}) = 0,
\end{aligned} \tag{D.25}$$

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 $\boldsymbol{C}_t[j]$ and l, respectively. (e) follows from the definition of $\Delta_{\min}$. Hence, we conclude that eq. (D.13) never happens.

By using eqs. (D.23) to (D.25), we achieve the following.

$$\begin{aligned}
\mathbb{E}[\mathfrak{T}_k(T)] &\leq \left\lceil \frac{4(m+1)\ln T}{(\frac{\Delta_{\min}}{m w_{\max}})^2} \right\rceil \\
&+ \sum_{t=1}^{\infty} \left[\sum_{n_{v_1}=1}^{t} \cdots \sum_{n_{v_m}=1}^{t} \sum_{n_{u_1}=l}^{t} \cdots \sum_{n_{u_m}=l}^{t} 2mt^{-2(m+1)} \right] \\
&\leq \frac{4w_{\max}^2 m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + m\sum_{t=1}^{\infty} 2t^{-2} \\
&\leq \frac{4w_{\max}^2 m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3}m. \qquad \text{(D.26)}
\end{aligned}$$

Therefore, the expected regret is upper bounded as

$$\begin{aligned}
\mathcal{R}(T) &\leq \Delta_{\max} \sum_{k=1}^{K} \mathbb{E}[\mathfrak{T}_k(T)] \\
&\leq \sum_{k=1}^{K} \Big[\frac{4w_{\max}^2 m^2(m+1)\ln T}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3}m \Big] \Delta_{\max} \\
&\leq \Big[\frac{4w_{\max}^2 m^2(m+1)K\ln T}{\Delta_{\min}^2} + K + \frac{\pi^2}{3}mK \Big] \Delta_{\max}. \qquad \text{(D.27)}
\end{aligned}$$

□

Proof of Theorem 6.4.2. 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:

$$\mathbb{E}\left[\mathcal{R}(T-\nu_1)\right] \leq \mathbb{E}\left[\mathcal{R}(T-\nu_1)|\mathcal{C}_1\right] + \Delta_{\max} T(1-\mathbb{P}(\mathcal{C}_1)) \tag{D.28}$$

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 eq. (D.28) 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:

$$\mathbb{E}\left[\mathcal{R}(T-\nu_2|\mathcal{C}_1)\right] \leq \mathbb{E}\left[\mathcal{R}(T-\nu_2)|\mathcal{C}_2\right] + \Delta_{\max} T(1-\mathbb{P}(\mathcal{F}_2 \cap D_2|\mathcal{C}_1)) \tag{D.29}$$

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

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

$$\mathcal{R}(T) \leq \sum_{g\in G}\sum_{i=1}^{N_g} \mathbb{E}\left[\mathcal{R}_g(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\mathcal{F}_i^g)|\mathcal{C}_{i-1}^g\right] \tag{D.30}$$

$$+ \mathbb{E}\left[\mathcal{R}_g(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\bar{\mathcal{F}}_i^g)|\mathcal{C}_{i-1}^g\right] \tag{D.31}$$

$$+ \Delta_{\max} T(1-\mathbb{P}(\mathcal{F}_i^g \cap D_i^g|\mathcal{C}_{i-1}^g)). \tag{D.32}$$

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 term in eqs. (D.30) to (D.32) 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 D.1.3. For the second term we have due to Lemma D.1.2:

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

.

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:

$$\begin{aligned}&\mathbb{E}\left[\mathcal{R}(\nu_i^g-\nu_{i-1}^g)\mathbb{1}(\mathcal{F}_i^g)|\mathcal{C}_{i-1}^g\right]\\ &\quad\leq K_g R_0(\nu_i^g-\nu_{i-1}^g)+\left[(\nu_i^g-\nu_{i-1}^g)\frac{p+N_{\boldsymbol{W}}K/T}{\zeta}+d+K_g+\frac{\pi^2}{3}mK_g\right]\Delta_{\max},\end{aligned}$$

where we have the result from Lemma 6.4.1
$R_0(T)=\frac{4\omega_{\max}^2m^2(m+1)\log(T)}{\Delta_{\min}^2}\Delta_{\max}$ 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:

$$
\begin{aligned}
\mathcal{R}(T) &\leq \sum_{g\in G}\sum_{i=1}^{N_g}\Big[K_g R_0(\nu_i^g-\nu_{i-1}^g)+\left[(\nu_i^g-\nu_{i-1}^g)p/\zeta+d\right]\Delta_{\max}\\
&\quad+[(\nu_i^g-\nu_{i-1}^g)K_g\delta+(K_g+1)T\delta+K_g+\frac{\pi^2 mK_g}{3}]\Delta_{\max}\Big]\\
&\quad+N_{\boldsymbol{W}}K\Delta_{\max}\\
&\leq \sum_{g\in G}\Big[N_gK_gR_0(T)+\Delta_{\max}\delta T(K_g+N_g+N_gK_g)\\
&\quad+\left(Tp/\zeta+dN_g+K_gN_g+\frac{\pi^2 mK_gN_g}{3}\right)\Delta_{\max}\Big]\\
&\quad+N_{\boldsymbol{W}}K\Delta_{\max}\\
&=\sum_{g\in G}\left[N_gK_gR_0(T)+(\delta T+1+\frac{\pi^2 m}{3})N_gK_g\Delta_{\max}\right]\\
&\quad+\left(Tp+dN_G+\delta T(K+N_G)+N_{\boldsymbol{W}}K\right)\Delta_{\max}
\end{aligned}
$$

□

D.1.1 Proof of Theorem 6.4.4

For the proof of the Corollary we make use of Assumption D.1.1.

Proof of Corollary 6.4.4. We insert $d \leq \frac{K\log T}{p\left(\Delta_{\min}^{\text{change}}\right)^2}$ into our expression of Theorem 6.4.2:

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

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

$$
\begin{aligned}
\mathcal{R}(T) \leq & \sum_{g \in G} N_g K_g \left[\frac{4\omega_{\max}^2 m^2 (m+1) \log(T)}{\Delta_{\min}^2} + 1 + \frac{\pi^2}{3} m \right] \Delta_{\max} \\
& + \left[K + N_G + \sqrt{N_G K T \log T} + \frac{N_G \sqrt{K T \log T}}{\sqrt{N_G} \left(\Delta_{\min}^{\text{change}} \right)^2} \right] \Delta_{\max} \\
& + K N_{\boldsymbol{W}} \Delta_{\max} \\
= & \mathcal{O} \left(\left(\frac{\sum_{g \in G} N_g K_g \log T}{\Delta_{\min}} + \frac{\sqrt{N_G K T \log T}}{\left(\Delta_{\min}^{\text{change}} \right)^2} + K N_{\boldsymbol{W}} \right) \Delta_{\max} \right)
\end{aligned}
\tag{D.33}
$$

□

D.2 Additional Information Regarding Numerical Experiments

D.2.1 Synthetic Data Experiment

Figure D.1 provides the expected values of the base arms' instantaneous reward distributions for each distribution stationary segment in our synthetic data experiment. Figure D.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 $t = 5000, 10000, 15000, 20000$.

D.2.2 Real-Data Experiment

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

Table D.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	Sicilia
TOS	Toscana
UMB	Umbria
VDA	Valle d'Aosta / Vallée d'Aoste
VEN	Veneto

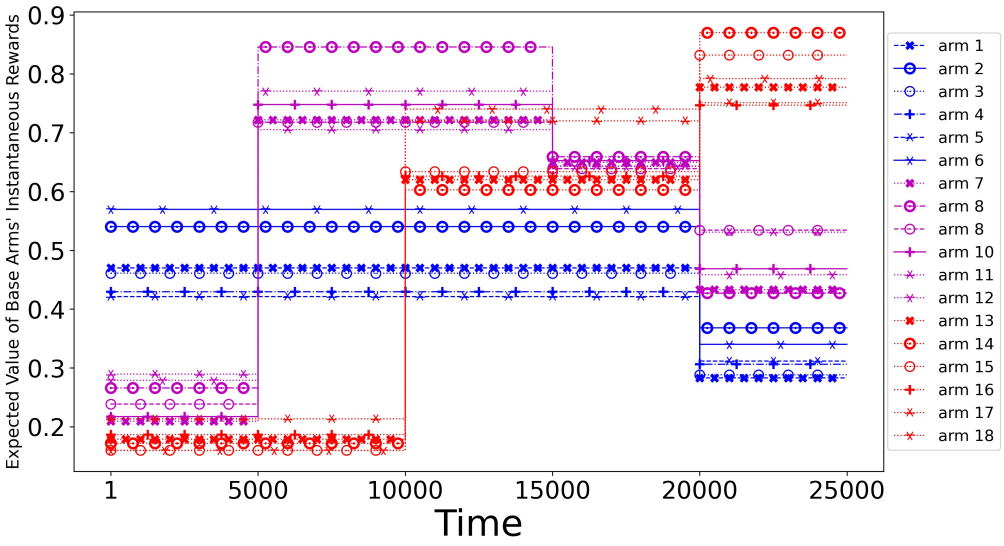


Figure D.1: Expected values of base arm's instantaneous rewards

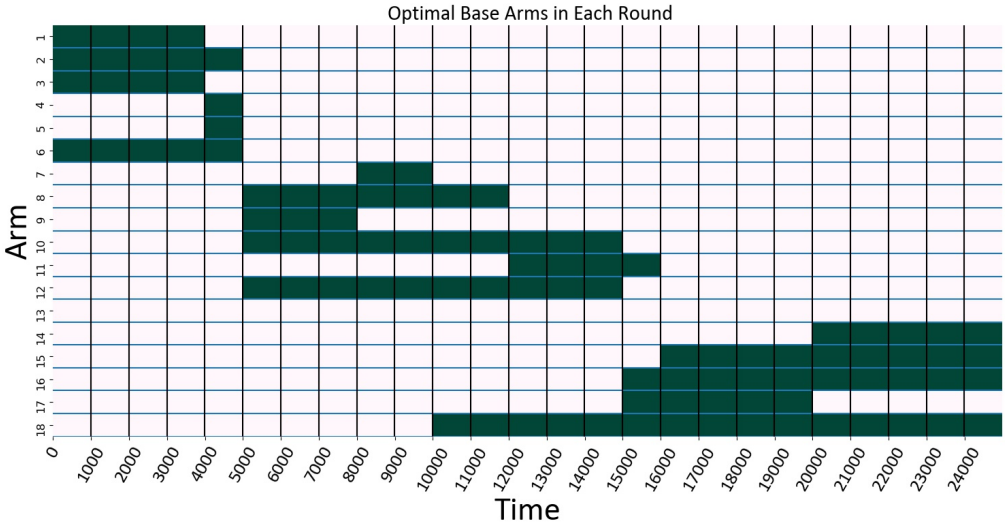


Figure D.2: Optimal super arms in synthetic dataset

Bibliography

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

Abeille, M. and Lazaric, A. (2017). Linear Thompson Sampling Revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*.

Aiolli, F. (2012). Transfer learning by kernel meta-learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*.

Amit, R. and Meir, R. (2018). Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proceedings of the 35th International Conference on Machine Learning*.

Amrallah, A., Mohamed, E. M., Tran, G. K., and Sakaguchi, K. (2020). Radio resource management aided multi-armed bandits for disaster surveillance system. In *Proc. 2020 International Conference on Emerging Technologies for Communications (ICETC2020), Virtual, K1-4*.

Ariu, K., Abe, K., and Proutiere, A. (2022). Thresholded Lasso Bandit. In *Proceedings of the 39th International Conference on Machine Learning*.

Atan, O., Ghoorchian, S., Maghsudi, S., and van der Schaar, M. (2023). Data-driven online recommender systems with costly information acquisition. *IEEE Trans. Serv. Comput.*

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The non-stochastic multiarmed bandit problem. *SIAM journal on computing*.

Auer, P., Gajane, P., and Ortner, R. (2019). Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*.

Azar, M., Lazaric, A., and Brunskill, E. (2013). Sequential transfer in multi-armed bandit with finite set of models. *Advances in Neural Information Processing Systems*.

Aziz, M., Kaufmann, E., and Riviere, M.-K. (2021). On multi-armed bandit designs for dose-finding clinical trials. *Journal of Machine Learning Research*.

Azizi, M., Duong, T., Abbasi-Yadkori, Y., György, A., Vernade, C., and Ghavamzadeh, M. (2022). Non-stationary bandits and meta-learning with a small set of optimal arms. *arXiv preprint arXiv:2202.13001*.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*.

Balcan, M.-F., Harris, K., Khodak, M., and Wu, Z. S. (2022). Meta-learning adversarial bandits. *arXiv preprint arXiv:2205.14128*.

Ban, Y. and He, J. (2021). Local clustering in contextual multi-armed bandits. In *Proceedings of the Web Conference 2021*.

Banerjee, D., Ghosh, A., Ray Chowdhury, S., and Gopalan, A. (2023). Exploration in linear bandits with rich action sets and its implications for inference. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*.

Bastani, H. and Bayati, M. (2019). Online Decision Making with High-Dimensional Covariates. *Operations Research*.

Bastani, H., Simchi-Levi, D., and Zhu, R. (2019). Meta Dynamic Pricing: Transfer Learning Across Experiments.

Bastani, H., Bayati, M., and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Manage. Sci.*

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*.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*.

Bazerque, J. A., Baingana, B., and Giannakis, G. B. (2013). Identifiability of sparse structural equation models for directed and cyclic networks. In *2013 IEEE Global Conference on Signal and Information Processing*.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*.

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*.

Besson, L., Kaufmann, E., Maillard, O.-A., and Seznec, J. (2022). Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*.

Bilaj, S., Dhouib, S., and Maghsudi, S. (2023). Hypothesis transfer in bandits by weighted models. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2022.*, pages 284–299.

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*.

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*.

Bouneffouf, D., Rish, I., and Aggarwal, C. (2020). Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*.

Boutilier, C., Hsu, C.-W., Kveton, B., Mladenov, M., Szepesvári, C., and Zaheer, M. (2020). Differentiable meta-learning of bandit policies. *Advances in Neural Information Processing Systems*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg.

Bush, R. R. and Mosteller, F. (1953). A stochastic model with applications to learning. *The Annals of Mathematical Statistics*.

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*.

Cardot, H. and Degras, D. (2015). Online principal component analysis in high dimension: Which algorithm to choose?

Cella, L. and Pontil, M. (2021). Multi-task and meta-learning with sparse linear bandits. In *Uncertainty in Artificial Intelligence*.

Cella, L., Lazaric, A., and Pontil, M. (2020). Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*.

Cella, L., Lounici, K., and Pontil, M. (2022). Meta representation learning with contextual linear bandits. *arXiv preprint arXiv:2205.15100*.

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*.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. (2013). A gang of bandits. *Advances in neural information processing systems*.

Cheeger, J. (1970). A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*.

Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*.

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*.

Chen, W., Wang, L., Zhao, H., and Zheng, K. (2021). Combinatorial semi-bandit in the non-stationary environment. In *Uncertainty in Artificial Intelligence*.

Cheng, X. and Maghsudi, S. (2023). Distributed consensus algorithm for decision-making in multi-agent multi-armed bandit. *arXiv preprint arXiv:2306.05998*.

Cheng, X., Pan, C., and Maghsudi, S. (2023). Parallel online clustering of bandits via hedonic game. In *International Conference on Machine Learning*.

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*.

DeepSeek-AI, :, Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., Ge, R., Guan, K., Guo, D., Guo, J., Hao, G., Hao, Z., He, Y., Hu, W., Huang, P., Li, E., Li, G., Li, J., Li, Y., Li, Y. K., Liang, W., Lin, F., Liu, A. X., Liu, B., Liu, W., Liu, X., Liu, X., Liu, Y., Lu, H., Lu, S., Luo, F., Ma, S., Nie, X., Pei, T., Piao, Y., Qiu, J., Qu, H., Ren, T., Ren, Z., Ruan, C., Sha, Z., Shao, Z., Song, J., Su, X., Sun, J., Sun, Y., Tang, M., Wang, B., Wang, P., Wang, S., Wang, Y., Wang, Y., Wu, T., Wu, Y., Xie, X., Xie, Z., Xie, Z., Xiong, Y., Xu, H., Xu, R. X., Xu, Y., Yang, D., You, Y., Yu, S., Yu, X., Zhang, B., Zhang, H., Zhang, L., Zhang, L., Zhang, M., Zhang, M., Zhang, W., Zhang, Y., Zhao, C., Zhao, Y., Zhou, S., Zhou, S., Zhu, Q., and Zou, Y. (2024). DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv e-prints*.

DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W. L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Wang, X., Liu, X., Xie, X., Yu, X., Song, X., Zhou, X., Yang, X., Lu, X., Su, X., Wu, Y., Li, Y. K., Wei, Y. X., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Zheng, Y., Zhang, Y., Xiong, Y., Zhao, Y., He, Y., Tang, Y., Piao, Y., Dong, Y., Tan, Y., Liu, Y., Wang, Y., Guo, Y., Zhu, Y., Wang, Y., Zou, Y., Zha, Y., Ma, Y., Yan, Y., You, Y., Liu, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Huang, Z., Zhang, Z., Xie, Z., Hao, Z., Shao, Z., Wen, Z., Xu, Z., Zhang, Z., Li, Z., Wang, Z., Gu, Z., Li, Z., and Xie, Z. (2024). Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018). Learning to learn around a common mean. *Advances in Neural Information Processing Systems*.

Dhouib, S., Bilaj, S., Nourani-Koliji, B., and Maghsudi, S. (2025). Cluster agnostic network lasso bandits. *Transactions on Machine Learning Research*.

Dong, X., Thanou, D., Rabbat, M., and Frossard, P. (2019). Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*.

Du, S. S., Koushik, J., Singh, A., and Póczos, B. (2017). Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*.

Duan, L., Tsang, I. W., Xu, D., and Chua, T.-S. (2009). Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th annual international conference on machine learning*.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *Proceedings of The 33rd International Conference on Machine Learning*.

Easley, D., Kleinberg, J., *et al.* (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press Cambridge.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International Conference on Machine Learning*.

Fontan, A. and Altafini, C. (2021). On the properties of laplacian pseudoinverses. In *2021 60th IEEE Conference on Decision and Control (CDC)*.

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*.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*.

Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *International Conference on Machine Learning*.

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*.

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*.

Glowacka, D. *et al.* (2019). Bandit algorithms in information retrieval. *Foundations and Trends® in Information Retrieval*.

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*.

Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*.

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*.

He, X., Alesiani, F., and Shaker, A. (2019). Efficient and scalable multi-task regression on massive number of tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Herbster, M., Pasteris, S., Vitale, F., and Pontil, M. (2021). A gang of adversarial bandits. *Advances in Neural Information Processing Systems*.

Hsu, D., Kakade, S., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*.

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*.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

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*.

Jiang, W., Kwok, J., and Zhang, Y. (2022). Subspace learning for effective meta-learning. In *International Conference on Machine Learning*.

Jung, A. (2020). Networked Exponential Families for Big Data Over Networks. *IEEE Access*.

Jung, A. and Vesselinova, N. (2019). Analysis of network lasso for semi-supervised regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

Jung, A., Tran, N., and Mara, A. (2018). When Is Network Lasso Accurate? *Frontiers in Applied Mathematics and Statistics*.

Kassraie, P., Rothfuss, J., and Krause, A. (2022). Meta-Learning Hypothesis Spaces for Sequential Decision-making. *ArXiv*.

Kim, G.-S. and Paik, M. C. (2019). Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., and Pérez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*.

Kuzborskij, I. and Orabona, F. (2013). Stability and hypothesis transfer learning. In *International Conference on Machine Learning*.

Kuzborskij, I. and Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. *Machine Learning*.

Kveton, B., Szepesvari, C., Wen, Z., and Ashkan, A. (2015). Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*.

Kveton, B., Mladenov, M., Hsu, C.-W., Zaheer, M., Szepesvari, C., and Boutilier, C. (2020). Meta-learning bandit policies by gradient ascent. *arXiv e-prints*.

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*.

Labille, K., Huang, W., and Wu, X. (2021). Transferable contextual bandits with prior observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*.

Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*.

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*.

Li, S., Chen, W., and Leung, K.-S. (2019). Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*.

Liau, D., Song, Z., Price, E., and Yang, G. (2018). Stochastic multi-armed bandits in constant space. In *International Conference on Artificial Intelligence and Statistics*.

Liu, B., Wei, Y., Zhang, Y., Yan, Z., and Yang, Q. (2018a). Transferable contextual bandit for cross-domain recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, F., Lee, J., and Shroff, N. (2018b). A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maghsudi, S. and Hossain, E. (2016). Multi-armed bandits with application to 5g small cells. *IEEE Wireless Communications*.

Maillard, O.-A. (2019). Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*.

Maiti, A., Patil, V., and Khan, A. (2021). Multi-armed bandits with bounded arm-memory: Near-optimal guarantees for best-arm identification and regret minimization. *Advances in Neural Information Processing Systems*.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*.

Mellor, J. and Shapiro, J. (2013). Thompson sampling in switching environments with bayesian online change detection. In *Artificial intelligence and statistics*.

Mezzadri, F. (2006). How to generate random matrices from the classical compact groups. *arXiv preprint math-ph/0609050*.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*.

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*.

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*.

Nourani-Koliji, B., Bilaj, S., Balef, A. R., and Maghsudi, S. (2023). Piecewise-stationary combinatorial semi-bandit with causally related rewards. In *ECAI 2023*. IOS Press.

Oh, M.-H., Iyengar, G., and Zeevi, A. (2021). Sparsity-Agnostic Lasso Bandit. In *Proceedings of the 38th International Conference on Machine Learning*.

Ortelli, F. and van de Geer, S. (2019). Synthesis and analysis in total variation regularization. *arXiv preprint arXiv:1901.06418*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*.

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*.

Perrot, M. and Habrard, A. (2015). A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*.

Ranjan, G. and Zhang, Z.-L. (2013). Geometry of complex networks and topological centrality. *Physica A: Statistical Mechanics and its Applications*.

Ras, Z. W., Wieczorkowska, A., and Tsumoto, S. (2021). *Recommender Systems for Medicine and Music*. Springer.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*.

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. (2021). Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*.

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*.

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*.

Schur, F., Kassraie, P., Rothfuss, J., and Krause, A. (2022). Lifelong Bandit Optimization: No Prior and No Regret.

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*.

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*.

Silva, N., Werneck, H., Silva, T., Pereira, A. C., and Rocha, L. (2022). Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Systems with Applications*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*.

Soare, M., Alsharif, O., Lazaric, A., and Pineau, J. (2014). Multi-task linear bandits. In *NIPS2014 Workshop on Transfer and Multi-task Learning: Theory meets Practice*.

Stark, B., Knahl, C., Aydin, M., and Elish, K. (2019). A literature review on medicine recommender systems. *International journal of advanced computer science and applications*.

Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*.

Suk, J. and Kpotufe, S. (2021). Self-tuning bandits over unknown covariate-shifts. In *Algorithmic Learning Theory*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press.

Thanou, D., Dong, X., Kressner, D., and Frossard, P. (2017). Learning heat diffusion graphs. *IEEE Transactions on Signal and Information Processing over Networks*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*.

Thrun, S. (1998). Lifelong learning algorithms. In *Learning to learn*. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Tommasi, T., Orabona, F., and Caputo, B. (2014). Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tropp, J. (2011). Freedman's inequality for matrix martingales. *Electronic Communications in Probability*.

Trovo, F., Paladino, S., Restelli, M., and Gatti, N. (2020). Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*.

Valko, M., Munos, R., Kveton, B., and Kocák, T. (2014). Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*.

Van Mieghem, P., Devriendt, K., and Cetinay, H. (2017). Pseudoinverse of the laplacian and best spreader node in a network. *Physical Review E*.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*.

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*.

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*.

Wang, Y.-X., Sharpnack, J., Smola, A. J., and Tibshirani, R. J. (2016). Trend filtering on graphs. *Journal of Machine Learning Research*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Xu, X. and Zhao, Q. (2021). Memory-constrained no-regret learning in adversarial multi-armed bandits. *IEEE Transactions on Signal Processing*.

Yang, J., Yan, R., and Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th ACM international conference on Multimedia*.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. (2020a). Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*.

Yang, K. and Toni, L. (2018). Graph-based recommendation system. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*.

Yang, K. and Toni, L. (2020). Differentiable linear bandit algorithm. *arXiv preprint arXiv:2006.03000*.

Yang, K., Toni, L., and Dong, X. (2020b). Laplacian-regularized graph bandits: Algorithms and theoretical analysis. In *International Conference on Artificial Intelligence and Statistics*.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis—kahan theorem for statisticians. *Biometrika*.

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*.

Zhao, P., Hoi, S. C., Wang, J., and Li, B. (2014). Online transfer learning. *Artificial Intelligence*.

Zhao, P., Cai, L.-W., and Zhou, Z.-H. (2020). Handling concept drift via model reuse. *Machine Learning*.

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*.

Zhou, Q., Zhang, X., Xu, J., and Liang, B. (2017). Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*.

Zhu, Z. and Van Roy, B. (2023). Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.