Contextual Combinatorial Bandits with Probabilistically Triggered Arms

Xutong Liu [1] **Jinhang Zuo** [2 3] **Siwei Wang** [4] **John C.S. Lui** [1] **Mohammad Hajiesmaili** [2] **Adam Wierman** [3] **Wei Chen** [4]

Abstract

We study contextual combinatorial bandits with probabilistically triggered arms (C^2MAB-T) under a variety of smoothness conditions that capture a wide range of applications, such as contextual cascading bandits and contextual influence maximization bandits. Under the triggering probability modulated (TPM) condition, we devise the C^2-UCB-T algorithm and propose a novel analysis that achieves an $\tilde{O}(d\sqrt{KT})$ regret bound, removing a potentially exponentially large factor $O(1/p_{\min})$, where d is the dimension of contexts, $p_{\min}$ is the minimum positive probability that any arm can be triggered, and batch-size K is the maximum number of arms that can be triggered per round. Under the variance modulated (VM) or triggering probability and variance modulated (TPVM) conditions, we propose a new variance-adaptive algorithm VAC^2-UCB and derive a regret bound $\tilde{O}(d\sqrt{T})$, which is independent of the batch-size K. As a valuable by-product, our analysis technique and variance-adaptive algorithm can be applied to the CMAB-T and C^2MAB setting, improving existing results there as well. We also include experiments that demonstrate the improved performance of our algorithms compared with benchmark algorithms on synthetic and real-world datasets.

1. Introduction

The stochastic multi-armed bandit (MAB) problem is a classical sequential decision-making problem that has been widely studied (Robbins, 1952; Auer et al., 2002; Bubeck et al., 2012). As an extension of MAB, combinatorial multi-armed bandits (CMAB) have drawn attention due to fruitful applications in online advertising, network optimization, and healthcare systems (Gai et al., 2012; Kveton et al., 2015a; Chen et al., 2013; 2016a; Wang & Chen, 2017; Merlis & Mannor, 2019). CMAB is a sequential decision-making game between a learning agent and an environment. In each round, the agent chooses a combinatorial action that triggers a set of base arms (i.e., a super-arm) to be pulled simultaneously, and the outcomes of these pulled base arms are observed as feedback (typically known as semi-bandit feedback). The goal of the agent is to minimize the expected *regret*, which is the difference in expectation for the overall rewards between always playing the best action (i.e., the action with the highest expected reward) and playing according to the agent's own policy.

Motivated by large-scale applications with a huge number of items (base arms), there exists a prominent line of work that advances the CMAB model: the combinatorial contextual bandits (or C^2MAB for short) (Qin et al., 2014; Li et al., 2016; Takemura et al., 2021). Specifically, C^2MAB incorporates contextual information and adds the simple yet effective linear structure assumption to allow scalability, which provides regret bounds that are independent of the number of base arms m. Despite C^2MAB's success in leveraging contextual information for better scalability, existing works fail to formulate the general arm triggering process, which is essential to model a wider range of applications, e.g., cascading bandits (CB) and influence maximization (IM), and more importantly, they do not provide satisfying results for settings with probabilistically triggered arms. For example, Qin et al. (2014); Takemura et al. (2021) only consider the deterministic semi-bandit feedback for C^2MAB. Li et al. (2016); Wen et al. (2017) implicitly consider the arm triggering process for specific CB or IM applications but only gives sub-optimal results with unsatisfying factors (e.g., $1/p_{\min}, K$ that could be as large as the number of base arms), owing to loose analysis, weak conditions, or inefficient algorithms that explore the unknown parameters too conservatively.

To handle the above issues, we enhance the C^2MAB framework by considering an arm triggering process. Specifically, we propose the general framework of contextual combi-

[1]The Chinese University of Hong Kong, Hong Kong SAR, China [2]University of Massachusetts Amherst, MA, United States [3]California Institute of Technology, CA, United States [4]Microsoft Research, Beijing, China. Correspondence to: Xutong Liu <liuxt@cse.cuhk.edu.hk>, Wei Chen <weic@microsoft.com>, John C.S. Lui <cslui@cse.cuhk.edu.hk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Table 1. Summary of the main results for C^{2}MAB-T, and additional results for CMAB-T and C^{2}MAB.

C^{2}MAB-T	Algorithm	Condition	Coefficient	Regret Bound
	C^3-UCB (Li et al., 2016)*	1-norm	B_1	$O(B_1 d\sqrt{KT}\cdot\log T/p_{\min})$
(Main Result 1)	C^2-UCB-T (Algorithm 1)	1-norm TPM	B_1	$O(B_1 d\sqrt{KT}\cdot\log T)$
(Main Result 2)	VAC^2-UCB (Algorithm 2)	VM	B_v †	$O(B_v d\sqrt{T}\cdot\log T/\sqrt{p_{\min}})$
(Main Result 3)	VAC^2-UCB (Algorithm 2)	TPVM	B_v, $\lambda\geq 1$ ‡	$O(B_v d\sqrt{T}\cdot\log T)$
CMAB-T	**Algorithm**	**Condition**	**Coefficient**	**Regret Bound**
	BCUCB-T (Liu et al., 2022)	TPVM	B_v, $\lambda\geq 1$	$O(B_v\sqrt{m(\log K)T}\cdot\log T)$
(Additional Result 1)	BCUCB-T (Our new analysis)	TPVM	B_v, $\lambda\geq 1$	$O(B_v\sqrt{m(\log K)T}\cdot(\log T)^{1/2})$**
C^{2}MAB	**Algorithm**	**Condition**	**Coefficient**	**Regret Bound**
	C^{2}UCB (Qin et al., 2014)	2-norm	B_2 §	$O(B_2 d\sqrt{T}\log T)$
	C^{2}UCB (Takemura et al., 2021)	1-norm	B_1	$O(B_1 d\sqrt{KT}\log T)$
(Additional Result 2)	VAC^2-UCB (Algorithm 2)	VM	B_v	$O(B_v d\sqrt{T}\log T)$

* This work is specified for contextual combinatorial cascading bandits, without formally defining the arm triggering process.
† Generally, coefficient $B_v = O(B_1\sqrt{K})$ and the existing regret bound is improved when $B_v = o(B_1\sqrt{K})$
‡ λ is a coefficient in TPVM condition: when λ is larger, the condition is stronger with smaller regret but can include less applications.
** We also show improved distribution-dependent regret bound in Appendix C; § Almost all applications satisfy $B_2 = \Theta(B_1\sqrt{K})$.

natorial bandits with probabilistically triggered arms (or C^{2}MAB-T for short). At the base arm level, C^{2}MAB-T uses a time-varying feature map $\boldsymbol{\phi}_t$ to model the contextual information at each round t, and the mean outcome of each arm $i\in[m]$ is a linear product of the feature vector $\boldsymbol{\phi}_t(i)\in\mathbb{R}^d$ and an unknown vector $\theta^*\in\mathbb{R}^d$ (where $d\ll m$ to handle large-scale applications). At the (combinatorial) action level, inspired by the non-contextual CMAB with probabilistically triggered arms (or CMAB-T) works (Chen et al., 2016b; Wang & Chen, 2017; Liu et al., 2022), we formally define an arm-triggering process to cover more general feedback models such as semi-bandit, cascading, and probabilistic feedback. We also inherit smoothness conditions for the non-linear reward function to cover different application scenarios, such as CB, IM, and online probabilistic maximum coverage (PMC) problems (Chen et al., 2016a; Wang & Chen, 2017). With this formulation, C^{2}MAB-T retains C^{2}MAB's scalability while also enjoying CMAB-T's rich reward functions and general feedback models.

Contributions. Our main results are shown in Table 1.

First, we study C^{2}MAB-T under the triggering probability modulated (TPM) smoothness condition, a condition introduced by Wang & Chen (2017) to remove a factor of $1/p_{\min}$ in the pioneer CMAB-T work (Chen et al., 2016a). This result follows a similar vein by devising C^2-UCB-T algorithm and proving a $\tilde{O}(d\sqrt{KT})$ regret, which removes a $1/p_{\min}$ factor for prior contextual CB applications (Li et al., 2016) (Main Result 1 in Table 1). The key technical challenge is that the triggering group (TG) analysis (Wang & Chen, 2017) for CMAB-T cannot handle the triggering probability determined by time-varying contexts. To tackle this issue, we devise a new technique, called the triggering probability equivalence (TPE), which links the triggering probabilities with the random triggering event under expectation. In this way, we no longer need to bound the regret caused by possibly triggered arms, but only need to bound the regret caused by actually triggered arms. As a result, we can then directly apply the simple non-triggering C^{2}MAB analysis to obtain the regret bound for C^{2}MAB-T. In addition, our TPE can reproduce the results for CMAB-T in a similar way.

Second, we study the C^{2}MAB-T under the variance modulated (VM) smoothness condition (Liu et al., 2022), in light of the recent variance-adaptive algorithms to remove the batch size dependence $O(\sqrt{K})$ for CMAB-T (Merlis & Mannor, 2019; Liu et al., 2022; Vial et al., 2022). We propose a new variance-adaptive algorithm VAC^2-UCB and prove a batch-size independent regret $\tilde{O}(d\sqrt{T/p_{\min}})$ under VM condition (Main Result 2 in Table 1). The main technical difficulty is to deal with the unknown variance. Inspired by Lattimore et al. (2015), we use the UCB/LCB value to construct an optimistic variance and on top of that, we prove a new concentration bound to incorporate the triggered arms and optimistic variance to get the desirable results.

Third, we investigate the stronger triggering probability and variance modulated (TPVM) condition (Liu et al., 2022) in order to remove the additional $1/\sqrt{p_{\min}}$ factor. The key challenge is that we cannot directly use TPE to link the *true* triggering probability with the random trigger event as before, since the TPVM condition only yields a mismatched triggering probability associated with the optimistic variance used in the algorithm. Our solution is to bound this additional mismatch by lower-order terms based on mild conditions on the triggering probability, which achieves the $\tilde{O}(d\sqrt{T})$ regret bounds (Main Result 3 in Table 1).

As a valuable by-product, our TPE analysis and VAC^2-UCB algorithm can be applied to non-contextual CMAB-T and C^{2}MAB, improving the existing results by a factor $\sqrt{\log T}$ (Additional Result 1 in Table 1) and $\sqrt{K}$ (Additional Result

2 in Table 1), respectively. Our empirical results on both synthetic and real data demonstrate that the VAC^2-UCB algorithm outperforms the state-of-art variance-agnostic and variance-aware bandit algorithms in the linear cascading bandit application that satisfies the TPVM condition.

Related Work. The stochastic CMAB model has received significant attention. The literature was initiated by (Gai et al., 2012). Since, its regret has been improved by Kveton et al. (2015b), Combes et al. (2015), Chen et al. (2016a). There exist two prominent lines of work in the literature related to our study: contextual CMAB and CMAB with probabilistically triggered arms (C^{2}MAB and CMAB-T).

For C^{2}MAB works, Qin et al. (2014) is the first study, which proposes C^{2}UCB algorithm that considers reward functions under 2-norm B_2 smoothness condition. Then Takemura et al. (2021) replaces the 2-norm smoothness condition with a new 1-norm B_1 smoothness condition, and proves a $O(B_1 d\sqrt{KT}\log T)$ regret bound. In this work, we extend the C^{2}MAB with triggering arms to cover more application scenarios (e.g., contextual CB and contextual IM). Moreover, we further consider the stronger VM condition and propose a new variance-adaptive algorithm that can achieve a $\sqrt{K}$ factor improvement in the regret upper bound for applications like PMC.

For CMAB-T works, Chen et al. (2016a) is the first work that considers the arm triggering process to cover CB and IM applications. The authors propose the CUCB algorithm, and give an $O(B_1\sqrt{mKT\log T}/p_{\min})$ regret bound under 1-norm B_1 smoothness condition. Then Wang & Chen (2017) proposes the stronger 1-norm triggering probability modulated (TPM) B_1 smoothness condition, and use the triggering group (TG) analysis to remove a $1/p_{\min}$ factor in the previous regret. Recently, Liu et al. (2022) incorporates the variance information, and proposes the variance-adaptive algorithm BCUCB-T, which also uses the TG analysis and further reduces the regret's dependency on batch-size from $O(K)$ to $O(\log K)$ under the new variance and triggering probability modulated (TPVM) condition. The smoothness conditions considered in this work are mostly inspired by the above works, but directly following their algorithm and TG analysis fail to obtain any meaningful result for our C^{2}MAB-T setting. Conversely, our new TPE analysis can be applied to CMAB-T, reproducing CMAB-T's result under the 1-norm TPM condition, and improving a factor of $(\sqrt{\log T})$ under the TPVM condition.

There are also many studies considering specific applications under the C^{2}MAB-T framework (by unifying C^{2}MAB and CMAB-T), including contextual CB (Li et al., 2016; Vial et al., 2022), contextual IM (Wen et al., 2017), etc. One can see that these applications fit into our framework by verifying that they satisfy the TPM, VM, or TPVM conditions; thus achieving improved results regarding $K, p_{\min}$ factors. We defer a detailed theoretical and empirical comparison to Sections 3 to 5. Zuo et al. (2022) study the online competitive IM and also uses C^{2}MAB-T to denote their contextual setting. However, their meaning of "contexts" is the action of the competitor, which acts at the action level and only affects the reward function (or regret) but not the base arms' estimation. This is very different from our setting, where contexts act at the base arm level and hence one cannot directly apply their results.

2. Problem Setting

We study contextual combinatorial bandits with probabilistically triggered arms (C^{2}MAB-T). We use $[n]$ to represent set $\{1, ..., n\}$. We use boldface lowercase letters and boldface CAPITALIZED letters for column vectors and matrices, respectively. $\|\boldsymbol{x}\|_p$ denotes the ℓ_p norm of vector $\boldsymbol{x}$. For any symmetric positive semi-definite (PSD) matrix $\boldsymbol{M}$ (i.e., $\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} \geq 0, \forall \boldsymbol{x}$), $\|\boldsymbol{x}\|_{\boldsymbol{M}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}}$ denotes the matrix norm of $\boldsymbol{x}$ regarding matrix $\boldsymbol{M}$.

We specify a C^{2}MAB-T problem instance using a tuple $([m], \mathcal{S}, \Phi, \Theta, D_{\text{trig}}, R)$, where $[m] = \{1, 2, ..., m\}$ is the set of base arms (or arms); $\mathcal{S}$ is the set of eligible actions where $S \in \mathcal{S}$ is an action;* Φ is the set of possible feature maps where any feature map $\boldsymbol{\phi} \in \Phi$ is a function $[m] \to \mathbb{R}^d$ that maps an arm to a d-dimensional feature vector (and w.l.o.g. we normalize $\|\boldsymbol{\phi}(i)\|_2 \leq 1$); $\Theta \subseteq R^d$ is the parameter space; D_{trig} is the probabilistic triggering function to characterize the arm triggering process (and feedback), and R is the reward function.

In C^{2}MAB-T, a learning game is played between a learning agent (or player) and the unknown environment in a sequential manner. Before the game starts, the environment chooses a parameter $\boldsymbol{\theta}^* \in \Theta$ unknown to the agent (and w.l.o.g. we also assume $\|\theta^*\|_2 \leq 1$). At the beginning of round t, the environment reveals feature vectors $(\boldsymbol{\phi}_t(1), ..., \boldsymbol{\phi}_t(m))$ for each arm, where $\boldsymbol{\phi}_t \in \Phi$ is the feature map known to the agent. Given $\boldsymbol{\phi}_t$, the agent selects an action $S_t \in \mathcal{S}$, and the environment draws Bernoulli outcomes $\boldsymbol{X}_t = (X_{t,1}, ...X_{t,m}) \in \{0,1\}^m$ for base arms†, with mean $\mathbb{E}[\boldsymbol{X}_{t,i}|\mathcal{H}_t] = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i)\rangle$ for each base arm i. Here $\mathcal{H}_t$ denotes the history before the agent chooses S_t and will be specified shortly after. Note that the outcome $\boldsymbol{X}_t$ is assumed to be conditional independent across arms given history $\mathcal{H}_t$, similar to previous works (Qin et al., 2014; Li et al., 2016; Vial et al., 2022). For convenience, we use $\boldsymbol{\mu}_t \triangleq (\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i)\rangle)_{i\in[m]}$ to denote the mean vector and $\mathcal{M} \triangleq \{\langle \boldsymbol{\theta}, \boldsymbol{\phi}(i)\rangle_{i\in[m]} : \boldsymbol{\phi} \in \Phi, \boldsymbol{\theta} \in \Theta\}$ to denote all possible mean vectors generated by Φ and Θ.

*$\mathcal{S}$ is a general action space. When $\mathcal{S}$ is a collection of subsets of $[m]$, we often refer to $S \in \mathcal{S}$ as a super arm.

†We assume $X_{t,i}$ are Bernoulli for the ease of exposition, yet our setting can handle any distribution with bounded $X_{t,i} \in [0, 1]$.

After the action S_t is played on the outcome $\boldsymbol{X}_t$, base arms in a random set $\tau_t \sim D_{\text{trig}}(S_t, \boldsymbol{X}_t)$ are triggered, meaning that the outcomes of arms in τ_t, i.e. $(X_{t,i})_{i\in\tau_t}$ are revealed as the feedback to the agent, and are involved in determining the reward of action S_t. At the end of round t, the agent will receive a non-negative reward $R(S_t, \boldsymbol{X}_t, \tau_t)$, determined by $S_t, \boldsymbol{X}_t$ and τ_t, and similar to (Wang & Chen, 2017), the expected reward is assumed to be $r(S_t; \boldsymbol{\mu}_t) \triangleq \mathbb{E}[R(S_t, \boldsymbol{X}_t, \tau_t)]$, a function of the unknown mean vector $\boldsymbol{\mu}_t$, where the expectation is taken over the randomness of $\boldsymbol{X}_t$ and $\tau_t \sim D_{\text{trig}}(S_t, \boldsymbol{X}_t)$. To this end, we can give the formal definition of the history $\mathcal{H}_t = (\boldsymbol{\phi}_s, S_s, \tau_s, (X_{s,i})_{i\in\tau_s})_{s<t} \bigcup \boldsymbol{\phi}_t$, which contains all information before round t, as well as the contextual information $\boldsymbol{\phi}_t$ at round t.

The goal of CMAB-T is to accumulate as much reward as possible over T rounds by learning the underlying parameter θ^*. The performance of an online learning algorithm A is measured by its *regret*, defined as the difference of the expected cumulative reward between always playing the best action $S_t^* \triangleq \text{argmax}_{S\in\mathcal{S}} r(S; \boldsymbol{\mu}_t)$ in each round t and playing actions chosen by algorithm A. For many reward functions, it is NP-hard to compute the exact S_t^* even when $\boldsymbol{\mu}_t$ is known, so similar to (Wang & Chen, 2017), we assume that the algorithm A has access to an offline (α, β)-approximation oracle, which for mean vector $\boldsymbol{\mu}$ outputs an action S such that $\Pr[r(S; \boldsymbol{\mu}) \geq \alpha \cdot r(S^*; \boldsymbol{\mu})] \geq \beta$. The T-round (α, β)-approximate regret is defined as

$$\text{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} (\alpha\beta \cdot r(S_t^*; \boldsymbol{\mu}_t) - r(S_t; \boldsymbol{\mu}_t))\right], \quad (1)$$

where the expectation is taken over the randomness of outcomes $\boldsymbol{X}_1, ..., \boldsymbol{X}_T$, the triggered sets $\tau_1, ..., \tau_T$, as well as the randomness of algorithm A itself.

Remark 1 (Difference with CMAB-T). C^{2}MAB-T strictly generalizes CMAB-T by allowing a probably time-varying feature map $\boldsymbol{\phi}_t$. Specifically, let $\boldsymbol{\theta}^* = (\mu_1, ..., \mu_m)$ and fix $\boldsymbol{\phi}_t(i) = \boldsymbol{e}_i$ where $\boldsymbol{e}_i \in \mathbb{R}^m$ is the one-hot vector with 1 at the i-th entry and 0 elsewhere, one can easily reproduce the CMAB-T setting in (Wang & Chen, 2017).

Remark 2 (Difference with C^{2}MAB). C^{2}MAB-T enhances the modeling power of prior C^{2}MAB (Qin et al., 2014; Takemura et al., 2021) by capturing the probabilistic nature of the feedback (v.s. the deterministic semi-bandit feedback). This enables a wider range of applications such as combinatorial CB, multi-layered network exploration, and online IM (Wang & Chen, 2017; Liu et al., 2022).

2.1. Key Quantities and Conditions

In the C^{2}MAB-T model, there are several quantities and assumptions that are crucial to the subsequent study. We define *triggering probability* $p_i^{\boldsymbol{\mu}, D_{\text{trig}}, S}$ as the probability that base arm i is triggered when the action is S, the mean vector is $\boldsymbol{\mu}$, and the probabilistic triggering function is D_{trig}. Since D_{trig} is always fixed in a given application context, we ignore it in the notation for simplicity, and use $p_i^{\boldsymbol{\mu},S}$ henceforth. Triggering probabilities $p_i^{\boldsymbol{\mu},S}$'s are crucial for the triggering probability modulated bounded smoothness conditions to be defined below. We define $\tilde{S}$ to be the set of arms that can be triggered by S, i.e.,$\{i \in [m] : p_i^{\boldsymbol{\mu},S} > 0, \text{ for any } \boldsymbol{\mu} \in \mathcal{M}\}$, the *batch size* K as the maximum number of arms that can be triggered, i.e., $K = \max_{S\in\mathcal{S}} |\tilde{S}|$, and $p_{\min} = \min_{i\in[m], \boldsymbol{\mu}\in\mathcal{M}, S\in\mathcal{S}, p_i^{\boldsymbol{\mu},S}>0} p_i^{\boldsymbol{\mu},S}$.

Owing to the nonlinearity and the combinatorial structure of the reward, it is essential to give some conditions for the reward function in order to achieve any meaningful regret bounds (Chen et al., 2013; 2016a; Wang & Chen, 2017; Merlis & Mannor, 2019; Liu et al., 2022). For C^{2}MAB-T, we consider the following conditions.

Condition 1 (Monotonicity). *We say that a C^{2}MAB-T problem instance satisfies monotonicity condition, if for any action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [0,1]^m$ such that $\mu_i \leq \mu_i'$ for all $i \in [m]$, we have $r(S; \boldsymbol{\mu}) \leq r(S; \boldsymbol{\mu}')$.*

Condition 2 (1-norm TPM Bounded Smoothness, (Wang & Chen, 2017)). *We say that a C^{2}MAB-T problem instance satisfies the triggering probability modulated (TPM) B_1-bounded smoothness condition, if for any action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [0,1]^m$, we have $|r(S; \boldsymbol{\mu}') - r(S; \boldsymbol{\mu})| \leq B_1 \sum_{i\in[m]} p_i^{\boldsymbol{\mu},S} |\mu_i - \mu_i'|$.*

Condition 3 (VM Bounded Smoothness, (Liu et al., 2022)). *We say that a C^{2}MAB-T problem instance satisfies the variance modulated (VM) (B_v, B_1)-bounded smoothness condition, if for any action $S \in \mathcal{S}$, mean vector $\boldsymbol{\mu}, \boldsymbol{\mu}' \in (0,1)^m$, for any $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^m$ s.t. $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$, we have $|r(S; \boldsymbol{\mu}') - r(S; \boldsymbol{\mu})| \leq B_v \sqrt{\sum_{i\in\tilde{S}} \frac{\zeta_i^2}{(1-\mu_i)\mu_i}} + B_1 \sum_{i\in\tilde{S}} |\eta_i|$.*

Condition 4 (TPVM Bounded Smoothness, (Liu et al., 2022)). *We say that a C^{2}MAB-T problem instance satisfies the triggering probability and variance modulated (TPVM) (B_v, B_1, λ)-bounded smoothness condition, if for any action $S \in \mathcal{S}$, mean vector $\boldsymbol{\mu}, \boldsymbol{\mu}' \in (0,1)^m$, for any $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^m$ s.t. $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$, we have $|r(S; \boldsymbol{\mu}') - r(S; \boldsymbol{\mu})| \leq B_v \sqrt{\sum_{i\in[m]} (p_i^{\boldsymbol{\mu},S})^\lambda \frac{\zeta_i^2}{(1-\mu_i)\mu_i}} + B_1 \sum_{i\in[m]} p_i^{\boldsymbol{\mu},S} |\eta_i|$.*

Condition 1 indicates the reward is monotonically increasing when the parameter $\boldsymbol{\mu}$ increases. Condition 2, 3 and 4 all bound the reward smoothness/sensitivity.

For Condition 2, the key feature is that the parameter change in each base arm i is modulated by the triggering probability $p_i^{\boldsymbol{\mu},S}$. Intuitively, for base arm i that is unlikely to be triggered/observed (small $p_i^{\boldsymbol{\mu},S}$), Condition 2 ensures that

a large change in μ_i (due to insufficient observation) only causes a small change (multiplied by $p_i^{\boldsymbol{\mu},S}$) in reward, which helps to save a $p_{\min}$ factor for non-contextual CMAB-T.

For Condition 3, intuitively if we ignore the denominator $(1-\mu_i)\mu_i$ of the leading B_v term, the reward change would be $O(B_v\sqrt{K}\Delta)$ when the amount of parameter change $|\mu_i'-\mu_i| = \Delta$ for each arm i. This introduces a $O(\sqrt{K})$ factor reduction in the reward change and translates to a $O(K)$ improvement in the regret, compared with $O(B_1 K\Delta)$ reward change when applying the non-triggering version of Condition 2 (i.e., $p_i^{\boldsymbol{\mu},\tilde{S}} = 1$ if $i \in \tilde{S}$ and $p_i^{\boldsymbol{\mu},\tilde{S}} = 0$ otherwise). However, for real applications, $B_1 = \Theta(B_1\sqrt{K})$ which cancels the $O(\sqrt{K})$ improvement. To reduce B_v coefficient, the leading B_v term is modulated by the inverse of the variance $V_i = (1-\mu_i)\mu_i$, and thus allow applications to achieve a B_v coefficient independent of K (or at least $B_v = o(B_1\sqrt{K})$, leading to significant savings in the regret bound for applications like PMC (Liu et al., 2022). The relation between Condition 2 and 3 is generally not comparable, but compared with Condition 2's non-triggering counterpart (i.e., 1-norm condition), Condition 3 is stronger.

Finally, for Condition 4, it combines both the triggering-probability modulation from Condition 2 and the variance modulation from Condition 3. The exponent λ of $p_i^{\boldsymbol{\mu},S}$ gives additional flexibility to trade-off between the strength of the condition and the regret, i.e., with a larger λ, one can obtain a smaller regret bound, while with a smaller λ, the condition is easier to satisfy to include more applications. In general, Condition 4 is stronger than Condition 2 and Condition 3, as the former degenerates to the other two conditions by setting $\boldsymbol{\zeta} = \mathbf{0}$ and the fact that $p_i^{\boldsymbol{\mu},S} \le 1$ for $i \in \tilde{S}$ and $p_i^{\boldsymbol{\mu},S} = 0$ otherwise, respectively. Conversely, by applying the Cauchy-Schwartz inequality, one can verify that if a reward function is TPM B_1-bounded smooth, then it is TPVM $(B_1\sqrt{K}/2, B_1, \lambda)$-bounded smooth for any $\lambda \le 2$ or similarly VM $(B_1\sqrt{K}/2, B_1)$-bounded smooth, respectively.

In light of the above conditions that significantly advance the non-contextual CMAB-T, the goal of subsequent sections is to design algorithms and conduct analysis to derive the (improved) results for the contextual setting. And later in Section 5, we demonstrate how these conditions are applied to applications, such as CB and online IM, to achieve both theoretical and empirical improvements. Due to the space limit, the detailed proofs are included in the Appendix.

3. Algorithm and Regret Analysis for C^{2}MAB-T under the TPM Condition

Our proposed algorithm C^2-UCB-T (Algorithm 1) is a generalization of the C^3-UCB algorithm originally designed for contextual combinatorial cascading bandits (Li et al., 2016). Our main contribution is to show an improved regret bound by a factor of $1/p_{\min}$ under the 1-norm TPM condition.

Algorithm 1 C^2-UCB-T: **C**ontextual **C**ombinatorial **U**pper **C**onfidence **B**ound Algorithm for C^{2}MAB-**T**

1: **Input:** Base arms $[m]$, dimension d, regularizer γ, failure probability $\delta = 1/T$, offline oracle ORACLE.
2: **Initialize:** Gram matrix $\boldsymbol{G}_1 = \gamma\boldsymbol{I}$, vector $\boldsymbol{b}_1 = \mathbf{0}$.
3: **for** $t = 1, ..., T$ **do**
4: $\quad\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1}\boldsymbol{b}_t$.
5: $\quad$**for** $i \in [m]$ **do**
6: $\quad\quad\bar{\mu}_{t,i} = \langle\boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t\rangle + \rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}$.
7: $\quad$**end for**
8: $\quad S_t = \text{ORACLE}(\bar{\mu}_{t,1}, ..., \bar{\mu}_{t,m})$.
9: $\quad$Play S_t and observe triggering arm set τ_t and observation set $(X_{t,i})_{i\in\tau_t}$.
10: $\quad\boldsymbol{G}_{t+1} = \boldsymbol{G}_t + \sum_{i\in\tau_t}\boldsymbol{\phi}_t(i)\boldsymbol{\phi}_t(i)^\top$.
11: $\quad\boldsymbol{b}_{t+1} = \boldsymbol{b}_t + \sum_{i\in\tau_t}\boldsymbol{\phi}_t(i)X_{t,i}$.
12: **end for**

Recall that we define the data about the history as $\mathcal{H}_t = (\boldsymbol{\phi}_s, S_s, \tau_s, (X_{s,i})_{i\in\tau_s})_{s<t} \bigcup \boldsymbol{\phi}_t$. Different from the CUCB algorithm (Wang & Chen, 2017) that directly estimates the mean $\boldsymbol{\mu}_{t,i}$ for each arm, Algorithm 1 estimates the underlying parameter $\boldsymbol{\theta}^*$ via a ridge regression problem over the history data $\mathcal{H}_t$. More specifically, we estimate $\boldsymbol{\theta}^*$ by solving the following ℓ_2-regularized least-square problem with regularization parameter $\gamma > 0$:

$$\hat{\boldsymbol{\theta}}_t = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta}\sum_{s<t}\sum_{i\in\tau_s}(\langle\boldsymbol{\theta}, \boldsymbol{\phi}_s(i)\rangle - X_{s,i})^2 + \gamma\left\|\boldsymbol{\theta}\right\|_2^2. \quad (2)$$

The closed form solution is precisely the $\hat{\theta}_t$ calculated in line 4, where the Gram matrix $\boldsymbol{G}_t = \sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top$ and the b-vector $\boldsymbol{b}_t = \sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)X_{s,i}$ are computed in line 10 and 11. We claim that $\hat{\theta}_t$ is a good estimator of θ^* by bounding their difference via the following lemma, which is also used in (Qin et al., 2014; Li et al., 2016).

Proposition 1 (Theorem 2, (Abbasi-Yadkori et al., 2011)). *Let* $\rho(\delta) = \sqrt{\log\left(\frac{(\gamma+KT/d)^d}{\gamma^d\cdot\delta^2}\right)} + \sqrt{\gamma}$, *then with probability at least* $1-\delta$, *for all* $t \in [T]$, $\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t} \le \rho(\delta)$.

Building on this, we construct an optimistic estimation of each arm's mean $\bar{\mu}_{t,i}$ in line 6, where $\rho(\delta)$ is in Proposition 1, $\langle\boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t\rangle$ and $\rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}$ are the empirical mean and confidence interval towards the direction $\boldsymbol{\phi}_t(i)$, respectively. As a convention, we clip $\bar{\mu}_{t,i}$ into $[0,1]$ if $\bar{\mu}_{t,i} > 1$ or $\bar{\mu}_{t,i} < 0$.

Thanks to Proposition 1, we have the following lemma for the desired amount of the base arm level optimism,

Lemma 1. *With probability at least* $1-\delta$, *we have* $\mu_{t,i} \le \bar{\mu}_{t,i} \le \mu_{t,i} + 2\rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}$ *for all* $i \in [m], t \in [T]$.

Proof. See Appendix A.2. ■

After computing the UCB values $\bar{\boldsymbol{\mu}}_t$, the agent selects action S_t via the offline oracle with $\bar{\boldsymbol{\mu}}_t$ as input. Then base arms in τ_t are triggered, and the agent receives observation set $(X_{t,i})_{i\in\tau_t}$ as feedback to improve future decisions.

Theorem 1. *For a C²MAB-T instance that satisfies monotonicity (Condition 1) and TPM smoothness (Condition 2) with coefficient B_1, C²-UCB-T (Algorithm 1) with an (α,β)-approximation oracle achieves an (α,β)-approximate regret bounded by* $O\left(B_1(\sqrt{d\log(KT/\gamma)}+\sqrt{\gamma})\sqrt{KTd\log(KT/\gamma)}\right)$.

Discussion. Looking at Theorem 1, we achieve an $O(B_1 d\sqrt{KT}\log T)$ regret bound when $d \ll K \le m \ll T$, which is independent of the number of arms m and the minimum triggering probability $p_{\min}$. Consider the combinatorial cascading bandits (Li et al., 2016) satisfying $B_1 = 1$ (see Section 5), our result improves the Li et al. (2016) by a factor of $1/p_{\min}$. Consider the linear reward function (Takemura et al., 2021) without triggering arms (i.e., $p_i^{\boldsymbol{\mu},S} = 1$ for $i \in S$, and 0 otherwise), one can easily verify $B_1 = 1$ and our regret matches the lower bound $\Omega(d\sqrt{KT})$ Takemura et al. (2021) up to logarithmic factors.

Analysis. Here we explain how to prove a regret bound that removes the $1/p_{\min}$ factor under the 1-norm TPM condition. The main challenge is that the mean vector $\boldsymbol{\mu}_t$ and the triggering probability $p_i^{\boldsymbol{\mu}_t,S}$ are dependent on *time-varying* contexts $\boldsymbol{\phi}_t(i)$, so it is impossible to derive any meaningful concentration inequality or regret bound based on $T_{t,i}$, which is the number of times that arm i is triggered, and has been used by the *triggering group (TG)* technique (Wang & Chen, 2017) to remove $1/p_{\min}$. To deal with this problem, we bypass the quantity $T_{t,i}$ and use the *triggering-probability equivalence (TPE)* technique that equalizes $p_i^{\boldsymbol{\mu}_t,S}$ with $\mathbb{E}_t[\mathbb{I}\{i\in\tau_t\}]$, which in turn replaces the expected regret produced by all possible arms with the expected regret produced by $i \in \tau_t$ to avoid $p_{\min}$. To sketch our proof idea, we assume the oracle is deterministic with $\beta = 1$ (the randomness of the oracle and $\beta < 1$ are handled in Appendix A), and let filtration $\mathcal{F}_{t-1}$ be the history data $\mathcal{H}_t$ (defined in Section 2). Denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$, the t-round regret $\mathbb{E}_t[\alpha\cdot r(S_t^*;\boldsymbol{\mu}_t) - r(S_t;\boldsymbol{\mu}_t)] \le \mathbb{E}_t[r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)]$, based on Condition 1, Lemma 1 and definition of S_t. Then

$$\begin{aligned}
&\mathbb{E}_t[r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)] \overset{(a)}{\le} \mathbb{E}_t\left[\textstyle\sum_{i\in\tilde{S}_t} B_1 p_i^{\boldsymbol{\mu}_t,S_t}(\bar{\mu}_{t,i}-\mu_{t,i})\right]\\
&\overset{(b)}{=} \mathbb{E}\left[\textstyle\sum_{i\in\tilde{S}_t} B_1 \mathbb{E}_{\tau_t}[\mathbb{I}\{i\in\tau_t\}](\bar{\mu}_{t,i}-\mu_{t,i}) \mid \mathcal{F}_{t-1}\right]\\
&\overset{(c)}{=} \mathbb{E}_t\left[\textstyle\sum_{i\in\tilde{S}_t} \mathbb{I}\{i\in\tau_t\} B_1(\bar{\mu}_{t,i}-\mu_{t,i})\right]\\
&\overset{(d)}{=} \mathbb{E}_t\left[\textstyle\sum_{i\in\tau_t} B_1(\bar{\mu}_{t,i}-\mu_{t,i})\right], \qquad (3)
\end{aligned}$$

where (a) is by Condition 2, (b) is because $\bar{\mu}_{t,i}, \mu_{t,i}, S_t$ are $\mathcal{F}_{t-1}$ measurable so that the only randomness is from triggering set τ_t and we can substitute $p_i^{\boldsymbol{\mu}_t,S_t}$ with event $\mathbb{I}\{i\in\tau_t\}$ under expectation, (c) is by absorbing the expectation over τ_t to $\mathbb{E}_t$, and (d) is a change of notation. After applying the TPE, we only need to bound the regret produced by $i \in \tau_t$. Hence

$$\begin{aligned}
\mathrm{Reg}(T) &\le \mathbb{E}\left[\textstyle\sum_{t\in[T]}\mathbb{E}_t\left[\sum_{i\in\tau_t} B_1(\bar{\mu}_{t,i}-\mu_{t,i})\right]\right]\\
&\overset{(a)}{\le} \mathbb{E}\left[\textstyle\sum_{t\in[T]}\mathbb{E}_t\left[\sum_{i\in\tau_t} 2B_1\rho(\delta)\,\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}\right]\right]\\
&\overset{(b)}{\le} 2B_1\rho(\delta)\mathbb{E}\left[\sqrt{KT\textstyle\sum_{t\in[T]}\sum_{i\in\tau_t}\|\boldsymbol{\phi}_t(i)\|^2_{\boldsymbol{G}_t^{-1}}}\right]\\
&\overset{(c)}{\le} O(B_1 d\sqrt{KT}\log T). \qquad (4)
\end{aligned}$$

where (a) follows from Lemma 1, (b) is by Cauchy Schwarz inequality over both i and t, and (c) is by the ellipsoidal potential lemma (Lemma 5) in the Appendix.

Remark 3. In addition to the general C²MAB-T setting, the TPE technique can also replace the more involved TG technique (Wang & Chen, 2017) for CMAB-T. Such replacement can save an unnecessary union bound over the group index, which in turn reproduce Theorem 1 of Wang & Chen (2017) under Condition 2, and improve Theorem 1 of Liu et al. (2022) under Condition 4 by a factor of $O(\sqrt{\log T})$, see Appendix C for details.

Algorithm 2 VAC²-UCB: **V**ariance-**A**daptive **C**ontextual **C**ombinatorial **U**pper **C**onfidence **B**ound Algorithm

Input: Base arms $[m]$, dimension d, regularizer γ, failure probability $\delta = 1/T$, offline oracle ORACLE.
Initialize: Gram matrix $\boldsymbol{G}_1 = \gamma\boldsymbol{I}$, regressand $\boldsymbol{b}_1 = \boldsymbol{0}$.
for $t = 1, ..., T$ **do**
  $\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1}\boldsymbol{b}_t$.
  for $i \in [m]$ **do**
    $\bar{\mu}_{t,i} = \langle\boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t\rangle + 2\rho(\delta)\,\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$
    $\underline{\mu}_{t,i} = \langle\boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t\rangle - 2\rho(\delta)\,\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$
    Set the optimistic variance $\bar{V}_{t,i}$ as Equation (6).
  end for
  $S_t = \mathrm{ORACLE}(\bar{\mu}_{t,1}, ..., \bar{\mu}_{t,m})$.
  Play S_t and observe triggering arm set τ_t and observation set $(X_{t,i})_{i\in\tau_t}$.
  $\boldsymbol{G}_{t+1} = \boldsymbol{G}_t + \sum_{i\in\tau_t}\bar{V}_{t,i}^{-1}\boldsymbol{\phi}_t(i)\boldsymbol{\phi}_t(i)^\top$.
  $\boldsymbol{b}_{t+1} = \boldsymbol{b}_t + \sum_{i\in\tau_t}\bar{V}_{t,i}^{-1}\boldsymbol{\phi}_t(i)X_{t,i}$.
end for

4. Variance-Adaptive Algorithm and Analysis for C²MAB-T under VM/TPVM Condition

In this section, we propose a new variance-adaptive algorithm VAC²-UCB (Algorithm 2) to further remove the $O(\sqrt{K})$ factor and achieve $\tilde{O}(B_v d\sqrt{T})$ regret bound for applications satisfying stronger VM/TPVM conditions.

Different from Algorithm 1, VAC^2-UCB leverages the second-order statistics (i.e., variance) to speed up the learning process. To get some intuition, we first assume the variance $V_{s,i} = \text{Var}[X_{s,i}]$ for each base arm i at round s is known in advance. In this case, VAC^2-UCB adopts the *weighted ridge-regression* to learn the parameter $\boldsymbol{\theta}^*$:

$$\hat{\boldsymbol{\theta}}_t = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{s<t}\sum_{i\in\tau_s} (\langle \boldsymbol{\theta}, \boldsymbol{\phi}_s(i)\rangle - X_{s,i})^2/V_{s,i} + \gamma \|\boldsymbol{\theta}\|_2^2, \quad (5)$$

where the first term is "weighted" by the true variance $V_{s,i}$. The closed-form solution of the above estimator is $\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1}\boldsymbol{b}_t$ where the Gram matrix $\boldsymbol{G}_t = \sum_{s<t}\sum_{i\in\tau_s} V_{s,i}^{-1}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top$ and the b-vector $\boldsymbol{b}_t = \sum_{s<t}\sum_{i\in\tau_s} V_{s,i}^{-1}\boldsymbol{\phi}_s(i)X_{s,i}$, which enjoys the similar form (but uses different weights $\bar{V}_{s,i}$) of line 12 and line 13.

The intuition of using the inverse of $V_{s,i}$ to re-weight the observation is that: the smaller the variance, the more accurate the observation $(\boldsymbol{\phi}_t(i), X_{t,i})$ is, and thus more important for the agent to learn unknown $\boldsymbol{\theta}^*$. In fact, the above estimator $\hat{\boldsymbol{\theta}}_t$ is *closely related* to the best linear unbiased estimator (BLUE) (Henderson, 1975). Concretely, in the literature of linear regression, Equation (5) is the lowest variance estimator of $\boldsymbol{\theta}^*$ among all unbiased linear estimators, when the regularization term $\gamma = 0$, $V_{s,i}$ are the true variance proxy of outcomes $(X_{s,i})_{s<t,i\in\tau_s}$ and the context sequence $(\boldsymbol{\phi}_s(i))_{s<t,i\in\tau_s}$ follows the fixed design in Equation (5).

For our C^{2}MAB-T setting, one new challenge arises since the variance $V_{s,i} = \mu_{s,i}(1-\mu_{s,i})$ is not known a priori. Inspired by (Lattimore et al., 2015; Zhou et al., 2021), we construct an optimistic estimation $\bar{V}_{s,i}$ to replace the true variance $V_{s,i}$ in Equation (5). Indeed, we construct $\bar{V}_{t,i}$ by solving the optimal value for the problem $\max_{\mu\in[\underline{\mu}_{t,i},\bar{\mu}_{t,i}]} \mu(1-\mu)$, whose closed form solution immediately follows from the equation below,

$$\bar{V}_{t,i} = \begin{cases} (1-\bar{\mu}_{t,i})\bar{\mu}_{t,i}, & \text{if } \bar{\mu}_{t,i} \le \frac{1}{2} \\ (1-\underline{\mu}_{t,i})\underline{\mu}_{t,i}, & \text{if } \underline{\mu}_{t,i} \ge \frac{1}{2} \\ \frac{1}{4}, & \text{otherwise} \end{cases} \quad (6)$$

where $\bar{\mu}_{t,i}$ and $\underline{\mu}_{t,i}$ are UCB and LCB values to be introduced later. Notice that with high probability the true $\mu_{t,i}$ lies within LCB and UCB values and as they are getting more accurate, the optimistic variance $V_{t,i}$ is also approaching the true variance $V_{t,i}$.

To guarantee $\hat{\boldsymbol{\theta}}_t$ is a good estimator, we prove a new lemma (similar to Proposition 1) to guarantee the concentration bound of $\boldsymbol{\theta}_t$ but in face of the unknown variance. Note that the sentinel work Lattimore et al. (2015) proves a similar concentration bound, the difference is that we have multiple arms triggered in each round instead of a single arm. To address this, we replaced the original concentration bound with the new one below that has an extra K^4 factor in N, which finally results in $\log K$ factor in the confidence radius $\rho(\delta)$.

Lemma 2. *Let $\gamma > 0, N = (4d^2K^4T^4)^d$ so that $\rho(\delta) = \left(1+\sqrt{\gamma}+4\sqrt{\log\left(\frac{6TN}{\delta}\log\left(\frac{3TN}{\delta}\right)\right)}\right)$. We have for all $t \le T$, with probability at least $1-\delta$, $\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t}^2 \le \rho(\delta)$.*

Proof. See Appendix B.1. ■

Building on this lemma, we construct $\bar{\mu}_{t,i}$ as an upper bound of $\boldsymbol{\mu}_{t,i}$ in line 6, and $\underline{\mu}_{t,i}$ as a lower bound of $\boldsymbol{\mu}_{t,i}$ in line 7, based on our variance-adaptive $\hat{\boldsymbol{\theta}}_t$, $\boldsymbol{G}_t$. Note that the doubling of the radius $2\rho(\delta)$ instead of using $\rho(\delta)$ in Lemma 2 is purely for the correctness of our technical analysis. As a convention, we clip $\bar{\mu}_{t,i}, \underline{\mu}_{t,i}$ into $[0,1]$ if they are above 1 or below 0.

Lemma 3. *With probability at least $1-\delta$, we have $\mu_{t,i} \le \bar{\mu}_{t,i} \le \mu_{t,i} + 3\rho(\delta)\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$, and $\mu_{t,i} \ge \underline{\mu}_{t,i} \ge \mu_{t,i} - 3\rho(\delta)\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$ for all $i \in [m]$.*

Proof. This lemma follows from the similar derivation of Lemma 1, where we have different definitions of $\bar{\mu}_{t,i}, \underline{\mu}_{t,i}$ and the concentration now relies on Lemma 2. ■

After the agent plays S_t, the base arms in τ_t are triggered, and the agent receives observation set $(X_{t,i})_{i\in\tau_t}$ as feedback. These observations (reweighted by optimistic variance $\bar{V}_{t,i}$) are then used to update $\boldsymbol{G}_t$ and $\boldsymbol{b}_t$ for future rounds.

4.1. Results and Analysis under VM condition

We first show a regret bound for VAC^2-UCB that is independent of batch size K when the VM condition holds.

Theorem 2. *For a C^{2}MAB-T instance that satisfies monotonicity (Condition 1) and VM smoothness (Condition 3) with coefficient (B_v, B_1), VAC^2-UCB (Algorithm 2) with an (α,β)-approximation oracle achieves an (α,β)-approximate regret bounded by $O\left(\frac{B_v}{\sqrt{p_{\min}}}(\sqrt{d\log(KT/\gamma)}+\sqrt{\gamma})\sqrt{Td\log(KT/\gamma)}\right)$.*

Discussion. Looking at Theorem 2, we achieve an $O(B_v d\sqrt{T}\log T/\sqrt{p_{\min}})$ regret bound when $d \ll K \le m \ll T$. For combinatorial cascading bandits (Li et al., 2016) with $B_v = 1$, our regret is independent of m, K and improves Li et al. (2016) by a factor $O(\sqrt{K/p_{\min}})$.

In addition to the general C^{2}MAB-T setting, one can verify that for non-triggering C^{2}MAB, $p_{\min} = 1$, and we obtain the batch-size independent regret bound $O(B_v d\sqrt{T}\log T)$. Recall $B_v = O(B_1\sqrt{K})$ for any C^{2}MAB-T instances, so our regret bound reproduces $O(B_1 d\sqrt{KT}\log T)$, and thus matches the similar lower bound (Takemura et al., 2021) for

Table 2. Summary of the coefficients, regret bounds and improvements for various applications.

Application	Condition	(B_v, B_1, λ)	Regret	Improvement
Online Influence Maximization (Wen et al., 2017)	TPM	$(-, \lvert V\rvert, -)$	$O(d\lvert V\rvert\sqrt{\lvert E\rvert T\log T})$	$\tilde{O}(\sqrt{\lvert E\rvert})$
Disjunctive Combinatorial Cascading Bandits (Li et al., 2016)	TPVM	$(1, 1, 2)$	$O(d\sqrt{T}\log T)$	$\tilde{O}(\sqrt{K}/p_{\min})$
Conjunctive Combinatorial Cascading Bandits (Li et al., 2016)	TPVM	$(1, 1, 1)$	$O(d\sqrt{T}\log T)$	$\tilde{O}(\sqrt{K}/r_{\max})$
Linear Cascading Bandits (Vial et al., 2022)*	TPVM	$(1, 1, 2)$	$O(d\sqrt{T}\log T)$	$\tilde{O}(\sqrt{K/d})$
Multi-layered Network Exploration (Liu et al., 2021b)	TPVM	$(\sqrt{1.25\lvert V\rvert}, 1, 2)$	$O(d\sqrt{\lvert V\rvert T}\log T)$	$\tilde{O}(\sqrt{n}/p_{\min})$
Probabilistic Maximum Coverage (Chen et al., 2013)**	VM	$(3\sqrt{2\lvert V\rvert}, 1, -)$	$O(d\sqrt{\lvert V\rvert T}\log T)$	$\tilde{O}(\sqrt{k})$

$|V|, |E|, n, k, L$ denotes the number of target nodes, the number of edges that can be triggered by the set of seed nodes, the number of layers, the number of seed nodes and the length of the longest directed path, respectively; K is the length of the ordered list, $r_{\max} = \alpha \cdot \max_{t\in[T], S\in\mathcal{S}} r(S; \boldsymbol{\mu}_t)$;
* A special case of disjunctive combinatorial cascading bandits. ** This row is for C^{2}MAB application and the rest of rows are for C^{2}MAB-T applications.

the linear reward functions. For the more interesting non-linear reward function with $B_v = o(B_1\sqrt{K})$, our regret improves non-variance-adaptive algorithm C^{2}UCB, whose regret is $O(B_1 d\sqrt{KT}\log T)$ (Qin et al., 2014; Takemura et al., 2021).

Analysis. At a high level, the improvement of $\sqrt{K}$ comes from the VM condition and the optimistic variance, which together save the use of Cauchy-Schwarz inequality that generates a $O(\sqrt{K})$ factor in the step (b) of Equation (4). In order to leverage the variance information, we decompose the regret into term (I) and (II),

$$\begin{aligned}\mathrm{Reg}(T) &\le \mathbb{E}\left[\textstyle\sum_{t=1}^T r(S_t; \bar{\boldsymbol{\mu}}_t) - r(S_t; \boldsymbol{\mu}_t)\right] \\ &\le \mathbb{E}[\textstyle\sum_{t=1}^T \underbrace{|r(S_t; \bar{\boldsymbol{\mu}}_t) - r(S_t; \tilde{\boldsymbol{\mu}}_t)|}_{\text{(I)}} + \underbrace{|r(S_t; \boldsymbol{\mu}_t) - r(S_t; \tilde{\boldsymbol{\mu}}_t)|}_{\text{(II)}}],\end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{\mu}}_t$ is the vector whose i-th entry is the maximizer that achieves optimistic variance $\bar{V}_{t,i}$, i.e., $\tilde{\mu}_{t,i} = \mathrm{argmax}_{\mu\in[\underline{\mu}_{t,i}, \bar{\mu}_{t,i}]}\, \mu(1-\mu)$. Now we show a sketched proof to bound the term (I) and one can bound the term (II) similarly.

$$\begin{aligned}&\mathbb{E}\left[\textstyle\sum_{t\in[T]} \text{(I)}\right] \overset{(a)}{\le} B_v \mathbb{E}\left[\textstyle\sum_{t=1}^T \sqrt{\sum_{i\in\tilde{S}_t} (\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2/\bar{V}_{t,i}}\right] \\ &\overset{(b)}{\le} \tfrac{B_v}{\sqrt{p_{\min}}} \cdot \mathbb{E}\left[\textstyle\sum_{t=1}^T \sqrt{\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t, S_t} (\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2/\bar{V}_{t,i}}\right] \\ &\overset{(c)}{\le} \tfrac{B_v}{\sqrt{p_{\min}}} \cdot \sqrt{T\mathbb{E}\left[\textstyle\sum_{t=1}^T \sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t, S_t} (\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2/\bar{V}_{t,i}\right]} \\ &\overset{(d)}{\le} \tfrac{B_v}{\sqrt{p_{\min}}} \cdot \sqrt{T\mathbb{E}\left[\textstyle\sum_{t=1}^T \sum_{i\in\tau_t} (6\rho(\delta)\, \|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}})^2/\bar{V}_{t,i}\right]} \\ &\overset{(e)}{\le} O(B_v d\sqrt{T\log T}/\sqrt{p_{\min}}),\end{aligned} \tag{8}$$

where (a) is by Condition 3, (b) is by the definition of $p_i^{\boldsymbol{\mu}_t, S_t} \ge p_{\min}$ for $i \in \tilde{S}_t$, (c) is by Cauchy–Schwarz over t and the Jensen's inequality, (d) follows from the TPE and Lemma 3, (e) follows from Lemma 6.

4.2. Results and Analysis under TPVM Condition

Next, we show that VAC^2-UCB can achieve regret bounds that remove the $O(\sqrt{K})$ and $O(1/\sqrt{p_{\min}})$ factor for applications satisfying the stronger TPVM conditions.

We first introduce a mild condition over the triggering probability (which is similar to Condition 2) to give our regret bounds and analysis.

Condition 5 (1-norm TPM Bounded Smoothness for Triggering Probability). *We say that a C^{2}MAB-T problem instance satisfies the triggering probability modulated B_p-bounded smoothness condition over the triggering probability, if for any action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [0,1]^m$, and any arm $i \in [m]$, we have $|p_i^{\boldsymbol{\mu}', S} - p_i^{\boldsymbol{\mu}, S}| \le B_p \sum_{j\in[m]} p_j^{\boldsymbol{\mu}, S} |\mu_j - \mu'_j|$.*

Now we state our main theorem as follows.

Theorem 3. *For a C^{2}MAB-T instance, when its reward function satisfies monotonicity (Condition 1) and TPVM smoothness (Condition 4) with coefficient (B_v, B_1, λ), and its triggering probability $p_i^{\boldsymbol{\mu}, S}$ satisfies 1-norm TPM smoothness with coefficient B_p (Condition 5), if $\lambda \ge 2$, then VAC^2-UCB (Algorithm 2) with an (α, β)-approximation oracle achieves an (α, β)-approximate regret bounded by*

$$O\left(B_v d\sqrt{T}\log T + B_v B_p d\sqrt{K}\log T/\sqrt{p_{\min}}\right), \tag{9}$$

and if $\lambda \ge 1$, then VAC^2-UCB (Algorithm 2) achieves an (α, β)-approximate regret bounded by

$$O\left(B_v d\sqrt{T}\log T + B_v\sqrt{B_p}(KT)^{1/4}(d\log T)^{3/4}/\sqrt{p_{\min}}\right). \tag{10}$$

Discussion. The leading term of Theorem 3 is $O(B_v d\sqrt{T}\log T)$ when $d \ll K \le m \ll T$, which removes the $1/\sqrt{p_{\min}}$ factor compared with Theorem 2. Also, notice that Theorem 3 relies on an additional B_p-smoothness condition over the triggering probability. However, we claim that this condition is mild and almost always satisfies with $B_p = B_1$ for applications considered in this paper (see Appendix D).

Analysis. We use the regret decomposition of Equation (7) to the same term (I) and (II), and leverage on TPVM condi-

tion (Condition 4) to obtain:

$$\mathbb{E}\left[\sum_{t\in[T]}(\mathrm{I})\right] \overset{(a)}{\leq} B_v\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}(p_i^{\tilde{\boldsymbol{\mu}}_t,S_t})^{\lambda}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2/\bar{V}_{t,i}}\right]. \tag{11}$$

However, we cannot use TPE as Equation (8) because $p_i^{\tilde{\boldsymbol{\mu}}_t,S_t} \neq p_i^{\boldsymbol{\mu}_t,S_t}$ in general. To handle this mismatch, we use the fact that triggering probability usually satisfies a smoothness condition in Condition 5, and prove that the mismatch only affect the lower-order term as follows:

By Condition 5, $(p_i^{\tilde{\boldsymbol{\mu}}_t,S_t})^{\lambda}$ is upper bounded by $(p_i^{\boldsymbol{\mu}_t,S_t} + \min\{1, \sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}|\mu_{t,j}-\tilde{\mu}_{t,j}|\})^2$ when $\lambda \geq 2$, and the regret is bounded by the terms as shown below:

$$\text{Eq. (11)} \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} B_v\sqrt{\sum_{i\in\tilde{S}_t} 3p_i^{\boldsymbol{\mu}_t,S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2/\bar{V}_{t,i}}\right]}_{\text{leading term}}$$
$$+\underbrace{\frac{B_vB_p\sqrt{K}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t,S_t}(\bar{\mu}_{t,i}-\underline{\mu}_{t,i})^2/\bar{V}_{t,i}\right]}_{\text{lower-order term}},$$

where the leading term is of order $O(B_v\sqrt{T}\log T)$ by using the same derivation of step (c)-(e) in Equation (8), and the lower order term is bounded by $O(B_vB_p\sqrt{K/p_{\min}}\log T)$ by TPE and the weighted ellipsoidal potential lemma (Lemma 6). For $\lambda \geq 1$, the lower-order term becomes $\frac{B_v\sqrt{B_p}K^{1/4}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu}_t,S_t}(\bar{\mu}_{t,i}-\underline{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right)^{3/4}\right]$, which results in a larger lower-order regret term. See Appendix B.3 for details.

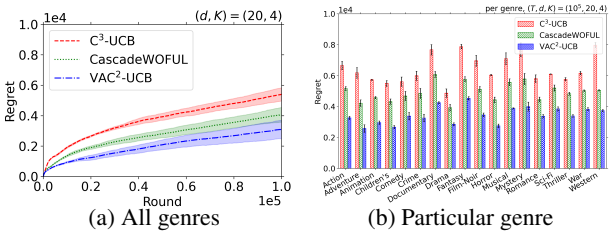
5. Applications and Experiments

We now move to applications and experimental results. We first show how our theoretical results improve various C^{2}MAB and C^{2}MAB-T applications under 1-norm TPM, TPVM and VM smoothness conditions with their corresponding B_1, B_v, λ coefficients. Then, we provide an empirical comparison in the context of the contextual cascading bandit application.

The instantiation of our theoretical results in the context of a variety of specific C^{2}MAB and C^{2}MAB-T applications is shown in Table 2. The final column of the table details the improvement in regret that our results yield in each case. For detailed settings, proofs, and the discussion of the application results, see Appendix D.

Our experimental results are summarized in Figure 1, which details experiments on the MovieLens-1M dataset[‡]. Experiments on other data are included in the Appendix. Figure 1 illustrates that our VAC^2-UCB algorithm outperforms C^3-UCB (Li et al., 2016), the variance-agnostic cascading bandit algorithm, and CascadeWOFUL (Vial et al., 2022), the state-of-the-art variance-aware cascading bandit algorithm, eventually incurring 45% and 25% less regret. For detailed settings, comparisons, and discussions, see Appendix E.

[‡]grouplens.org/datasets/movielens/1m/



(a) All genres (b) Particular genre

Figure 1. Regret results for MovieLens data.

6. Conclusion

This paper studies contextual combinatorial bandits with probabilistically triggered arms (C^{2}MAB-T) under a variety of smoothness conditions. Under the triggering probability modulated (TPM) condition, we design the C^2-UCB-T algorithm and propose a novel analysis to achieve an $\tilde{O}(d\sqrt{KT})$ regret bound, removing a potentially exponentially large factor $O(1/p_{\min})$. Under the variance modulated conditions (VM or TPVM), we propose a new variance-adaptive algorithm VAC^2-UCB and derive a regret bound $\tilde{O}(d\sqrt{T})$, which removes the batch-size K dependence. As valuable by-product, we find our TPE analysis technique and variance-adaptive algorithm can be applied to the CMAB-T and C^{2}MAB setting, improving existing results as well. Experiments show that our algorithm can achieve at least 13% and 25% improvement compared with benchmark algorithms on synthetic and real-world datasets, respectively. For the future study, it would be interesting to extend our application scenarios. One could also relax the perfectly linear assumption by introducing model mis-specifications or corruptions.

Acknowledgement

The work of John C.S. Lui was supported in part by RGC's GRF 14215722. The work of Mohammad Hajiesmaili is supported by NSF CAREER-2045641, CPS-2136199, CNS-2106299, and CNS-2102963. Wierman is supported by NSF grants CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Bernstein, S. *The Theory of Probabilities (Russian)*. Moscow, 1946.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159. PMLR, 2013.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016a.

Chen, X., Li, Y., Wang, P., and Lui, J. A general framework for estimating graphlet statistics via random walk. *Proceedings of the VLDB Endowment*, 10(3):253–264, 2016b.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28, 2015.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5): 1466–1478, 2012.

Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pp. 423–447, 1975.

Kempe, D., Kleinberg, J., and Tardos, É. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1450–1458, 2015a.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015b.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Crammer, K., and Szepesvári, C. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28, 2015.

Li, S., Wang, B., Zhang, S., and Chen, W. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pp. 1245–1253. PMLR, 2016.

Li, S., Kong, F., Tang, K., Li, Q., and Chen, W. Online influence maximization under linear threshold model. *Advances in Neural Information Processing Systems*, 33: 1192–1204, 2020.

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021a.

Liu, X., Zuo, J., Chen, X., Chen, W., and Lui, J. C. Multi-layered network exploration via random walks: From offline optimization to online learning. In *International Conference on Machine Learning*, pp. 7057–7066. PMLR, 2021b.

Liu, X., Zuo, J., Wang, S., Joe-Wong, C., Lui, J., and Chen, W. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. In *Advances in Neural Information Processing Systems*, 2022.

Merlis, N. and Mannor, S. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. In *Conference on Learning Theory*, pp. 2465–2489. PMLR, 2019.

Qin, L., Chen, S., and Zhu, X. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469. SIAM, 2014.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9791–9798, 2021.

Vial, D., Shakkottai, S., and Srikant, R. Minimax regret for cascading bandits. In *Advances in Neural Information Processing Systems*, 2022.

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

Wen, Z., Kveton, B., Valko, M., and Vaswani, S. Online influence maximization under independent cascade model with semi-bandit feedback. *Advances in neural information processing systems*, 30, 2017.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

Zong, S., Ni, H., Sung, K., Ke, N. R., Wen, Z., and Kveton, B. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.

Zuo, J., Liu, X., Joe-Wong, C., Lui, J. C., and Chen, W. Online competitive influence maximization. In *International Conference on Artificial Intelligence and Statistics*, pp. 11472–11502. PMLR, 2022.

Appendix

The Appendix is organized as follows. Appendix A gives the detailed proofs for theorems and lemmas in Section 3. Appendix B provides the detailed proofs for theorems and lemmas in Section 4. Appendix C shows how the triggering probability equivalence technique can be applied to non-contextual CMAB-T to obtain improved results. Appendix D gives the detailed settings, results and comparisons included in Table 2. Appendix E provides detailed experimental setups and additional results. Appendix F summarizes the concentration bounds, facts, and technical lemmas used in this paper.

A. Proofs for C^{2}MAB-T under the TPM Condition (Section 3)

A.1. Proof of Theorem 1

We first give/recall some definitions and events. Recall that in Algorithm 1, the gram matrix, the b-vector and the estimator are

$$\boldsymbol{G}_t = \gamma \boldsymbol{I} + \sum_{s<t} \sum_{i \in \tau_s} \boldsymbol{\phi}_s(i) \boldsymbol{\phi}_s(i)^\top \tag{12}$$

$$\boldsymbol{b}_t = \sum_{s<t} \sum_{i \in \tau_s} \boldsymbol{\phi}_s(i) X_{s,i} \tag{13}$$

$$\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1} \boldsymbol{b}_t. \tag{14}$$

Let us use $\mathcal{W}_t$ to denote the nice event when the oracle can output solution S with $r(S; \boldsymbol{\mu}) \geq \alpha \cdot r(S^*; \boldsymbol{\mu})$ where $S^* = \operatorname{argmax}_{S \in \mathcal{S}} r(S; \boldsymbol{\mu})$ for any $\boldsymbol{\mu}$ at round t. We use $\mathcal{N}_t$ to denote the nice event when the $\left\| \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^* \right\|_{\boldsymbol{G}_t} \leq \rho(\delta)$ holds for any $t \in [T]$. Define the filtration to be $\mathcal{F}_{t-1} = (S_1, \boldsymbol{\phi}_1, \tau_1, (X_{1,i})_{t \in \tau_1}, ..., S_{t-1}, \boldsymbol{\phi}_{t-1}, \tau_{t-1}, (X_{t-1,i})_{t \in \tau_{t-1}}, S_t, \boldsymbol{\phi}_t)$ that takes both history data $\mathcal{H}_t$ and action S_t to handle the randomness of the oracle, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Now we bound the regret under nice event $\mathcal{W}_t$ and $\mathcal{N}_t$,

$$\begin{aligned}
\mathrm{Reg}(T) &\overset{(a)}{=} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t[\alpha \cdot r(S_t^*; \boldsymbol{\mu}_t) - r(S_t; \boldsymbol{\mu}_t)]\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t[\alpha \cdot r(S_t^*; \bar{\boldsymbol{\mu}}_t) - r(S_t; \boldsymbol{\mu}_t)]\right] \overset{(c)}{\leq} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t[r(S_t; \bar{\boldsymbol{\mu}}_t) - r(S_t; \boldsymbol{\mu}_t)]\right] \\
&\overset{(d)}{\leq} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t\left[\sum_{i \in \tilde{S}_t} B_1 p_i^{\boldsymbol{\mu}_t, S_t} (\bar{\mu}_{t,i} - \mu_{t,i})\right]\right] \\
&\overset{(e)}{=} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t\left[\sum_{i \in \tau_t} B_1 (\bar{\mu}_{t,i} - \mu_{t,i})\right]\right] \\
&\overset{(f)}{\leq} \mathbb{E}\left[\sum_{t \in [T]} \mathbb{E}_t\left[\sum_{i \in \tau_t} 2B_1 \rho(\delta) \left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}\right]\right] \overset{(g)}{=} 2B_1 \rho(\delta) \mathbb{E}\left[\sum_{t \in [T]} \sum_{i \in \tau_t} \left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}\right] \\
&\overset{(h)}{\leq} 2B_1 \rho(\delta) \mathbb{E}\left[\sqrt{KT \sum_{t \in [T]} \sum_{i \in \tau_t} \left\|\boldsymbol{\phi}_t(i)\right\|^2_{\boldsymbol{G}_t^{-1}}}\right] \\
&\overset{(i)}{\leq} O\left(B_1 (\sqrt{2d \log T} + \sqrt{\gamma}) \sqrt{2KT \log T}\right) \leq O(B_1 d \sqrt{KT} \log T).
\end{aligned} \tag{15}$$

where (a) follows from the regret definition and the tower rule, (b) is by Condition 1 and Lemma 1 saying that $\mu_{t,i} \leq \bar{\mu}_{t,i}$, (c) is by nice event $\mathcal{W}_t$ and the definition of S_t, (d) is by Condition 2 (e) follows from by the TPE trick Lemma 4, (f) by Lemma 1, (g) by tower rule, (h) by Cauchy Schwarz inequality, and (i) by the ellipsoidal potential lemma (Lemma 5). Similar to (Wang & Chen, 2017) The theorem is concluded by the definition of the (α, β)-approximate regret, and considering event $\neg \mathcal{W}_t$ or $\neg \mathcal{N}_t$, which contributes to at most $(1-\beta) T \Delta_{\max} + \delta T \Delta_{\max}$ regret.

A.2. Important Lemmas used for proving Theorem 1

Lemma 1. *With probability at least* $1-\delta$*, we have* $\mu_{t,i} \le \bar{\mu}_{t,i} \le \mu_{t,i} + 2\rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}$ *for all* $i \in [m], t \in [T]$.

Proof. For any $i \in [m], t \in [T]$, we have

$$\begin{aligned}
&\left|\langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t(i)\rangle - \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i)\rangle\right| \\
=&\left|\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i)\rangle\right| \\
\overset{(a)}{\le}& \left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t} \cdot \left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}} \\
\overset{(b)}{\le}& \rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}},
\end{aligned}$$

where (a) by Cauchy-Schwartz, (b) by Proposition 1. Now use the definition of $\mu_{t,i} = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i)\rangle$ and $\bar{\mu}_{t,i} = \langle \hat{\boldsymbol{\theta}}_t, \boldsymbol{\phi}_t(i)\rangle + \rho(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{G}_t^{-1}}$. ■

Lemma 4 (Triggering Probability Equivalence (TPE))**.** $\mathbb{E}_t\left[\sum_{i\in\tilde{S}_t} B_1 p_i^{\boldsymbol{\mu}_t, S_t}(\bar{\mu}_{t,i} - \mu_{t,i})\right] = \mathbb{E}_t\left[\sum_{i\in\tau_t} B_1(\bar{\mu}_{t,i} - \mu_{t,i})\right]$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{i\in\tilde{S}_t} B_1 p_i^{\boldsymbol{\mu}_t, S_t}(\bar{\mu}_{t,i} - \mu_{t,i})\right] \\
\overset{(a)}{=}& \mathbb{E}\left[\sum_{i\in\tilde{S}_t} B_1 \mathbb{E}_{\tau_t}[\mathbb{I}\{i\in\tau_t\}](\bar{\mu}_{t,i} - \mu_{t,i}) \mid \mathcal{F}_{t-1}\right] \\
\overset{(b)}{=}& \mathbb{E}_t\left[\sum_{i\in\tilde{S}_t} \mathbb{I}\{i\in\tau_t\} B_1(\bar{\mu}_{t,i} - \mu_{t,i})\right] \\
\overset{(c)}{=}& \mathbb{E}_t\left[\sum_{i\in\tau_t} B_1(\bar{\mu}_{t,i} - \mu_{t,i})\right],
\end{aligned} \tag{16}$$

(a) is because $\bar{\mu}_{t,i}, \mu_{t,i}, S_t$ are $\mathcal{F}_{t-1}$-measurable so that the only randomness is from triggering set τ_t and we can substitute $p_i^{\boldsymbol{\mu}_t, S_t}$ with event $\mathbb{I}\{i\in\tau_t\}$ under expectation, (b) is by absorbing the expectation over τ_t to $\mathbb{E}_t$, and (c) is a simple change of notation. Actually, TPE can be applied whenever the quantities (other than $p_i^{D,S}$) are $\mathcal{F}_{t-1}$-measurable, which would be helpful for later sections. ■

Lemma 5 (Ellipsoidal Potential Lemma)**.** $\sum_{t=1}^{T}\sum_{i\in\tau_t}\left\|\boldsymbol{\phi}_t(i)\right\|^2_{\boldsymbol{G}_t^{-1}} \le 2d\log(1 + KT/(\gamma d)) \le 2d\log T$ *when* $\gamma \ge K$.

Proof.

$$\begin{aligned}
\det(\boldsymbol{G}_{t+1}) \overset{(a)}{=}& \det\left(\boldsymbol{G}_t + \sum_{i\in\tau_t}\boldsymbol{\phi}_t(i)\boldsymbol{\phi}_t(i)^\top\right) \\
\overset{(b)}{=}& \det(\boldsymbol{G}_t)\cdot\det\left(\boldsymbol{I} + \sum_{i\in\tau_t} G_t^{-1/2}\boldsymbol{\phi}_t(i)(G_t^{-1/2}\boldsymbol{\phi}_t(i))^\top\right) \\
\overset{(c)}{\ge}& \det(\boldsymbol{G}_t)\cdot\left(1 + \sum_{i\in\tau_t}\left\|\boldsymbol{\phi}_t(i)\right\|^2_{\boldsymbol{G}_t^{-1}}\right) \\
\overset{(d)}{\ge}& \det(\gamma\boldsymbol{I})\prod_{s=1}^{t}\left(1 + \sum_{i\in\tau_s}\left\|\boldsymbol{\phi}_s(i)\right\|^2_{\boldsymbol{G}_s^{-1}}\right),
\end{aligned} \tag{17}$$

where (a) follows from the definition, (b) follows from $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$ and $\boldsymbol{A}+\boldsymbol{B} = \boldsymbol{A}^{1/2}(\boldsymbol{I}+\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2})$, (c) follows from Lemma 14, (d) follows from repeatedly applying (c).

Since $\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}} \leq \frac{\|\boldsymbol{\phi}_s(i)\|^2}{\lambda_{\min}(\boldsymbol{G}_s)} \leq 1/\gamma \leq 1/K$, we have $\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}} \leq 1$. Using the fact that $2\log(1+x) \geq x$ for any $[0,1]$, we have

$$
\begin{aligned}
&\sum_{s\in t}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}} \\
\leq& 2\sum_{s=1}^{t}\log\left(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}\right) \\
=& 2\log\prod_{s=1}^{t}\left(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}\right) \\
\overset{(a)}{\leq}& 2\log\left(\frac{\det(\boldsymbol{G}_{t+1})}{\det(\gamma\boldsymbol{I})}\right) \\
\overset{(b)}{\leq}& 2\log\left(\frac{(\gamma+KT/d)^d}{\gamma^d}\right) = 2d\log(1+KT/(\gamma d)) \leq 2d\log(T),
\end{aligned}
$$

where the (a) follows from Equation (17), (b) follows from Lemma 15. ■

B. Proofs for C^{2}MAB-T under the VM or TPVM Condition (Section 4)

B.1. Proof of Lemma 2

Our analysis is inspired by the derivation of Theorem 3 by (Lattimore et al., 2015) to bound the key ellipsoidal radius $\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{G}_t} \leq \rho$ for the C^{2}MAB-T setting, where multiple arms can be triggered in each round. Before we going into the main proof, we first introduce some notations and events as follows.

Recall that for $t \geq 1$, $X_{t,i}$ is a Bernoulli random variable with mean $\mu_{t,i} = \langle\theta^*, \boldsymbol{\phi}_t(i)\rangle$, suppose $\|\boldsymbol{\theta}^*\|_2 \leq 1, \|\boldsymbol{\phi}_t(i)\| \leq 1$, we can represent $X_{t,i}$ by $X_{t,i} = \mu_{t,i} + \eta_{t,i}$, where noise $\eta_{t,i} \in [-1,1]$, its mean $\mathbb{E}[\eta_{t,i} \mid \mathcal{F}_{t-1}] = 0$, and its variance $\mathrm{Var}[\eta_{t,i} \mid \mathcal{F}_{t-1}] = \mu_{t,i}(1-\mu_{t,i})$. Also note that in Algorithm 2, the gram matrices, the b-vector and the weighted least-square estimator are the following.

$$
\boldsymbol{G}_t = \gamma\cdot\boldsymbol{I} + \sum_{s=1}^{t-1}\sum_{i\in\tau_s}\bar{V}_{s,i}^{-1}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top, \tag{18}
$$

$$
\boldsymbol{b}_t = \sum_{s=1}^{t-1}\sum_{i\in\tau_s}\bar{V}_{s,i}^{-1}\boldsymbol{\phi}_s(i)X_{s,i}, \tag{19}
$$

$$
\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1}\boldsymbol{b}_t, \tag{20}
$$

where we set $G_0 = \gamma\boldsymbol{I}$, and optimistic variances $\bar{V}_{s,i}$ are defined as in Equation (6) of Algorithm 2.

Let us define $\boldsymbol{Z}_t = \sum_{s<t}\sum_{i\in\tau_s}\eta_{s,i}\boldsymbol{\phi}_t(i)/\bar{V}_{s,i}$, and the key of this proof is to bound $\boldsymbol{Z}_t$ (this quantity is often denoted as S_t in the self-normalized bound (Abbasi-Yadkori et al., 2011), but S_t is occupied to denote actions at round t in this work).

We finally define failure events $F_0 \subseteq F_1 \subseteq ... \subseteq F_T$ be a sequence of events defined by

$$
F_t = \{\exists s \leq t \text{ such that } \|\boldsymbol{Z}_s\|_{\boldsymbol{G}_s} + \sqrt{\gamma} \geq \rho\}. \tag{21}
$$

These failure events are crucial in the sense that $\boldsymbol{\theta}^*$ lies in the confidence ellipsoid $\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{G}_t} \leq \rho$ (see Lemma 8 for its proof).

Next, we can prove by induction that the probability of $\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t} + \sqrt{\gamma} \geq \rho$ given $\neg F_{t-1}$ is very small, for $t = 1, ..., T$ (see its proof in Lemma 7). Based on this, we can have $\Pr[\neg F_T] = 1 - \Pr[F_0] - \sum_{t=1}^{T}\Pr\left[\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t} + \sqrt{\gamma} \geq \rho \text{ and } \neg F_{t-1}\right] \geq 1-\delta$ (as $\neg F_0$ always holds), and thus by Equation (148), Lemma 2 is proved as desired.

B.2. Proof of Theorem 2 under VM condition

Similar to Appendix A, we first give/recall some definitions and events. Recall that in Algorithm 2, the gram matrices, the b-vector, and the weighted least-square estimator are defined in Equation (18) The optimistic variances $\bar{V}_{s,i}$ are defined as in Equation (6) of Algorithm 2. Let us use $\mathcal{W}_t$ to denote the nice event when the oracle can output solution S with $r(S;\boldsymbol{\mu}) \geq \alpha \cdot r(S^*;\boldsymbol{\mu})$ where $S^* = \text{argmax}_{S\in\mathcal{S}} r(S;\boldsymbol{\mu})$ for any $\boldsymbol{\mu}$ at round t. We use $\mathcal{N}_t$ to denote the nice event when the $\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t} \leq \rho(\delta)$ holds for any $t \in [T]$ (which can be implied by $\neg F_T$). Define the filtration to be $\mathcal{F}_{t-1} = (S_1, \boldsymbol{\phi}_1, \tau_1, (X_{1,i})_{t\in\tau_1}, ..., S_{t-1}, \boldsymbol{\phi}_{t-1}, \tau_{t-1}, (X_{t-1,i})_{t\in\tau_{t-1}}, S_t, \boldsymbol{\phi}_t)$ that takes both history data $\mathcal{H}_t$ and action S_t to handle the randomness of the oracle, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$.

Let $\tilde{\boldsymbol{\mu}}_t$ be the vector whose i-th entry is the maximizer that achieves $\bar{V}_{t,i}$, i.e., $\tilde{\mu}_{t,i} = \text{argmax}_{\mu\in[\underline{\mu}_{t,i},\bar{\mu}_{t,i}]} \mu(1-\mu)$. Now we bound the regret under nice event $\mathcal{W}_t$ and $\mathcal{N}$ (where $\mathcal{N}_t$ can be implied from $\neg F_T$ by derivation in Lemma 8),

$$\text{Reg}(T) \overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T} \alpha r(S_t^*;\boldsymbol{\mu}_t) - r(S_t;\boldsymbol{\mu}_t)\right] \tag{22}$$

$$\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} \alpha r(S_t^*;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)\right] \tag{23}$$

$$\overset{(c)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)\right] \tag{24}$$

$$\overset{(d)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} \underbrace{|r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\tilde{\boldsymbol{\mu}}_t)|}_{\text{(I)}} + \underbrace{|r(S_t;\boldsymbol{\mu}_t) - r(S_t;\tilde{\boldsymbol{\mu}}_t)|}_{\text{(II)}}\right], \tag{25}$$

where (a) is by definition, (b) follows from Condition 1 and Lemma 3, (c) from event $\mathcal{W}$ and the definition of S_t, (d) from triangle inequality.

Now we show how to bound term (I),

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in[T]} \text{(I)}\right] &\overset{(a)}{\leq} B_v \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\sum_{i\in\tilde{S}_t} \frac{(\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \\
&\overset{(b)}{\leq} \frac{B_v}{\sqrt{p_{\min}}} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t,S_t} \frac{(\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \\
&\overset{(c)}{\leq} \frac{B_v}{\sqrt{p_{\min}}} \cdot \mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t,S_t} \frac{(\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \\
&\overset{(d)}{\leq} \frac{B_v}{\sqrt{p_{\min}}} \cdot \sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t,S_t} \frac{(\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right]} \\
&\overset{(e)}{=} \frac{B_v}{\sqrt{p_{\min}}} \cdot \sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tau_t} \frac{(\bar{\mu}_{t,i} - \tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right]} \\
&\overset{(f)}{\leq} \frac{B_v}{\sqrt{p_{\min}}} \cdot \sqrt{T\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tau_t} \frac{(6\rho(\delta)\,\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}})^2}{\bar{V}_{t,i}}\right]} \\
&\overset{(g)}{\leq} O(B_v d\sqrt{T}\log(KT)/\sqrt{p_{\min}}),
\end{aligned} \tag{26}$$

where (a) follows from Condition 3, (b) follows from the definition of $p_{\min}$ s.t. $p_i^{\boldsymbol{\mu}_t,S_t} \geq p_{\min}$ for $i \in \tilde{S}_t$, (c) follows from

Cauchy–Schwarz, (d) follows from Jensen's inequality, (e) follows from the TPE trick, (f) follows from Lemma 3, (g) follows from Lemma 6.

Now for the term (II)$\leq O(B_v d\sqrt{T}\log(KT)/\sqrt{p_{\min}})$ follows from the similar derivation of Equation (26) by replacing $(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2$ with $(\mu_{t,i}-\tilde{\mu}_{t,i})^2$. And the theorem is concluded by considering $\neg\mathcal{W}_t$ and $\neg\mathcal{N}_t$, similar to Appendix A.

Lemma 6 (Weighted Ellipsoidal Potential Lemma)**.** $\sum_{t=1}^{T}\sum_{i\in\tau_t}\|\boldsymbol{\phi}_t(i)\|^2_{\boldsymbol{G}_t^{-1}}/\bar{V}_{t,i}\leq 2d\log(1+KT/(\gamma d))\leq 2d\log T$ *when* $\neg F_T$ *and* $\gamma\geq 4K$.

Proof.

$$\begin{aligned}
\det(\boldsymbol{G}_{t+1}) &\overset{(a)}{=} \det\left(\boldsymbol{G}_t+\sum_{i\in\tau_t}\boldsymbol{\phi}_t(i)\boldsymbol{\phi}_t(i)^\top/\bar{V}_{t,i}\right)\\
&\overset{(b)}{=} \det(\boldsymbol{G}_t)\cdot\det\left(\boldsymbol{I}+\sum_{i\in\tau_t}G_t^{-1/2}\boldsymbol{\phi}_t(i)(G_t^{-1/2}\boldsymbol{\phi}_t(i))^\top/\bar{V}_{t,i}\right)\\
&\overset{(c)}{\geq} \det(\boldsymbol{G}_t)\cdot\left(1+\sum_{i\in\tau_t}\|\boldsymbol{\phi}_t(i)\|^2_{\boldsymbol{G}_t^{-1}}/\bar{V}_{t,i}\right)\\
&\overset{(d)}{\geq} \det(\gamma\boldsymbol{I})\prod_{s=1}^{t}\left(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{t,i}\right),
\end{aligned}\tag{27}$$

where (a) follows from the definition, (b) follows from $\det(\boldsymbol{AB})=\det(\boldsymbol{A})\det(\boldsymbol{B})$ and $\boldsymbol{A}+\boldsymbol{B}=\boldsymbol{A}^{1/2}(\boldsymbol{I}+\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2})$, (c) follows from Lemma 14, (d) follows from repeatedly applying (c).

If $\bar{V}_{s,i}=\frac{1}{4}$, $\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}\leq\frac{4\|\boldsymbol{\phi}_s(i)\|^2}{\lambda_{\min}(\boldsymbol{G}_s)}\leq 4/\gamma\leq 1/K$, else if $\bar{V}_{s,i}<\frac{1}{4}$, and since $\neg\mathcal{F}_T$, by Lemma 9, $\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}\leq\frac{1}{\rho(\delta)}\frac{1}{\sqrt{\gamma}}\leq\frac{1}{\gamma}\leq 1/(4K)$. Therefore, we have $\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}\leq 1$. Using the fact that $2\log(1+x)\geq x$ for any $[0,1]$, we have

$$\begin{aligned}
&\sum_{s\in t}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}\\
\leq& 2\sum_{s=1}^{t}\log\left(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}\right)\\
=& 2\log\prod_{s=1}^{t}\left(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|^2_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}\right)\\
\overset{(a)}{\leq}& 2\log\left(\frac{\det(\boldsymbol{G}_{t+1})}{\det(\gamma\boldsymbol{I})}\right)\\
\overset{(b)}{\leq}& 2\log\left(\frac{(\gamma+KT/d)^d}{\gamma^d}\right)=2d\log(1+4dK^2T^2/(\gamma d))\leq 4d\log(KT),
\end{aligned}$$

where the (a) follows from Equation (17), (b) follows from Lemma 15 by setting $L=\|\boldsymbol{\phi}_s(i)\|^2/\bar{V}_{s,i}\leq 4dKs$ (from Lemma 11). ■

B.3. Proof of Theorem 3 Under TPVM Condition

In this section, we consider two cases when $\lambda\geq 2$ and $\lambda\geq 1$. Recall that to use the TPVM condition (Condition 4), we need one additional condition over the triggering probability (Condition 5).

B.3.1. When $\lambda\geq 2$

.

We inherit the same notation and events as in Appendix A.1, and start to bound term (I) in Equation (26) differently,

$$\mathbb{E}\left[\sum_{t\in[T]}(\mathrm{I})\right] \overset{(a)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}(p_i^{\tilde{\boldsymbol{\mu}}_t,S_t})^2 \frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{28}$$

$$\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}\left(p_i^{\boldsymbol{\mu}_t,S_t} + \min\left\{1, \sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right\}\right)^2 \cdot \frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{29}$$

$$\overset{(c)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t} 3p_i^{\boldsymbol{\mu}_t,S_t} \cdot \frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right]$$

$$+\mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}\left(\sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right)^2 \cdot \frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{30}$$

$$\overset{(d)}{=} O(B_v d\sqrt{T}\log(KT)) + B_v\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}} \cdot \sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right] \tag{31}$$

$$\overset{(e)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\frac{1}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}} \cdot \sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right] \tag{32}$$

$$\overset{(f)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\frac{1}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}} \cdot \sqrt{K\sum_{j\in\tilde{S}_t} B_p^2 p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|^2}\right] \tag{33}$$

$$\overset{(g)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v B_p\frac{\sqrt{K}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}} \cdot \sqrt{\sum_{j\in\tilde{S}_t} p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|^2/\bar{V}_{t,i}}\right] \tag{34}$$

$$\overset{(h)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v B_p\frac{\sqrt{K}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\underline{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right] \tag{35}$$

$$\overset{(i)}{\leq} O(B_v d\sqrt{T}\log(KT) + B_v B_p\frac{\sqrt{K}}{\sqrt{p_{\min}}}d\log(KT)) = O(B_v d\sqrt{T}\log(KT)), \tag{36}$$

where (a) follows from Condition 4, (b) is by applying Condition 5 for triggering probability $p_i^{\bar{\boldsymbol{\mu}}_t,\tilde{S}_t}$ and $p_i^{\bar{\boldsymbol{\mu}}_t,\tilde{S}_t}, p_i^{\boldsymbol{\mu}_t,\tilde{S}_t} \leq 1$, (c) follows from $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$, (d) follows from same derivation from Equation (26), (e) follows from $p_i^{\tilde{\boldsymbol{\mu}}_t,\tilde{S}_t} \geq p_{\min}^i$, (f) follows from Cauchy-Schwarz, (g) follows from $\bar{V}_{t,i} \leq 1/4$, (h) follows from $\tilde{\mu}_{t,i}, \mu_{t,i} \in [\bar{\mu}_{t,i}, \underline{\mu}_{t,i}]$ by event $\mathcal{N}_t$, (i) follows from the similar analysis of (d)-(g) in Equation (26) inside the square-root without considering the additional $B_v\sqrt{T}/\sqrt{p_{\min}}$.

For the term (II), one can easily verify it follows from the similar deviation of the term (I) with the difference in constant terms. And Theorem 3 is concluded by considering small probability $\neg\mathcal{W}_t$ and $\mathcal{N}_t$ events.

B.3.2. When $\lambda \geq 1$

We inherit the same notation and events as in Appendix A.1, and start to bound term (I) in Equation (28) as follows,

$$\mathbb{E}\left[\sum_{t\in[T]}(\mathrm{I})\right] \overset{(a)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}(p_i^{\tilde{\boldsymbol{\mu}}_t,S_t}) \frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{37}$$

$$\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}\left(p_i^{\boldsymbol{\mu}_t,S_t} + \min\left\{1, \sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right\}\right)\cdot\frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{38}$$

$$\overset{(c)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t} p_i^{\boldsymbol{\mu}_t,S_t}\cdot\frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right]$$

$$+\mathbb{E}\left[\sum_{t=1}^{T} B_v \sqrt{\sum_{i\in\tilde{S}_t}\left(\sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|\right)\cdot\frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\right] \tag{39}$$

$$\overset{(d)}{=} O(B_v d\sqrt{T}\log(KT)) + B_v\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\cdot\sqrt{\sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|}\right] \tag{40}$$

$$\overset{(e)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\frac{1}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\cdot\sqrt{\sum_{j\in\tilde{S}_t} B_p p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|}\right] \tag{41}$$

$$\overset{(f)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\frac{1}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\cdot\left(K\sum_{j\in\tilde{S}_t} B_p^2 p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|^2\right)^{1/4}\right] \tag{42}$$

$$\overset{(g)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\sqrt{B_p}\frac{K^{1/4}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\sqrt{\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\tilde{\mu}_{t,i})^2}{\bar{V}_{t,i}}}\cdot\left(\sum_{j\in\tilde{S}_t} p_j^{\boldsymbol{\mu},S_t}\left|\mu_{t,j}-\tilde{\mu}_{t,j}\right|^2/\bar{V}_{t,i}\right)^{1/4}\right] \tag{43}$$

$$\overset{(h)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\sqrt{B_p}\frac{K^{1/4}}{\sqrt{p_{\min}}}\mathbb{E}\left[\sum_{t=1}^{T}\left(\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\underline{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right)^{3/4}\right] \tag{44}$$

$$\overset{(i)}{\leq} O(B_v d\sqrt{T}\log(KT)) + B_v\sqrt{B_p}\frac{(KT)^{1/4}}{\sqrt{p_{\min}}}\mathbb{E}\left[\left(\sum_{t=1}^{T}\sum_{i\in\tilde{S}_t}\frac{p_i^{\boldsymbol{\mu},S_t}(\bar{\mu}_{t,i}-\underline{\mu}_{t,i})^2}{\bar{V}_{t,i}}\right)^{3/4}\right] \tag{45}$$

$$\overset{(j)}{\leq} O\left(B_v d\sqrt{T}\log(KT) + B_v\sqrt{B_p}\frac{(KT)^{1/4}}{\sqrt{p_{\min}}}\left(d\log(KT)\right)^{3/4}\right) = O(B_v d\sqrt{T}\log(KT)), \tag{46}$$

where (a) follows from Condition 4, (b) is by applying Condition 5 for triggering probability $p_i^{\bar{\boldsymbol{\mu}}_t,\tilde{S}_t}$ and $p_i^{\bar{\boldsymbol{\mu}}_t,\tilde{S}_t}, p_i^{\boldsymbol{\mu}_t,\tilde{S}_t} \leq 1$, (c) follows from $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$, (d) follows from same derivation from Equation (26), (e) follows from $p_i^{\tilde{\boldsymbol{\mu}}_t,\tilde{S}_t} \geq p_{\min}^i$, (f) follows from Cauchy-Schwarz, (g) follows from $\bar{V}_{t,i}\leq 1/4$, (h) follows from $\tilde{\mu}_{t,i},\mu_{t,i}\in[\bar{\mu}_{t,i},\underline{\mu}_{t,i}]$ by event $\mathcal{N}_t$, (i) follows from Holder's inequality with $p=4, q=4/3$, (j) follows from the similar analysis of (d)-(g) in Equation (26) inside the square-root without considering the additional $B_v\sqrt{T}/\sqrt{p_{\min}}$.

For the term (II), one can easily verify it follows from the similar deviation of the term (I) with the difference in constant terms. And Theorem 3 is concluded by considering small probability $\neg\mathcal{W}_t$ and $\mathcal{N}_t$ events.

C. Proofs for TPE Trick to Improve Non-Contextual CMAB-T

We first introduce some definitions that are used in (Wang & Chen, 2017) and (Liu et al., 2022). Recall that non-contextual CMAB-T is a degenerate case when $\boldsymbol{\phi}_t(i)=\boldsymbol{e}_i$ and $\boldsymbol{\theta}^*=\boldsymbol{\mu}$, where $\boldsymbol{\mu}\triangleq\mathbb{E}_{\boldsymbol{X}_t\sim\mathcal{D}}[\boldsymbol{X}_t\mid\mathcal{H}_t]$ is the mean of the true outcome distribution $\mathcal{D}$.

Definition 1 ((Approximation) Gap)**.** *Fix a distribution $D\in\mathcal{D}$ and its mean vector $\boldsymbol{\mu}$, for each action $S\in\mathcal{S}$, we define the (approximation) gap as $\Delta_S=\max\{0,\alpha r(S^*;\boldsymbol{\mu})-r(S;\boldsymbol{\mu})\}$. For each arm i, we define $\Delta_i^{\min}=\inf_{S\in\mathcal{S}:p_i^{D,S}>0,\,\Delta_S>0}\Delta_S$,*

$\Delta_i^{\max} = \sup_{S\in\mathcal{S}:p_i^{D,S}>0,\Delta_S>0} \Delta_S$. *As a convention, if there is no action* $S \in \mathcal{S}$ *such that* $p_i^{D,S} > 0$ *and* $\Delta_S > 0$*, then* $\Delta_i^{\min} = +\infty, \Delta_i^{\max} = 0$. *We define* $\Delta_{\min} = \min_{i\in[m]} \Delta_i^{\min}$ *and* $\Delta_{\max} = \max_{i\in[m]} \Delta_i^{\max}$.

Definition 2 (Event-Filtered Regret)**.** *For any series of events* $(\mathcal{E}_t)_{t\in[T]}$ *indexed by round number* t*, we define the* $Reg^A_{\alpha,\boldsymbol{\mu}}(T, (\mathcal{E}_t)_{t\in[T]})$ *as the regret filtered by events* $(\mathcal{E}_t)_{t\in[T]}$*, or the regret is only counted in* t *if* $\mathcal{E}$ *happens in* t*. Formally,*

$$Reg^A_{\alpha,\boldsymbol{\mu}}(T, (\mathcal{E}_t)_{t\in[T]}) = \mathbb{E}\left[\sum_{t\in[T]} \mathbb{I}(\mathcal{E}_t)(\alpha \cdot r(S^*;\boldsymbol{\mu}) - r(S_t;\boldsymbol{\mu}))\right]. \tag{47}$$

For simplicity, we will omit $A, \alpha, \boldsymbol{\mu}, T$ *and rewrite* $Reg^A_{\alpha,\boldsymbol{\mu}}(T, (\mathcal{E}_t)_{t\in[T]})$ *as* $Reg(T, \mathcal{E}_t)$ *when contexts are clear.*

C.1. Reproducing Theorem 1 of (Wang & Chen, 2017) under 1-norm TPM Condition

Theorem 4. *For a CMAB-T problem instance* $([m], \mathcal{S}, \mathcal{D}, D_{trig}, R)$ *that satisfies monotonicity (Condition 1), and TPM bounded smoothness (Condition 2) with coefficient* B_1*, if* $\lambda \geq 1$*, CUCB (Wang & Chen, 2017) with an* (α, β)*-approximation oracle achieves an* (α, β)*-approximate distribution-dependent regret bounded by*

$$Reg(T) \leq \sum_{i\in[m]} \frac{48B_1^2 K \log T}{\Delta_i^{\min}} + 2B_1 m + \frac{\pi^2}{3} \cdot \Delta_{\max}. \tag{48}$$

And the distribution-independent regret,

$$Reg(T) \leq 14B_1\sqrt{mKT\log T} + 2B_1 m + \frac{\pi^2}{3} \cdot \Delta_{\max}. \tag{49}$$

The main idea is to use TPE trick to replace $\tilde{S}_t$ (arms that could be probabilistically triggered by action S_t) with τ_t (arms that are actually triggered by action S_t) under conditional expectation, so that we can use the simpler Appendix B.2 of Wang & Chen (2017) to avoid the much more involved Appendix B.3 of Wang & Chen (2017). Such replacement bypasses the triggering group analysis (and its counter $N_{t,i,j}$) in Appendix B.3, which uses $N_{t,i,j}$ to associate $T_{t,i}$ with the counters for $\tilde{S}_t$. For our simplified analysis, we can directly associate the $T_{t,i}$ with the arm triggering for the arms τ_t that are actually triggered/observed and eventually reproduce the regret bounds of (Wang & Chen, 2017).

We follow exactly the same CUCB algorithm (Algorithm 1 (Wang & Chen, 2017)), conditions (Condition 1, 2 (Wang & Chen, 2017)). We also inherit the event definitions of $\mathcal{N}_t^s$ (Definition 4 (Wang & Chen, 2017)) that for every arm $i \in [m]$, $|\hat{\mu}_{t-1,i} - \mu_i| < \rho_{t,i} = \sqrt{\frac{3\log t}{2T_{t-1,i}}}$, and the event F_t being $\{r(S_t;\bar{\mu}) < \alpha \cdot \text{opt}(\boldsymbol{\mu})\}$. Let us further denote $\Delta_{S_t} = \alpha r(S^*;\boldsymbol{\mu}) - r(S_t;\boldsymbol{\mu})$, τ_t be the arms actually triggered by S_t at time t. Let filtration $\mathcal{F}_{t-1}$ be $(\phi_1, S_1, \tau_1, (X_{1,i})_{i\in\tau_1}, ..., \phi_{t-1}, S_{t-1}, \tau_{t-1}, (X_{t-1,i})_{i\in\tau_{t-1}}, \phi_t, S_t)$, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ We also have that $\mathcal{F}_{t-1}$, $T_{t-1,i}, \hat{\mu}_{t,i}$ are measurable.

Proof. Under event $\mathcal{N}_t^s$ and $\neg F_t$, and given filtration $\mathcal{F}_{t-1}$, we have

$$\Delta_{S_t} \overset{(a)}{\leq} B_1 \sum_{i\in[m]} p_i^{D,S_t}(\bar{\mu}_{t,i} - \mu_i) \tag{50}$$

$$\overset{(b)}{\leq} -\Delta_{S_t} + 2B_1 \sum_{i\in[m]} p_i^{D,S_t}(\bar{\mu}_{t,i} - \mu_i) \tag{51}$$

$$= -\frac{\sum_{i\in[m]} p_i^{D,S_t}\Delta_{S_t}}{\sum_{i\in[m]} p_i^{D,S_t}} + 2B_1 \sum_{i\in[m]} p_i^{D,S_t}(\bar{\mu}_{t,i} - \mu_i) \tag{52}$$

$$\overset{(c)}{\leq} 2B_1 \sum_{i\in[m]} p_i^{D,S_t}\left(-\frac{\Delta_i^{\min}}{2B_1K} + (\bar{\mu}_{t,i} - \mu_i)\right) \tag{53}$$

$$\overset{(d)}{\leq} 2B_1 \sum_{i\in[m]} p_i^{D,S_t}\left(-\frac{\Delta_i^{\min}}{2B_1K} + \min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\}\right), \tag{54}$$

where (a) follows from exactly the Equation (10) of Appendix B.3 in Wang & Chen (2017), (b) is by the reverse amortization trick that multiplies two to both sides of (a) and rearranges the terms, (c) is by $p_i^{D,S_t} \leq 1$ and $\Delta_i^{\min} \leq \Delta_{S_t}$, (d) by event $\mathcal{N}_t^s$ so that $(\bar{\mu}_{t,i} - \mu_i) \leq \min\{1, 2\rho_{t,i}\} = \left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\}$.

Let

$$\kappa_{i,T}(\ell) = \begin{cases} 2B_1, & \text{if } \ell = 0, \\ 2B_1\sqrt{\frac{6\log T}{\ell}}, & \text{if } 1 \leq \ell \leq L_{i,T}, \\ 0, & \text{if } \ell > L_{i,T}, \end{cases} \tag{55}$$

where $L_{i,T} = \frac{24B_1^2K^2\log T}{(\Delta_i^{\min})^2}$.

It follows that

$$\Delta_{S_t} = \mathbb{E}_t[\Delta_{S_t}] \overset{(a)}{\leq} \mathbb{E}_t\left[2B_1 \sum_{i\in[m]} p_i^{D,S_t}\left(-\frac{\Delta_i^{\min}}{2B_1K} + \min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\}\right)\right] \tag{56}$$

$$\overset{(b)}{=} \mathbb{E}_t\left[2B_1 \sum_{i\in[m]} \mathbb{I}\{i \in \tau_t\}\left(-\frac{\Delta_i^{\min}}{2B_1K} + \min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\}\right)\right] \tag{57}$$

$$= \mathbb{E}_t\left[2B_1 \sum_{i\in\tau_t}\left(-\frac{\Delta_i^{\min}}{2B_1K} + \min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\}\right)\right] \tag{58}$$

$$\overset{(c)}{\leq} \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right], \tag{59}$$

where (a) follows from Equation (54), (b) follows from the TPE trick to replace $p_i^{D,S_t} = \mathbb{E}_t[\mathbb{I}\{i \in \tau_t\}]$, (c) follows from that if $T_{t-1,i} \leq L_{i,T}$, we have $\min\{\sqrt{\frac{6\log T}{T_{t-1,i}}}, 1\} \leq \frac{1}{2B_1}\kappa_{i,T}(T_{t-1,i})$, and if $T_{t-1,i} \geq L_{i,T} + 1$, then $\min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\} \leq \frac{\Delta_i^{\min}}{2B_1K}$, so $-\frac{\Delta_i^{\min}}{2B_1K} + \min\left\{1, \sqrt{\frac{6\log T}{T_{t-1,i}}}\right\} \leq 0 = \kappa_{i,T}(T_{t-1,i})$.

Now we apply the definition of the event-filtered regret,

$$Reg(\mathcal{N}_t^s, \neg F_t) = \mathbb{E}\left[\sum_{t=1}^T \Delta_{S_t}\right] \tag{60}$$

$$\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right]\right] \tag{61}$$

$$\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^T \sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right] \tag{62}$$

$$\overset{(c)}{=} \mathbb{E}\left[\sum_{i\in[m]} \sum_{s=0}^{T_{T-1,i}} \kappa_{i,T}(s)\right] \tag{63}$$

$$\leq \sum_{i\in[m]} \sum_{s=0}^{L_{i,T}} \kappa_{i,T}(s) \tag{64}$$

$$= 2B_1m + \sum_{i\in[m]} \sum_{s=1}^{L_{i,T}} 2B_1\sqrt{\frac{6\log T}{s}} \tag{65}$$

$$\overset{(d)}{\leq} 2B_1m + \sum_{i\in[m]} \int_{s=0}^{L_{i,T}} 2B_1\sqrt{\frac{6\log T}{s}} \cdot ds \tag{66}$$

$$\leq 2B_1 m + \sum_{i\in[m]} \frac{48B_1^2 K \log T}{\Delta_i^{\min}}, \tag{67}$$

where (a) follows from Equation (59), (b) follows from the tower rule, (c) follows from that $T_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$, (d) is by the sum & integral inequality $\int_{L-1}^{U} f(x)dx \geq \sum_{i=L}^{U} f(i) \geq \int_{L}^{U+1} f(x)dx$ for non-increasing function f.

Following Wang & Chen (2017) to handle small probability events $\neg\mathcal{N}_t^s$ and F_t, we have

$$Reg(T) \leq \sum_{i\in[m]} \frac{48B_1^2 K \log T}{\Delta_i^{\min}} + 2B_1 m + \frac{\pi^2}{3} \cdot \Delta_{\max}, \tag{68}$$

and the distribution-independent regret is

$$Reg(T) \leq 14B_1\sqrt{mKT\log T} + 2B_1 m + \frac{\pi^2}{3} \cdot \Delta_{\max}. \tag{69}$$

■

C.2. Improving Theorem 1 of (Liu et al., 2022) under TPVM Condition

We first show the regret bound of using our TPE technique in Theorem 5 and its prior result in Proposition 2.

Theorem 5. *For a CMAB-T problem instance $([m], \mathcal{S}, \mathcal{D}, D_{trig}, R)$ that satisfies monotonicity (Condition 1), and TPVM bounded smoothness (Condition 4) with coefficient (B_v, B_1, λ), if $\lambda \geq 1$, BCUCB-T (Liu et al., 2022) with an (α, β)-approximation oracle achieves an (α, β)-approximate distribution-dependent regret bounded by*

$$O\left(\sum_{i\in[m]} \frac{B_v^2 \log K \log T}{\tilde{\Delta}_{i,\lambda}^{\min}} + \sum_{i\in[m]} B_1 \log\left(\frac{B_1 K}{\Delta_i^{\min}}\right)\log T\right), \tag{70}$$

where $\tilde{\Delta}_{i,\lambda}^{\min} = \min_{S:p_i^{D,S}>0,\Delta_S>0} \Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}$. And the distribution-independent regret,

$$Reg(T) \leq O\left(B_v\sqrt{m(\log K)T\log T} + B_1 m \log(KT)\log T\right). \tag{71}$$

Proposition 2 (Theorem 1, Liu et al. (2022))**.** *For a CMAB-T problem instance $([m], \mathcal{S}, \mathcal{D}, D_{trig}, R)$ that satisfies monotonicity (Condition 1), and TPVM bounded smoothness (Condition 4) with coefficient (B_v, B_1, λ), if $\lambda \geq 1$, BCUCB-T with an (α, β)-approximation oracle achieves an (α, β)-approximate regret bounded by*

$$O\left(\sum_{i\in[m]} \log\left(\frac{B_v K}{\Delta_i^{\min}}\right)\frac{B_v^2 \log K \log T}{\Delta_i^{\min}} + \sum_{i\in[m]} B_1 \log^2\left(\frac{B_1 K}{\Delta_i^{\min}}\right)\log T\right). \tag{72}$$

And the distribution-independent regret,

$$Reg(T) \leq O\left(B_v\sqrt{m(\log K)T}\log(KT) + B_1 m \log^2(KT)\log T\right). \tag{73}$$

Looking at our regret bound (Theorem 5), there are two improvements compared with Proposition 2: (1) the min gap is improved to $\tilde{\Delta}_{i,\lambda}^{\min} \geq \Delta_i^{\min}$, (2) we remove a $O(\log(\frac{B_v K}{\Delta_i^{\min}}))$ for the leading term. For (2), it translates to a $O(\sqrt{\log T})$ improvement for the distribution-independent bound.

Proof. Similar to Appendix C.1, the main idea is to use TPE trick to replace $\tilde{S}_t$ (arms that could be probabilistically triggered by action S_t) with τ_t (arms that are actually triggered by action S_t) under conditional expectation to avoid the usage of much more involved triggering group analysis (Wang & Chen, 2017). Such replacement bypasses the triggering group analysis (and its counter $N_{t,i,j}$) (Liu et al., 2022), which uses $N_{t,i,j}$ to associate $T_{t,i}$ with the counters for $\tilde{S}_t$. By doing so, we do not need a union bound over the group index j, which saves a $\log(B_v K/\Delta_i^{\min}$ (or $\log(B_1 K/\Delta_i^{\min})$ factor.

We follow exactly the same BCUCB-T algorithm (Algorithm 1 (Liu et al., 2022)), conditions (Condition 1, 2, 3 (Liu et al., 2022)). We also inherit the event definitions of $\mathcal{N}_t^s$ (Definition 6 (Liu et al., 2022)) that (1) for every base arm $i \in [m]$, $|\hat{\mu}_{t-1,i} - \mu_i| \leq \rho_{t,i}$, where $\rho_{t,i} = \sqrt{\frac{6\hat{V}_{t-1,i}\log t}{T_{t-1,i}}} + \frac{9\log t}{T_{t-1,i}}$; (2) for every base arm $i \in [m]$, $\hat{V}_{t-1,i} \leq 2\mu_i(1-\mu_i) + \frac{3.5\log t}{T_{t-1,i}}$. We use the event F_t being $\{r(S_t;\bar{\mu}) < \alpha \cdot \text{opt}(\boldsymbol{\mu})\}$. Let us further denote $\Delta_{S_t} = \alpha r(S^*;\boldsymbol{\mu}) - r(S_t;\boldsymbol{\mu})$, τ_t be the arms actually triggered by S_t at time t. Let filtration $\mathcal{F}_{t-1}$ be $(\phi_1, S_1, \tau_1, (X_{1,i})_{i\in\tau_1}, ..., \phi_{t-1}, S_{t-1}, \tau_{t-1}, (X_{t-1,i})_{i\in\tau_{t-1}}, \phi_t, S_t)$, and let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ We also have that $\mathcal{F}_{t-1}, T_{t-1,i}, \hat{\mu}_{t,i}$ are measurable.

We follow the same regret decomposition as in Lemma 9 of Liu et al. (2022), to decompose the event-filtered regret $Reg(T, \mathcal{N}_t^s, \neg F_t)$ into two event-filtered regret $Reg(T, E_{t,1})$ and $Reg(T, E_{t,2})$ under events $\mathcal{N}_t^s, \neg F_t$.

$$Reg(T) \leq Reg(T, E_{t,1}) + Reg(T, E_{t,2}), \tag{74}$$

where event $E_{t,1} = \{\Delta_{S_t} \leq 2e_{t,1}(S_t)\}$, event $E_{t,2} = \{\Delta_{S_t} \leq 2e_{t,2}(S_t)\}$, $e_{t,1}(S_t) = 4\sqrt{3}B_v\sqrt{\sum_{i\in\tilde{S}_t}(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})(p_i^{D,S_t})^\lambda}$,$e_{t,2}(S_t) = 28B_1\sum_{i\in\tilde{S}_t}(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})(p_i^{D,S_t})$.

C.2.1. Bounding the $Reg(T, E_{t,1})$ term

We bound the leading $Reg(T, E_{t,1})$ term under two cases when $\lambda \in [1, 2)$ and $\lambda \in [2, \infty)$.

(a) When $\lambda \in [1, 2)$,

Let $c_1 = 4\sqrt{3}$, $\tilde{\Delta}_{S_t} = \Delta_{S_t}/(p_i^{D,S})^{\lambda-1}$. Given filtration $\mathcal{F}_{t-1}$ and event $E_{t,1}$, we have

$$\Delta_{S_t} \overset{(a)}{\leq} \sum_{i\in[m]} \frac{4c_1^2B_v^2(p_i^{D,S_t})^\lambda \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} \tag{75}$$

$$\overset{(b)}{=} -\Delta_{S_t} + 2\sum_{i\in[m]} \frac{4c_1^2B_v^2(p_i^{D,S_t})^\lambda \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} \tag{76}$$

$$= -\frac{\sum_{i\in\tilde{S}_t} p_i^{D,S_t}\Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}}{\sum_{i\in[m]}(p_i^{D,S_t})^{2-\lambda}} + 2\sum_{i\in[m]} \frac{4c_1^2B_v^2(p_i^{D,S_t})^\lambda \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} \tag{77}$$

$$\overset{(c)}{\leq} \sum_{i\in[m]} p_i^{D,S_t}\left(\frac{8c_1^2B_v^2\frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}} - \frac{\Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}}{K}\right) \tag{78}$$

$$\overset{(d)}{=} \sum_{i\in[m]} p_i^{D,S_t}\left(\frac{8c_1^2B_v^2\frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K}\right), \tag{79}$$

where (a) follows from event $E_{t,1}$, (b) is by the reverse amortization trick that multiplies two to both sides of (a) and rearranges the terms, (c), (d) are by definition of $K, \tilde{\Delta}_{S_t}$.

It follows that

$$\Delta_{S_t} = \mathbb{E}_t[\Delta_{S_t}] \overset{(a)}{\leq} \mathbb{E}_t\left[\sum_{i\in[m]} p_i^{D,S_t}\left(\frac{8c_1^2B_v^2\frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K}\right)\right] \tag{80}$$

$$= \mathbb{E}_t\left[\sum_{i\in\tau_t}\left(\frac{8c_1^2B_v^2\frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K}\right)\right] \tag{81}$$

$$\leq \mathbb{E}_t\left[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right] \tag{82}$$

where (a) follows from Equation (79), (b) follows from TPE trick to replace $p_i^{D,S_t} = \mathbb{E}_t[\mathbb{I}\{i \in \tau_t\}]$. (c) is because we define a regret allocation function

$$\kappa_{i,T}(\ell) = \begin{cases} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}}, & \text{if } \ell = 0, \\ 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{\ell}}, & \text{if } 1 \le \ell \le L_{i,T,1}, \\ \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min}} \frac{1}{\ell}, & \text{if } L_{i,T,1} < \ell \le L_{i,T,2}, \\ 0, & \text{if } \ell > L_{i,j,T,2}, \end{cases} \tag{83}$$

where $L_{i,T,1} = \frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_i^{\min})^2}$, $L_{i,T,2} = \frac{8c_1^2 B_v^2 K \log T}{(\tilde{\Delta}_i^{\min})^2}$, $\tilde{\Delta}_i^{\min} = \min_{S: p_i^{D,S} > 0, \Delta_S > 0} \Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}$, and (c) holds due to Lemma 16.

$$\begin{aligned}
Reg(T, E_{t,1}) &= \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{S_t}\right] && (84)\\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{t\in[T]} \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right]\right] && (85)\\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t\in[T]} \sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right] && (86)\\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{i\in[m]} \sum_{s=0}^{T_{T-1,i}} \kappa_{i,T}(s)\right] && (87)\\
&\le \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} + \sum_{i\in[m]} \sum_{s=1}^{L_{i,T,1}} 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{s}} + \sum_{i\in[m]} \sum_{s=L_{i,T,1}+1}^{L_{i,T,2}} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min}} \frac{1}{s} && (88)\\
&\le \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} + \sum_{i\in[m]} \int_{s=0}^{L_{i,T,1}} 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{s}} \cdot ds + \sum_{i\in[m]} \sum_{s=L_{i,T,1}}^{L_{i,T,2}} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min}} \frac{1}{s} \cdot ds && (89)\\
&\le \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} + \sum_{i\in[m]} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min}} (3 + \log K), && (90)
\end{aligned}$$

where (a) follows from Equation (82), (b) follows from the tower rule, (c) follows from that $T_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$.

(b) When $\lambda \in [2, \infty)$,

Let $\tilde{\Delta}_{S_t,\lambda} = \Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}$, $\tilde{\Delta}_{S_t} = \Delta_{S_t}/p_i^{D,S_t}$. Note that $\tilde{\Delta}_{S,\lambda} = \tilde{\Delta}_S$ when $\lambda = 2$, $\tilde{\Delta}_{S,\lambda} \ge \tilde{\Delta}_S$ when $\lambda \ge 2$, and $\tilde{\Delta}_{S,\lambda} \le \tilde{\Delta}_S$, when $\lambda \le 2$, for any i, S. Given filtration $\mathcal{F}_{t-1}$ and under event $E_{t,1}$, we have

$$\begin{aligned}
\Delta_{S_t} &\le \sum_{i\in[m]} \frac{4c_1^2 B_v^2 (p_i^{D,S_t})^{\lambda} \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} && (91)\\
&= -\Delta_{S_t} + 2\sum_{i\in[m]} \frac{4c_1^2 B_v^2 (p_i^{D,S_t})^{\lambda} \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} && (92)\\
&= -\frac{\sum_{i\in\tilde{S}_t} p_i^{D,S_t} \Delta_{S_t}/p_i^{D,S_t}}{K} + 2\sum_{i\in[m]} \frac{4c_1^2 B_v^2 (p_i^{D,S_t})^{\lambda} \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}} && (93)
\end{aligned}$$

$$\leq \sum_{i\in[m]} p_i^{D,S_t} \left(\frac{8c_1^2 B_v^2 \frac{\log t}{T_{i,t-1}}}{\Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}} - \frac{\Delta_{S_t}/p_i^{D,S_t}}{K} \right) \tag{94}$$

$$= \sum_{i\in[m]} p_i^{D,S_t} \left(\frac{8c_1^2 B_v^2 \frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} \right). \tag{95}$$

$$\Delta_{S_t} = \mathbb{E}_t[\Delta_{S_t}] \leq \mathbb{E}_t \left[\sum_{i\in[m]} p_i^{D,S_t} \left(\frac{8c_1^2 B_v^2 \frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} \right) \right] \tag{96}$$

$$= \mathbb{E}_t \left[\sum_{i\in\tau_t} \left(\frac{8c_1^2 B_v^2 \frac{\log t}{T_{i,t-1}}}{\tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} \right) \right] \tag{97}$$

$$\leq \mathbb{E}_t \left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i}) \right] \tag{98}$$

where the last inequality is by Lemma 17 and we define a regret allocation function

$$\kappa_{i,T}(\ell) = \begin{cases} \frac{c_1^2 B_v^2}{\tilde{\Delta}_{i,\lambda}^{\min}}, & \text{if } \ell = 0, \\ 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{\ell}}, & \text{if } 1 \leq \ell \leq L_{i,T,1}, \\ \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_{i,\lambda}^{\min}} \frac{1}{\ell}, & \text{if } L_{i,T,1} < \ell \leq L_{i,T,2}, \\ 0, & \text{if } \ell > L_{i,j,T,2}, \end{cases} \tag{99}$$

where $L_{i,T,1} = \frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min} \cdot \tilde{\Delta}_{i,\lambda}^{\min}}$, $L_{i,T,2} = \frac{8c_1^2 B_v^2 K \log T}{\tilde{\Delta}_i^{\min} \cdot \tilde{\Delta}_{i,\lambda}^{\min}}$, $\tilde{\Delta}_i^{\min} = \min_{S:p_i^{D,S}>0,\Delta_S>0} \Delta_{S_t}/p_i^{D,S_t}$, $\tilde{\Delta}_{i,\lambda}^{\min} = \min_{S:p_i^{D,S}>0,\Delta_S>0} \Delta_{S_t}/(p_i^{D,S_t})^{\lambda-1}$.

$$Reg(T, E_{t,1}) = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{S_t}\right] \tag{100}$$

$$\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t\in[T]} \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right]\right] \tag{101}$$

$$\overset{(b)}{=} \mathbb{E}\left[\sum_{t\in[T]} \sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right] \tag{102}$$

$$\overset{(c)}{=} \mathbb{E}\left[\sum_{i\in[m]} \sum_{s=0}^{T_{T-1,i}} \kappa_{i,T}(s)\right] \tag{103}$$

$$\leq \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} + \sum_{i\in[m]} \sum_{s=1}^{L_{i,T,1}} 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{s}} + \sum_{i\in[m]} \sum_{s=L_{i,T,1}+1}^{L_{i,T,2}} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_i^{\min}} \frac{1}{s} \tag{104}$$

$$\leq \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} + \sum_{i\in[m]} \int_{s=0}^{L_{i,T,1}} 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{s}} \cdot ds + \sum_{i\in[m]} \sum_{s=L_{i,T,1}}^{L_{i,T,2}} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_{i,\lambda}^{\min}} \frac{1}{s} \cdot ds \tag{105}$$

$$\leq \sum_{i\in[m]} \frac{c_1^2 B_v^2}{\tilde{\Delta}_{i,\lambda}^{\min}} + \sum_{i\in[m]} \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_{i,\lambda}^{\min}} (1 + \log K) + \sum_{i\in[m]} \frac{16c_1^2 B_v^2 \log T}{\sqrt{\tilde{\Delta}_{i,\lambda}^{\min} \cdot \tilde{\Delta}_i^{\min}}}, \tag{106}$$

where (a) follows from Equation (98), (b) follows from the tower rule, (c) follows from that $T_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$.

C.2.2. Bounding the $Reg(T, E_{t,2})$ term

Let $c_2 = 28$. Given filtration $\mathcal{F}_{t-1}$ and event $E_{t,2}$, we have

$$\begin{aligned}
\Delta_{S_t} &\overset{(a)}{\leq} \sum_{i \in \tilde{S}_t} 2c_2 B_1 p_i^{D,S_t} \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\} \\
&\overset{(b)}{\leq} -\Delta_{S_t} + 2\sum_{i \in \tilde{S}_t} 2c_2 B_1 p_i^{D,S_t} \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\} \\
&= -\frac{\sum_{i\in[m]} p_i^{D,S_t}\Delta_{S_t}}{\sum_{i\in[m]} p_i^{D,S_t}} + 2\sum_{i\in[m]} 2c_2 B_1 p_i^{D,S_t} \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\} \qquad (107) \\
&\overset{(c)}{\leq} \sum_{i\in[m]} p_i^{D,S_t}\left(-\frac{\Delta_{S_t}}{K} + 4c_2 B_1 \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\}\right), \qquad (108)
\end{aligned}$$

where (a) follows from event $E_{t,2}$, (b) is by the reverse amortization trick that multiplies two to both sides of (a) and rearranges the terms, (c) follows from $p_i^{D,S_t} \leq 1$.

It follows that

$$\begin{aligned}
\Delta_{S_t} = \mathbb{E}_t[\Delta_{S_t}] &\overset{(a)}{\leq} \mathbb{E}_t\left[\sum_{i\in[m]} p_i^{D,S_t}\left(-\frac{\Delta_{S_t}}{K} + 4c_2 B_1 \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\}\right)\right] \\
&\overset{(b)}{=} \mathbb{E}_t\left[\sum_{i\in\tau_t}\left(-\frac{\Delta_{S_t}}{K} + 4c_2 B_1 \min\left\{1/28, \frac{\log T}{T_{t-1,i}}\right\}\right)\right] \\
&\overset{(c)}{\leq} \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right] \qquad (109)
\end{aligned}$$

regret allocation function

$$\kappa_{i,T}(\ell) = \begin{cases} \Delta_i^{\max}, & \text{if } 0 \leq \ell \leq L_{i,T,1} \\ \frac{4c_2 B_1 \log T}{\ell}, & \text{if } L_{i,j,1} < \ell \leq L_{i,j,2} \\ 0, & \text{if } \ell > L_{i,T,2} + 1, \end{cases} \qquad (110)$$

where $L_{i,T,1} = \frac{4c_2 B_1 \log T}{\Delta_i^{\max}}$, $L_{i,T,2} = \frac{4c_2 B_1 K \log T}{\Delta_i^{\min}}$. And (a) follows from Equation (109), (b) from the TPE, (c) follows from Lemma 18.

$$\begin{aligned}
Reg(T, E_{t,2}) &= \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{S_t}\right] \qquad (111) \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sum_{t\in[T]} \mathbb{E}_t\left[\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right]\right] \qquad (112) \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\tau_t} \kappa_{i,T}(T_{t-1,i})\right] \qquad (113) \\
&\overset{(c)}{=} \mathbb{E}\left[\sum_{i\in[m]}\sum_{s=0}^{T_{T-1,i}} \kappa_{i,T}(s)\right] \qquad (114) \\
&\leq m\Delta_{\max} + \sum_{i\in[m]}\sum_{\ell=1}^{L_{i,T,1}} \Delta_i^{\max} + \sum_{i\in[m]}\sum_{L_{i,T,1}+1}^{L_{i,T,2}} \frac{4c_2 B_1 \log T}{\ell} \qquad (115)
\end{aligned}$$

$$\leq m\Delta_{\max} + \sum_{i\in[m]} 4c_2 B_1 \log T + \sum_{i\in[m]} 4c_2 B_1 \log(\frac{K\Delta_i^{\max}}{\Delta_i^{\min}}) \log T \tag{116}$$

$$= m\Delta_{\max} + \sum_{i\in[m]} 4c_2 B_1 \left(1 + \log(\frac{K\Delta_i^{\max}}{\Delta_i^{\min}})\right) \log T \tag{117}$$

$$\leq m\Delta_{\max} + \sum_{i\in[m]} 4c_2 B_1 \left(1 + \log(\frac{K\Delta_i^{\max}}{\Delta_i^{\min}})\right) \log T, \tag{118}$$

where (a) follows from Equation (109), (b) follows from the tower rule, (c) follows from that $T_{t-1,i}$ is increased by 1 if and only if $i \in \tau_t$. ■

Similar to (Wang & Chen, 2017, Appendix B.3), for the distribution-independent regret bound, we fix a gap Δ to be decided later and we consider two events on Δ_{S_t}: $\{\Delta_{S_t} \leq \Delta\}$ and $\{\Delta_{S_t} > \Delta\}$.

For the former case, the regret is trivially $Reg(T, \{\Delta_{S_t} \leq \Delta\}) \leq T\Delta$. For the later case, under $\{\Delta_{S_t} > \Delta\}$ it is also straightforward to replace all $\Delta_i^{\min}$ with Δ and derive $Reg(T, \{\Delta_{S_t} > \Delta\}) \leq O\left(\frac{mB_v^2 \log K \log T}{\Delta} + mB_1 \log(\frac{B_1 K}{\Delta}) \log T\right)$.

By selecting $\Delta = \Theta\left(\sqrt{\frac{mB_v^2 \log T \log K}{T}} + \frac{B_1 m \log K \log T}{T}\right)$, we have

$$Reg(T) \leq O\left(B_v \sqrt{m(\log K)T \log T} + B_1 m \log(KT) \log T\right) \tag{119}$$

D. Applications

For convenience, we show our table again in Table 3.

Table 3. Summary of the coefficients, regret bounds and improvements for various applications.

Application	Condition	(B_v, B_1, λ)	Regret	Improvement
Online Influence Maximization (Wen et al., 2017)	TPM	$(-, \lvert V\rvert, -)$ †	$O(d\lvert V\rvert\sqrt{\lvert E\rvert T \log T})$	$\tilde{O}(\sqrt{\lvert E\rvert})$
Disjunctive Combinatorial Cascading Bandits (Li et al., 2016)	TPVM	$(1, 1, 2)$	$O(d\sqrt{T} \log T)$	$\tilde{O}(\sqrt{K}/p_{\min})$‡
Conjunctive Combinatorial Cascading Bandits (Li et al., 2016)	TPVM	$(1, 1, 1)$	$O(d\sqrt{T} \log T)$	$\tilde{O}(\sqrt{K}/r_{\max})$
Linear Cascading Bandits (Vial et al., 2022)*	TPVM	$(1, 1, 2)$	$O(d\sqrt{T} \log T)$	$\tilde{O}(\sqrt{K/d})$‡
Multi-layered Network Exploration (Liu et al., 2021b)	TPVM	$(\sqrt{1.25\lvert V\rvert}, 1, 2)$ †	$O(d\sqrt{\lvert V\rvert T} \log T)$	$\tilde{O}(\sqrt{n}/p_{\min})$
Probabilistic Maximum Coverage (Chen et al., 2013)*∗	VM	$(3\sqrt{2\lvert V\rvert}, 1, -)$	$O(d\sqrt{\lvert V\rvert T} \log T)$	$\tilde{O}(\sqrt{k})$

† $|V|, |E|, n, k, L$ denotes the number of target nodes, the number of edges that can be triggered by the set of seed nodes, the number of layers, the number of seed nodes and the length of the longest directed path, respectively; ‡ K is the length of the ordered list, $r_{\max} = \alpha \cdot \max_{t\in[T], S\in\mathcal{S}} r(S; \boldsymbol{\mu}_t)$;
* A special case of disjunctive combinatorial cascading bandits. ** This row is for C^{2}MAB application and the rest of rows are for C^{2}MAB-T applications.

D.1. Online Influence Maximization Bandit (Wang & Chen, 2017) and Its Contextual Generalization (Wen et al., 2017)

Following the setting of (Wang & Chen, 2017, Section 2.1), we consider a weighted directed graph $G(V, E, p)$, where V is the set of vertices, E is the set of directed edges, and each edge $(u, v) \in E$ is associated with a probability $p(u, v) \in [0, 1]$. When the agent selects a set of seed nodes $S \subseteq V$, the influence propagates as follows: At time 0, the seed nodes S are activated; At time $t > 1$, a node u activated at time $t - 1$ will have one chance to activate its inactive out-neighbor v with independent probability $p(u, v)$. The influence spread of S is denoted as $\sigma(S)$ and is defined as the expected number of activated nodes after the propagation process ends. The problem of Influence Maximization is to find seed nodes S with $|S| \leq k$ so that the influence spread $\sigma(S)$ is maximized.

For the problem of online influence maximization (OIM), we consider T rounds repeated influence maximization tasks and the edge probabilities $p(u, v)$ are assumed to be unknown initially. For each round $t \in [T]$, the agent selects k seed nodes as S_t, the influence propagation of S_t is observed and the reward is the number of nodes activated in round t. The agent's goal is to accumulate as much reward as possible in T rounds. The OIM fits into CMAB-T framework: the edges E are the set of base arms $[m]$, the (unknown) outcome distribution D is the joint of m independent Bernoulli random variables for the edge

set E, the action S are any seed node sets with size k at most k. For the arm triggering, the triggered set τ_t is the set of edges (u, v) whose source node u is reachable from S_t. Let X_t be the outcomes of the edges E according to probability $p(u, v)$ and the live-edge graph $G_t^{\text{live}}(V, E^{\text{live}})$ be a induced graph with edges that are alive, i.e., $e \in E^{\text{live}}$ iff $X_{t,e} = 1$ for $e \in E$. The triggering probability distribution $D_{\text{trig}}(S_t, X_t)$ degenerates to a deterministic triggered set, i.e., τ_t is deterministically decided given S_t and X_t. The reward $R(S_t, X_t, \tau_t)$ equals to the number activated nodes at the end of t, i.e., the nodes that are reachable from S_t in the live-edge graph G_t^{live}. The offline oracle is a $(1 - 1/e - \varepsilon, 1/|V|)$-approximation algorithm given by the greedy algorithm from (Kempe et al., 2003).

Now consider OIM's contextual generalization for large-scale OIM, we follow Wen et al. (2017), where each edge $e = (u, v)$ is associated with a known feature vector $\boldsymbol{x}_e \in \mathbb{R}^d$ and an unknown parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$, and the edge probability is $p(u, v) = \langle \boldsymbol{x}_e, \boldsymbol{\theta}^* \rangle$. By Lemma 2 of (Wang & Chen, 2017), $B_1 = \tilde{C} \leq |V|$, where $\tilde{C}$ is the largest number of nodes any node can reach and batch size $K \leq |E|$, so by Theorem 1, C^2-UCB-T obtain a worst-case $\tilde{O}(d|V|\sqrt{|E|T})$ regret bound. Compared with IMLinUCB algorithm (Wen et al., 2017) that achieves $\tilde{O}(d(|V| - k)|E|\sqrt{T})$, our regret achieves a improvement up to a factor of $\tilde{O}(\sqrt{|E|})$.

Now for the claim of the triggering probability satisfies $B_p = B_1$, it follows from the Theorem 4 of Li et al. (2020) by identifying $f(S, w, v) = p_i^{w,S}$.

D.2. Contextual Combinatorial Cascading Bandits (Li et al., 2016)

Contextual Combinatorial cascading bandits have two categories: conjunctive cascading bandits and disjunctive cascading bandits (Li et al., 2016). We also compare with a special case of linear cascading bandits that also uses variance-adaptive algorithms and achieve very competitive results.

Disjunctive form. For the disjunctive form, we want to select an ordered list S of K items from total L items, so as to maximize the probability that at least one of the outcomes of the selected items are 1. Each item is associated with a Bernoulli random variable with mean $\mu_{t,i} \in [0, 1]$ at round t, indicating whether the user will be satisfied with the item if he scans the item. To leverage the contextual information, Li et al. (2016) assumes $\boldsymbol{\mu}_{t,i} = \langle \boldsymbol{x}_{t,i}, \boldsymbol{\theta}^* \rangle$, where $\boldsymbol{x}_{t,i} \in \mathbb{R}^d$ is the known context at round t for arm i, $\boldsymbol{\theta} \in \mathbb{R}^d$ is the unknown parameter to be learned. This setting models the movie recommendation system where the user sequentially scans a list of recommended items and the system is rewarded when the user is satisfied with any recommended item. After the user is satisfied with any item or scans all K items but is not satisfied with any of them, the user leaves the system. Due to this stopping rule, the agent can only observe the outcome of items until (including) the first item whose outcome is 1. If there are no satisfactory items, the outcomes must be all 0. In other words, the triggered set is the prefix set of items until the stopping condition holds.

Without loss of generality, let the action be $\{1, ..., K\}$, then the reward function is $r(S; \boldsymbol{\mu}) = 1 - \prod_{j=1}^{K}(1 - \mu_j)$ and the triggering probability is $p_i^{\boldsymbol{\mu},S} = \prod_{j=1}^{i-1}(1 - \mu_j)$. Let $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, ..., \bar{\mu}_K)$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$, where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$ with $\bar{\boldsymbol{\mu}}, \boldsymbol{\mu} \in (0, 1)^K, \boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1, 1]^K$. By Lemma 19 in Liu et al. (2021a), disjunctive CB satisfies Condition 4 with $(B_v, B_1, \lambda) = (1, 1, 2)$. Also, we can verify that disjunctive CB also satisfies $B_p = B_1 = 1$ as follows:

$$\begin{aligned}
&\left| p_i^{\bar{\boldsymbol{\mu}},S} - p_i^{\boldsymbol{\mu},S} \right| \\
&= \left| \prod_{j=1}^{i}(1 - \mu_j) - \prod_{j=1}^{i}(1 - \bar{\mu}_j) \right| && (120) \\
&= \sum_{j=1}^{i} |\bar{\mu}_j - \mu_j| \, (1 - \mu_1)...(1 - \mu_{j-1})(1 - \bar{\mu}_{j+1})...(1 - \bar{\mu}_i) && (121) \\
&\leq \sum_{j=1}^{i} |\bar{\mu}_j - \mu_j| \, (1 - \mu_1)...(1 - \mu_{j-1}) && (122) \\
&= \sum_{j=1}^{i} |\bar{\mu}_j - \mu_j| \, p_j^{\boldsymbol{\mu},S}. && (123)
\end{aligned}$$

Now by Theorem 3, VAC^2-UCB obtains a regret bound of $O(d\sqrt{T}\log T)$. Compared with Corollary 4.5 in Li et al. (2016) that yields a $O(d\sqrt{KT}\log T/p_{\min})$ regret, our results improves by a factor of $O(\sqrt{K}/p_{\min})$.

Conjunctive form. For the conjunctive form, the learning agent wants to select K paths from total L paths (i.e., base arms) so as to maximize the probability that the outcomes of the selected paths are all 1. Each item is associated with a Bernoulli random variable with mean $\mu_{t,i}$ at round t, indicating whether the path will be live if the package will transmit via this path. Such a setting models the network routing problem (Kveton et al., 2015a), where the items are routing paths and the package is delivered when all paths are alive. The learning agent will observe the outcome of the first few paths till the first one that is down, since the transmission will stop if any of the path is down. In other words, the triggered set is the prefix set of paths until the stopping condition holds.

Without loss of generality, let the action be $\{1, ..., K\}$, then the reward function is $r(S; \boldsymbol{\mu}) = 1 - \prod_{j=1}^{K}(\mu_j)$ and the triggering probability is $p_i^{\boldsymbol{\mu},S} = \prod_{j=1}^{i-1}(\mu_j)$. Let $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, ..., \bar{\mu}_K)$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$, where $\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$ with $\bar{\boldsymbol{\mu}}, \boldsymbol{\mu} \in (0,1)^K, \boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^K$. By Lemma 20 in Liu et al. (2021a), conjunctive CB satisfies Condition 4 with $(B_v, B_1, \lambda) = (1, 1, 1)$. Also, we can verify that conjunctive CB also satisfies $B_p = B_1 = 1$ as follows:

$$\begin{aligned}
&\left|p_i^{\bar{\boldsymbol{\mu}},S} - p_i^{\boldsymbol{\mu},S}\right| \\
&= \left|\prod_{j=1}^{i}\mu_j - \prod_{j=1}^{i}\bar{\mu}_j\right| && (124)\\
&= \sum_{j=1}^{i}|\bar{\mu}_j - \mu_j|\,(\mu_1)...(\mu_{j-1})(\bar{\mu}_{j+1})...(\bar{\mu}_i) && (125)\\
&\leq \sum_{j=1}^{i}|\bar{\mu}_j - \mu_j|\,(\mu_1)...(\mu_{j-1}) && (126)\\
&= \sum_{j=1}^{i}|\bar{\mu}_j - \mu_j|\,p_j^{\boldsymbol{\mu},S}. && (127)
\end{aligned}$$

Now by Theorem 3, VAC^2-UCB obtains a regret bound of $O(d\sqrt{T}\log T)$. Compared with Corollary 4.6 in Li et al. (2016) that yields a $O(d\sqrt{KT}\log T/r_{\max})$ regret, our results improves by a factor of $O(\sqrt{K}/r_{\max})$.

Linear Cascading Bandit. Linear cascading bandit (Vial et al., 2022) is a special case of combinatorial cascading bandit (Li et al., 2016). The former assumes that action space $\mathcal{S}$ is the collection of all permutations whose size equals to K (i.e., a uniform matroid). In this case, the items in the feasible solutions are **exchangeable** (a critical property for matroids), i.e., $S - \{e_1\} + \{e_2\} \in \mathcal{S}$, for any $S \in \mathcal{S}, e_1, e_2 \in [m]$. Based on this property, their analysis can get the correct results. For the latter, however, $\mathcal{S}$ (i.e., Θ in [16]) consists of arbitrary feasible actions (perhaps with different sizes), e.g., $S \in \mathcal{S}$ could refer to any path that connects the source and the destination in network routing applications.

Other than the above difference, linear cascading bandits follow the same setting as disjunctive contextual combinatorial bandits. Following the similar argument of disjunctive contextual combinatorial bandits, the regret bound of VAC^2-UCB is $O(d\sqrt{T}\log T)$. Compared with CascadeWOFUL that achieves $\tilde{O}(\sqrt{d(d+K)T})$ by Theorem 4 in Vial et al. (2022), our regret improves a factor of $\tilde{O}(\sqrt{1+K/d})$. For the empirical comparison, see Section 5 for details.

D.3. Multi-layered Network Exploration Problem (MuLaNE) (Liu et al., 2021b)

We consider the MuLaNE problem with random node weights. After we apply the bipartite coverage graph, the corresponding graph is a tri-partite graph (n, V, R) (i.e., a 3-layered graph where the first layer and the second layer forms a bipartite graph, and the second and the third layer forms another bipartite graph), where the left nodes represent n random walkers; Middle nodes are $|V|$ possible targets V to be explored; Right nodes R are V nodes, each of which has only one edge connecting the middle edge. The MuLaNE task is to allocate B budgets into n layers to explore target nodes V and the base arms are $\mathcal{A} = \{(i, u, b) : i \in [n], u \in V, b \in [B]\}$.

With budget allocation $k_1, ..., k_L$, the (effective) base arms consist of two parts:

(1) $\{(i, j) : i \in [n], j \in V\}$, each of which is associated with visiting probability $x_{i,j} \in [0, 1]$ indicating whether node j will be visited by explorer i given k_i budgets. All these base arms corresponds to budget $k_i, i \in [n]$ are triggered.

(2) $y_j \in [0, 1]$ for $j \in V$ represents the random node weight. The triggering probability $p_j^{\boldsymbol{\mu},S} = 1 - \prod_{i\in[n]}(1 - x_{i,j})$.

For its contextual generalization, we assume $x_{i,j} = \langle \boldsymbol{\phi}_x(i,j), \boldsymbol{\theta}^* \rangle$, $y_j = \langle \boldsymbol{\phi}_y(j), \boldsymbol{\theta}^* \rangle$, where $\boldsymbol{\phi}_x(i,j), \boldsymbol{\phi}_y(j)$ are the known features for visiting probability and the node weights for large-scale MuLaNE applications, respectively. Let effective base arms $\boldsymbol{\mu} = (\boldsymbol{x}, \boldsymbol{y}) \in (0,1)^{(n|V|+|V|)}$, $\bar{\boldsymbol{\mu}} = (\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \in (0,1)^{(n|V|+|V|)}$, where $\bar{\boldsymbol{x}} = \boldsymbol{\zeta}_x + \boldsymbol{\eta}_x + \boldsymbol{x}, \bar{\boldsymbol{y}} = \boldsymbol{\zeta}_y + \boldsymbol{\eta}_y + \boldsymbol{y}$, for $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^{(n|V|+|V|)}$. For the target node $j \in V$, the per-target reward function $r_j(S; \boldsymbol{x}, \boldsymbol{y}) = y_j(1 - \prod_{i\in[n]}(1 - x_{i,j}))$. Denote $\bar{p}_j^{\boldsymbol{\mu},S} = 1 - \prod_{i\in[n]}(1 - \bar{x}_{i,j})$. Based on Lemma 21 in Liu et al. (2022), contextual MuLaNE satisfies Condition 4 with $(B_v, B_1, \lambda) = (\sqrt{1.25|V|}, 1, 2)$. To validate that this application satisfies Condition 5 with $B_p = B_1 = 1$, we have

$$\begin{aligned}
&\left| p_j^{\bar{\boldsymbol{\mu}},S} - p_j^{\boldsymbol{\mu},S} \right| \\
&= \left| \prod_{i\in[n]}(1 - x_{i,j}) - \prod_{i\in[n]}(1 - \bar{x}_{i,j}) \right| && (128)\\
&= \sum_{i\in[n]} |\bar{x}_{i,j} - x_{i,j}| \, (1 - x_{1,j})...(1 - x_{i-1,j})(1 - \bar{x}_{i+1,j})...(1 - \bar{x}_{i,j}) && (129)\\
&= \sum_{i\in[n]} |\bar{x}_{i,j} - x_{i,j}| \, . && (130)
\end{aligned}$$

By Theorem 3, we obtain $O(d\sqrt{|V|T}\log T)$, which improves the result $O(d\sqrt{n|V|T}\log T/p_{\min})$ that follows the result of C^{3}UCB algorithm (Li et al., 2016) by a factor of $O(\sqrt{n/p_{\min}})$.

D.4. Probabilistic Maximum Coverage Bandit (Chen et al., 2016a; Merlis & Mannor, 2019)

In this section, we consider the probabilistic maximum coverage (PMC) problem. PMC is modeled by a weighted bipartite graph $G = (L, V, E)$, where L are the source nodes, V is the target nodes and each edge $(u,v) \in E$ is associated with a probability $p(u,v)$. The task of PMC is to select a set $S \subseteq L$ of size k so as to maximize the expected number of nodes activated in V, where a node $v \in V$ can be activated by a node $u \in S$ with an independent probability $p(u,v)$. PMC can naturally models the advertisement placement application, where L are candidate web-pages, V are the set of users, and $p(u,v)$ is the probability that a user v will click on web-page u.

PMC fits into the non-triggering CMAB framework: each edge $(u,v) \in E$ corresponds to a base arm, the action is the set of edges that are incident to the set $S \subseteq L$, the unknown mean vectors $\boldsymbol{\mu} \in (0,1)^E$ with $\mu_{u,v} = p(u,v)$ and we assume they are independent across all base arms. In this context, the reward function $r(S; \boldsymbol{\mu}) = \sum_{v\in V}(1 - \prod_{u\in S}(1 - \mu_{u,v}))$.

In this paper, we consider a contextual generalization by assuming that $p(u,v) = \langle \boldsymbol{\phi}(u,v), \boldsymbol{\theta}^* \rangle$, where $\boldsymbol{\phi}(u,v) \in \mathbb{R}^d$ is the known context and $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is the unknown parameter. By Lemma 24 in Liu et al. (2022), PMC satisfies Condition 3 with $(B_v, B_1) = (3\sqrt{2|V|}, 1)$. Following Theorem 2, VAC^2-UCB obtains $O(d\sqrt{|V|T}\log T)$, which improves the C^{3}UCB algorithm's bound $O(d\sqrt{k|V|T}\log T)$ (Li et al., 2016) by a factor of $O(\sqrt{k})$.

E. Experiments

Synthetic data. We consider the same disjunctive linear cascading bandit setting as in (Vial et al., 2022), where the goal is to choose $K \in \{2i\}_{i=2}^8$ out of $m = 100$ items to maximize the reward. Notice that the linear cascading bandit problem is a simplified version of the contextual cascading bandit problem where the feature vectors of base arms are fixed in all rounds (see Appendix D.2 for details). For each K, we sample the click probability μ_i of item i uniformly in $[\frac{2}{3K}, \frac{1}{K}]$ for $i \le K$ and in $[0, \frac{1}{3K}]$ for $i > K$. We vary $d \in \{2i\}_{i=2}^8$ to generate the same $\boldsymbol{\mu}$ and compute unit-norm vectors θ^* and $\phi(i)$ satisfying $\mu_i = \langle \theta^*, \phi(i) \rangle$. We compare VAC^2-UCB to C^3-UCB (Li et al., 2016) and CascadeWOFUL (Vial et al., 2022): C^3-UCB is the variance-agnostic cascading bandit algorithm (essentially the same as CascadeLinUCB (Zong et al., 2016) in the linear cascading setting by using the tunable parameter $\sigma = 1$) and CascadeWOFUL is the state-of-the-art variance-aware cascading bandit algorithm. As shown in Figure 2, the regret of our VAC^2-UCB algorithm has superior dependence on K and d over that of C^3-UCB. When $d = K = 10$, VAC^2-UCB achieves sublinear regret; it incurs 75% and 13% less regret than C^3-UCB and CascadeWOFUL after $100,000$ rounds. Notice that CascadeWOFUL is also variance-aware but specifically designed for cascading bandits, while our algorithm can be applied to general C^{2}MAB-T.

Real data. We conduct experiments on the MovieLens-1M dataset which contains user ratings for $m \approx 4000$ movies. Following the same experimental setup in (Vial et al., 2022), we set $d = 20, K = 4$, and the goal is to choose K out

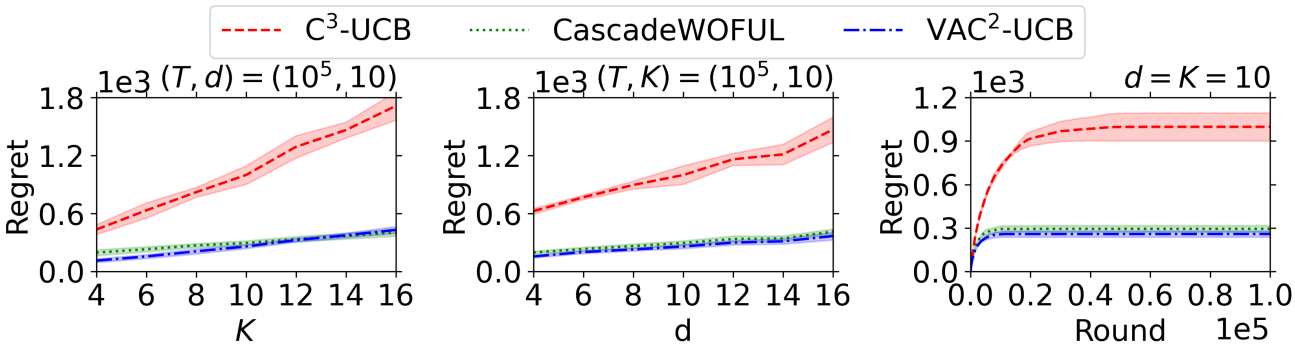


Figure 2. Results for synthetic data

of m movies to maximize the reward of the cascading recommendation. We use their learned feature mapping ϕ from movies to the probability that a uniformly random user rated the movie more than three stars. We point the reader to Section 6 of (Vial et al., 2022) for more details. In each round, we sample a random user J_t and define the potential click result $X_{t,i} = \mathbb{I}\{\text{user } J_t \text{ rated movie } i \text{ more than 3 stars}\}$. In other words, we observe the actual feedback of user J_t instead of using the Bernoulli click model. Figure 1a shows that VAC^2-UCB outperforms C^3-UCB and CascadeWOFUL, incurring 45% and 25% less regret after $100,000$ rounds. To model platforms like Netflix that recommend movies in specific categories, we also run experiments while restricting the candidate items to movies of a particular genre. Figure 1b shows that VAC^2-UCB is superior for all genres compared to C^3-UCB and CascadeWOFUL.

F. Concentration Bounds, Facts, and Technical Lemmas

In this section, we first give key concentration bounds and then provide lemmas that are useful for the analysis.

F.1. Concentration Bounds

We mainly use the following concentration bounds, which is essentially a modification of the Freedman's version of the Bernstein's inequality (Bernstein, 1946; Freedman, 1975).

Proposition 3 (Theorem 9, Lattimore et al. (2015))**.** *Let $\delta \in (0,1)$ and $X_1, ..., X_n$ be a sequence of random variables adapted to filtration $\{\mathcal{F}_t\}$ with $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$. Let $Z \subseteq [n]$ be such that $\mathbb{I}\{t \in Z\}$ is $\mathcal{F}_{t-1}$-measurable and let R_t be $\mathcal{F}_{t-1}$ measurable such that $|X_t| \leq R_t$ almost surely. Let $V = \sum_{t \in Z} \mathrm{Var}[X_t \mid \mathcal{F}_{t-1}] + \sum_{t \notin Z} R_t^2/2$, $R = \max_{t \in Z} R_t$, and $S = \sum_{t=1}^n X_t$. Then $\Pr[S \geq f(R,V)] \leq \delta$, where $f(r,v) = \frac{2(r+1)}{3} \log \frac{2}{\delta_{r,v}} + \sqrt{2(v+1) \log \frac{2}{\delta_{r,v}}}$, and $\delta_{r,v} = \frac{\delta}{3(r+1)^2(v+1)}$.*

F.2. Facts

Fact 1. *For any positive-definite matrices $\boldsymbol{A}, \boldsymbol{B} > \boldsymbol{0}^d$ and any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. It holds that*

1. *If $\boldsymbol{A} \leq \boldsymbol{B}$, then $\boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$.*
2. *If $\boldsymbol{A} \leq \boldsymbol{B}$, then $\|\boldsymbol{x}\|_{\boldsymbol{A}} \leq \|\boldsymbol{x}\|_{\boldsymbol{B}}$.*
3. *Suppose $\boldsymbol{A}$ has maximum eigenvalue $\lambda_{\max}$, then $\|\boldsymbol{A}\boldsymbol{x}\|_2 \leq \lambda_{\max} \cdot \|\boldsymbol{x}\|_2$ and $\lambda_{\max} \leq \mathrm{trace}(\boldsymbol{A})$.*

F.3. Technical Lemmas

Recall that event F_t is defined in Equation (21), gram matrix $\boldsymbol{G}_t$ is defined in Equation (18), optimistic variance $\bar{V}_{t,i}$ is defined in Equation (6), $R_{\boldsymbol{v}}$ is defined in Equation (132).

Lemma 7. $\Pr\left[\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t} + \sqrt{\gamma} \geq \rho \text{ and } \neg F_{t-1}\right] \leq \delta/T$, *for* $t = 1, ..., T$.

Proof of lemma 7. Let $\boldsymbol{v} \in \mathbb{R}^d$ and define

$$V_{s,i,\boldsymbol{v}} = \begin{cases} \mathrm{Var}[\eta_{s,i} \mid \mathcal{F}_{s-1}]\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle^2/\bar{V}_{s,i}^2, & \text{if } \bar{V}_{s,i} < \frac{1}{4} \\ \langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle^2/\bar{V}_{s,i}, & \text{otherwise.} \end{cases} \tag{131}$$

$$R_{\boldsymbol{v}} = \max_{s<t, i\in\tau_s} \{\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle/\bar{V}_{s,i} : \bar{V}_{s,i} < \frac{1}{4}\} \tag{132}$$

By applying the Proposition 3, with probability at least $1-\delta/T$ it holds that

$$\langle Z_t, \boldsymbol{v}\rangle = \sum_{s<t}\sum_{i\in\tau_s} \eta_{s,i}\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle/\bar{V}_{s,i} \le \frac{2(R_{\boldsymbol{v}}+1)}{3}\log\frac{1}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s} V_{s,i,\boldsymbol{v}})\log\frac{1}{\delta_{\boldsymbol{v}}}} \tag{133}$$

where $\delta_{\boldsymbol{v}} = \frac{3\delta}{T(1+R_{\boldsymbol{v}})^2(1+\sum_{s<t}\sum_{i\in\tau_s} V_{s,i,\boldsymbol{v}})^2}$.

Since $\boldsymbol{v}$ could be a random variable in later proofs, we use the covering argument trick (Chap.20, Lattimore & Szepesvári (2020)) to handle $\boldsymbol{v}$. Specifically, we define the covering set $\Lambda = \{j\cdot\varepsilon : j = -\frac{C}{\varepsilon}, -\frac{C}{\varepsilon}+1, ..., \frac{C}{\varepsilon}-1, \frac{C}{\varepsilon}\}^d$, with size $N = |\Lambda| = (2C/\varepsilon)^d$ and parameters C, ε will be determined shortly after. By applying union bound on Equation (133), we have with probability at least $1-\delta$ that

$$\langle Z_t, \boldsymbol{v}\rangle \le \frac{2(R_{\boldsymbol{v}}+1)}{3}\log\frac{N}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s} V_{s,i,\boldsymbol{v}})\log\frac{N}{\delta_{\boldsymbol{v}}}} \text{ for all } \boldsymbol{v}\in\Lambda. \tag{134}$$

Now we can set $\boldsymbol{v} = \boldsymbol{G}_t^{-1}\boldsymbol{Z}_t$, and it follows from Lemma 12 that $\|\boldsymbol{v}\|_\infty \le \|\boldsymbol{Z}_t\|_1 \le 2dK^2t^2 = C$, where the last inequality follows from Lemma 13. Based on our construction of the covering set Λ, there exists $\boldsymbol{v}'\in\Lambda$ with $\boldsymbol{v}'\le\boldsymbol{v}$, and $\|\boldsymbol{v}'-\boldsymbol{v}\|_\infty \le \varepsilon$, such that

$$\|\boldsymbol{Z}_t\|^2_{\boldsymbol{G}_t^{-1}} = \langle \boldsymbol{Z}_t, \boldsymbol{v}\rangle \le \|\boldsymbol{Z}_t\|_1\varepsilon + \langle \boldsymbol{Z}_t, \boldsymbol{v}'\rangle \tag{135}$$

$$\le \|\boldsymbol{Z}_t\|_1\varepsilon + \frac{2(R_{\boldsymbol{v}}+1)}{3}\log\frac{N}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s} V_{s,i,\boldsymbol{v}})\log\frac{N}{\delta_{\boldsymbol{v}}}} \tag{136}$$

$$\le \|\boldsymbol{Z}_t\|_1\varepsilon + \frac{2(R_{\boldsymbol{v}}+1)}{3}\log\frac{N}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\|\boldsymbol{Z}_t\|^2_{\boldsymbol{G}_t^{-1}})\log\frac{N}{\delta_{\boldsymbol{v}}}} \tag{137}$$

where Equation (136) uses the fact that $R_{\boldsymbol{v}'} \le R_{\boldsymbol{v}}, V_{s,i,\boldsymbol{v}'} \le V_{s,i,\boldsymbol{v}}, \frac{1}{\delta_{\boldsymbol{v}'}} \le \frac{1}{\delta_{\boldsymbol{v}'}}$ for any $\boldsymbol{v}'\le\boldsymbol{v}$, Equation (137) follows from the following derivation,

$$\sum_{s<t}\sum_{i\in\tau_s} V_{s,i,\boldsymbol{v}} \le \sum_{s<t}\sum_{i\in\tau_s}\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle^2/\bar{V}_{s,i} \tag{138}$$

$$= \sum_{s<t}\sum_{i\in\tau_s}(\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t)^\top\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t/\bar{V}_{s,i} \tag{139}$$

$$= (\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t)^\top\left(\sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top/\bar{V}_{s,i}\right)\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t \tag{140}$$

$$\le (\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t)^\top\boldsymbol{G}_t(\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t) \tag{141}$$

$$= \|\boldsymbol{Z}_t\|^2_{\boldsymbol{G}_t}, \tag{142}$$

where Equation (138) follows from $\neg F_{s-1}$ implies $\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_s\right\|_{\boldsymbol{G}_s} \le \rho$ for $s<t$ by Lemma 8 and thus $\bar{V}_{t,i} \ge \mathrm{Var}[\eta_{s,i} \mid \mathcal{F}_{s-1}]$, Equation (139) follows from definition of $\boldsymbol{v}$, Equation (141) follows from $\sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top/\bar{V}_{s,i} < \boldsymbol{G}_t$.

Now we set $\varepsilon = 1/C = 1/(2K^2t^2d)$, we have

$$\|\boldsymbol{Z}_t\|^2_{\boldsymbol{G}_t^{-1}} \le \text{RHS of Equation (137)} \tag{143}$$

$$\leq C\varepsilon + \frac{2(2\left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}}/\rho+1)}{3}\log\frac{N}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}}^2)\log\frac{N}{\delta_{\boldsymbol{v}}}} \tag{144}$$

$$\leq 1 + 2\log\frac{N}{\delta_{\boldsymbol{v}}} + \sqrt{2(1+\left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}}^2)\log\frac{N}{\delta_{\boldsymbol{v}}}} \tag{145}$$

where Equation (144) is to bound $R_{\boldsymbol{v}}$ by Lemma 10, Equation (145) is by the definition of ρ as an upper bound.

By rearranging and simplifying Equation (145), we have

$$\left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}} + \sqrt{\gamma} \leq 1 + \sqrt{\gamma} + 4\sqrt{\log\frac{N}{\delta_{\boldsymbol{v}}}} \tag{146}$$

$$\leq 1 + \sqrt{\gamma} + 4\sqrt{\log\left(\frac{6TN}{\delta}(1+\left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}}^2)\right)}, \tag{147}$$

where the last inequality is because of $\delta_{\boldsymbol{v}} \leq \frac{3\delta}{T(1+\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t^{-1}}^2)}$ from the definition of $\delta_{\boldsymbol{v}}$, Lemma 10, and Equation (141).

Finally, we solve the above equation and set $\rho = 1 + \sqrt{\gamma} + 4\sqrt{\log\left(\frac{6TN}{\delta}\log(\frac{3TN}{\delta})\right)}$, which completes the reduction on t to show the probability $\Pr[\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t^{-1}} + \sqrt{\gamma} \geq \rho] \geq 1 - \delta/T$ under event $\neg F_{t-1}$. ■

Lemma 8. *If $\neg F_t$ holds, then it holds that,*

$$\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{G}_t} \leq \rho. \tag{148}$$

Proof. We have

$$\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{G}_t} = \left\|\boldsymbol{G}_t^{-1}(\sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)X_{s,i}/\bar{V}_{s,i}) - \boldsymbol{G}_t^{-1}\boldsymbol{G}_t\boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t} \tag{149}$$

$$= \left\|\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t + \boldsymbol{G}_t^{-1}(\sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top\boldsymbol{\theta}^*/\bar{V}_{s,i}) - \boldsymbol{G}_t^{-1}\boldsymbol{G}_t\boldsymbol{\theta}^*)\right\|_{\boldsymbol{G}_t} \tag{150}$$

$$= \left\|\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t - \gamma\boldsymbol{G}_t^{-1}\boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t} \tag{151}$$

$$\leq \left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}} + \gamma\left\|\boldsymbol{\theta}^*\right\|_{\boldsymbol{G}_t^{-1}} \tag{152}$$

$$\leq \left\|\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t^{-1}} + \sqrt{\gamma} \tag{153}$$

$$\leq \rho - \sqrt{\gamma} + \sqrt{\gamma} = \rho, \tag{154}$$

where Equation (149)-(151) follow from definition and math calculation, Equation (152) from $\boldsymbol{G}_t \geq \boldsymbol{G}_0 = \gamma\boldsymbol{I}$ and $\|\boldsymbol{\theta}\|_2 \leq 1$, Equation (153) from that if $\neg F_t$ holds, then $\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t} + \sqrt{\gamma} \leq \rho$. ■

Lemma 9. *For any $s < t$, $\frac{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}}}{\bar{V}_{s,i}} \leq \frac{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}}{\bar{V}_{s,i}}$, and if $\neg F_{t-1}$ holds and $\bar{V}_{s,i} < \frac{1}{4}$, $\frac{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}}{\bar{V}_{s,i}} \leq \frac{2}{\rho} \leq 1$ for any $i \in [m]$.*

Proof. The first inequality is by $\boldsymbol{G}_t \geq \boldsymbol{G}_s$ and Fact 1. For the second inequality, when $\neg F_{t-1}$ holds, $\left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_s\right\|_{\boldsymbol{G}_s} \leq \rho$ as in Equation (148), and since $\bar{V}_{s,i} < \frac{1}{4}$, it follows from the definition of $\bar{V}_{s,i}$ Equation (6) that at least one of the following is true:

$$\bar{V}_{s,i} \geq \frac{1}{2}(\langle\boldsymbol{\phi}_s(i), \hat{\boldsymbol{\theta}}_s + 2\rho\left\|\boldsymbol{\phi}_s(i)\right\|_{\boldsymbol{G}_s^{-1}}\rangle) \geq \rho\left\|\boldsymbol{\phi}_s(i)\right\|_{\boldsymbol{G}_s^{-1}}/2, \tag{155}$$

$$\bar{V}_{s,i} \geq \frac{1}{2}(1 - \langle\boldsymbol{\phi}_s(i), \hat{\boldsymbol{\theta}}_s + 2\rho\left\|\boldsymbol{\phi}_s(i)\right\|_{\boldsymbol{G}_s^{-1}}\rangle) \geq \rho\left\|\boldsymbol{\phi}_s(i)\right\|_{\boldsymbol{G}_s^{-1}}/2, \tag{156}$$

which concludes the second inequality. ■

Lemma 10. *Let* $\boldsymbol{v} = \boldsymbol{G}_t^{-1}\boldsymbol{b}_t$*, if* $\neg F_{t-1}$*, then* $R_{\boldsymbol{v}} \leq \frac{2\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t^{-1}}}{\rho}$.

Proof. For all $s < t$ and $i \in [m]$, we have $\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle / \bar{V}_{s,i} \leq \frac{2\langle \boldsymbol{\phi}_s(i), \boldsymbol{v}\rangle}{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}\cdot\rho}} = \frac{2\langle \boldsymbol{\phi}_s(i), \boldsymbol{G}_t^{-1}\boldsymbol{Z}_t\rangle}{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}\cdot\rho}} \leq \frac{2\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}}\cdot\left\|\boldsymbol{G}_t^{-1}\boldsymbol{Z}_t\right\|_{\boldsymbol{G}_t}}{\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}\cdot\rho}} = \frac{2\|\boldsymbol{Z}_t\|_{\boldsymbol{G}_t^{-1}}}{\rho}$, where the first inequality follows from Lemma 9, the last inequality follows from the Cauchy-Schwarz inequality. ■

Lemma 11. *If* $\neg F_t$*, then* $\|\boldsymbol{\phi}_t(i)\|_2^2 / \bar{V}_{t,i} \leq 4dKt$.

Proof. If $\bar{V}_{t,i} = \frac{1}{4}$, the inequality trivially holds since $\|\boldsymbol{\phi}_t(i)\| \leq 1$. Consider $\bar{V}_{t,i} < \frac{1}{4}$, and $\lambda_{\max}$ be the maximum eigenvalue of $\boldsymbol{G}_t$. Then, it holds that $\|\boldsymbol{\phi}_t(i)\|_2^2 / \bar{V}_{t,i} \leq \frac{\|\boldsymbol{\phi}_t(i)\|_2^2}{\rho\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} \leq \frac{\|\boldsymbol{\phi}_t(i)\|_2}{\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} = \frac{\left\|\boldsymbol{G}_t^{1/2}\boldsymbol{G}_t^{-1/2}\boldsymbol{\phi}_t(i)\right\|_2}{\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} \leq \sqrt{\lambda_{\max}}$, where the first inequality follows from Lemma 9, the second inequality is by $\rho \geq 1$, $\|\boldsymbol{\phi}_t(i)\| \leq 1$, and the last is by **Fact 1**.3.

Now Assume $\|\boldsymbol{\phi}_s(i)\|_2^2 / \bar{V}_{s,i} \leq 4s$ for $s < t$, which always holds for $t = 1$. By reduction, we consider round t, it holds that $\|\boldsymbol{\phi}_t(i)\|_2^2 / \bar{V}_{t,i} \leq \sqrt{\lambda_{\max}} \leq \sqrt{\mathrm{trace}(\boldsymbol{G}_t)} = \sqrt{\gamma d + \sum_{s=1}^{t-1}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_2^2/\bar{V}_{s,i}} \leq \sqrt{Kd + \sum_{s=1}^{t-1} 4dK^2 s} \leq \sqrt{d(K + 2K^2 t(t-1))} \leq 4dKt$, where the first inequality follows from the analysis in the last paragraph, the third inequality follows from reduction over $s < t$, and the last inequality is by math calculation. ■

Lemma 12. *If* $\neg F_t$*, then* $\|\boldsymbol{\phi}_t(i)\|_1 / \bar{V}_{t,i} \leq 4dKt$.

Proof. Similar to the proof of Lemma 11, $\|\boldsymbol{\phi}_t(i)\|_1 / \bar{V}_{t,i} \leq \sqrt{d}\,\|\boldsymbol{\phi}_t(i)\|_2 / \bar{V}_{t,i} \leq \frac{\sqrt{d}\|\boldsymbol{\phi}_t(i)\|_2}{\rho\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} \leq \frac{\|\boldsymbol{\phi}_t(i)\|_2}{\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} = \frac{\left\|\boldsymbol{G}_t^{1/2}\boldsymbol{G}_t^{-1/2}\boldsymbol{\phi}_t(i)\right\|_2}{\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}} \leq \sqrt{\lambda_{\max}} \leq 4dKt$, where the first inequality uses Cauchy-Schwarz, the second inequality uses $\rho \geq \sqrt{d}$, and the rest follows from the proof of Lemma 11. ■

Lemma 13. *If* $\neg F_{t-1}$*, then* $\|\boldsymbol{Z}_t\|_1 \leq 2dK^2t^2$.

Proof. $\|\boldsymbol{Z}_t\|_1 = \left\|\sum_{s<t}\sum_{i\in\tau_s}\eta_{s,i}\boldsymbol{\phi}_s(i)/\bar{V}_{s,i}\right\|_1 \leq \sum_{s<t}\sum_{i\in\tau_s}\left\|\boldsymbol{\phi}_s(i)/\bar{V}_{s,i}\right\|_1 \leq \sum_{s<t}\sum_{i\in\tau_s} 4dKt \leq 2dK^2t^2$, where the first inequality follows from $\eta_{s,i} \in [-1, 1]$, the second inequality follows from Lemma 12. ■

Lemma 14 (Lemma A.3, (Li et al., 2016))**.** *Let* $\boldsymbol{x}_i \in \mathbb{R}^d$*,* $1 \leq i \leq n$*. Then we have*

$$\det\left(\boldsymbol{I} + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\right) \geq 1 + \sum_{i=1}^{n}\|\boldsymbol{x}\|_2^2.$$

Lemma 15 (Lemma 11, (Abbasi-Yadkori et al., 2011))**.** *Let* $\boldsymbol{x}_i \in \mathbb{R}^d$ *with* $\|\boldsymbol{x}_i\|_2^2 \leq L$*,* $1 \leq i \leq n$ *and let* $\boldsymbol{G}_t = \gamma\boldsymbol{I} + \sum_{i=1}^{t-1}\boldsymbol{x}_i\boldsymbol{x}_i^\top$*, then*

$$\det(\boldsymbol{G}_{t+1}) \leq (\gamma + tL^2/d)^d.$$

Lemma 16. *Equation* (82) *holds.*

Proof. **When** $T_{t-1,i} > L_{i,T,2} = \frac{8c_1^2B_v^2K\log T}{(\tilde{\Delta}_i^{\min})^2}$**,**

we have $(82, i) \leq \frac{8c_1^2B_v^2\log T}{T_{t-1,i,}\cdot\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K} < \frac{(\tilde{\Delta}_i^{\min})^2}{K\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K} \leq 0 = \kappa_{i,T}(T_{t-1,i})$.

When $L_{i,T,1} < T_{t-1,i} \leq L_{i,T,2}$**,**

We have $(82, i) \leq \frac{8c_1^2B_v^2\log T}{T_{t-1,i,}\cdot\tilde{\Delta}_{S_t}} - \frac{\tilde{\Delta}_{S_t}}{K} < \frac{8c_1^2B_v^2\log T}{T_{t-1,i,}\cdot\tilde{\Delta}_{S_t}} \leq \frac{8c_1^2B_v^2\log T}{T_{t-1,i}\cdot\tilde{\Delta}_{\min}^i} = \kappa_{i,T}(T_{t-1,i,j_i^{S_t}})$.

When $T_{t-1,i} \leq L_{i,T,1}$,

We further consider two different cases $T_{t-1,i} \leq \frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t})^2}$ or $\frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t})^2} < T_{t-1,i} \leq L_{i,T,1} = \frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_i^{\min})^2}$.

For the former case, if there exists $i \in \tau_t$ so that $T_{t-1,i} \leq \frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t})^2}$, then we know $\sum_{q \in \tilde{S}_t} \kappa_{q,T}(T_{t-1,q}) \geq \kappa_{i,T}(T_{t-1,i}) = 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} \geq 2\tilde{\Delta}_{S_t} > \Delta_{S_t}$, which makes eq. (82) holds no matter what. This means we do not need to consider this case for good.

For the later case, when $\frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t})^2} < T_{t-1,i}$, we know that $(82, i) \leq \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t}} \frac{1}{T_{t-1,i}} = 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t})^2} \frac{1}{T_{t-1,i}}} \sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} \leq 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} = \kappa_{i,T}(T_{t-1,i})$.

When $\ell = 0$,

We have $(82, i) \leq \frac{8c_1^2 B_v^2}{\tilde{\Delta}_{S_t}} \cdot \frac{1}{28} - \frac{\tilde{\Delta}_{S_t}}{K} \leq \frac{c_1^2 B_v^2}{\tilde{\Delta}_{S_t}} \leq \frac{c_1^2 B_v^2}{\tilde{\Delta}_i^{\min}} = \kappa_{i,T}(T_{t-1,i})$.

Combining all above cases, we have $\Delta_{S_t} \leq \mathbb{E}[\sum_{i \in \tau_t} \kappa_{i,T}(T_{t-1,i})]$. ■

Lemma 17. *Equation* (98) *holds.*

Proof. **When $T_{t-1,i} > L_{i,T,2} = \frac{8c_1^2 B_v^2 K \log T}{\tilde{\Delta}_i^{\min} \cdot \tilde{\Delta}_{i,\lambda}^{\min}}$,**

we have $(98, i) \leq \frac{8c_1^2 B_v^2 \log T}{T_{t-1,i,} \cdot \tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} < \frac{\tilde{\Delta}_i^{\min} \cdot \tilde{\Delta}_{i,\lambda}^{\min}}{K \tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} \leq 0 = \kappa_{i,T}(T_{t-1,i})$.

When $L_{i,T,1} < T_{t-1,i} \leq L_{i,T,2}$,

We have $(98, i) \leq \frac{8c_1^2 B_v^2 \log T}{T_{t-1,i} \cdot \tilde{\Delta}_{S_t,\lambda}} - \frac{\tilde{\Delta}_{S_t}}{K} < \frac{8c_1^2 B_v^2 \log T}{T_{t-1,i,} \cdot \tilde{\Delta}_{S_t,\lambda}} \leq \frac{8c_1^2 B_v^2 \log T}{T_{t-1,i} \cdot \tilde{\Delta}_{\min}^{i,\lambda}} = \kappa_{i,T}(T_{t-1,i})$.

When $T_{t-1,i} \leq L_{i,T,1}$,

We further consider two different cases $T_{t-1,i} \leq \frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t,\lambda} \cdot \tilde{\Delta}_{S_t}}$ or $\frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t,\lambda} \cdot \tilde{\Delta}_{S_t}} < T_{t-1,i} \leq L_{i,T,1} = \frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_{i,\lambda}^{\min} \cdot \tilde{\Delta}_i^{\min}}$.

For the former case, if there exists $i \in \tau_t$ so that $T_{t-1,i} \leq \frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t,\lambda} \cdot \tilde{\Delta}_{S_t}}$, then we know $\sum_{q \in \tilde{S}_t} \kappa_{q,T}(T_{t-1,q}) \geq \kappa_{i,T}(T_{t-1,i}) = 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} \geq 2\sqrt{\tilde{\Delta}_{S_t,\lambda} \cdot \tilde{\Delta}_{S_t}} \geq \Delta_{S_t}$, which makes eq. (98) holds no matter what. This means we do not need to consider this case for good.

For the later case, when $\frac{4c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t,\lambda} \cdot \tilde{\Delta}_{S_t}} < T_{t-1,i}$, we know that $(98, i) \leq \frac{8c_1^2 B_v^2 \log T}{\tilde{\Delta}_{S_t,\lambda}} \frac{1}{T_{t-1,i}} = 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{(\tilde{\Delta}_{S_t,\lambda})^2} \frac{1}{T_{t-1,i}}} \sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} \leq 2\sqrt{\frac{\tilde{\Delta}_{S_t}}{\tilde{\Delta}_{S_t,\lambda}}} \sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} \leq 2\sqrt{\frac{4c_1^2 B_v^2 \log T}{T_{t-1,i}}} = \kappa_{i,T}(T_{t-1,i})$.

When $\ell = 0$,

We have $(98, i) \leq \frac{8c_1^2 B_v^2}{\tilde{\Delta}_{S_t,\lambda}} \cdot \frac{1}{28} - \frac{\tilde{\Delta}_{S_t}}{K} \leq \frac{c_1^2 B_v^2}{\tilde{\Delta}_{S_t,\lambda}} \leq \frac{c_1^2 B_v^2}{\tilde{\Delta}_{i,\lambda}^{\min}} = \kappa_{i,T}(T_{t-1,i})$.

Combining all above cases, we have $\Delta_{S_t} \leq \mathbb{E}[\sum_{i \in \tau_t} \kappa_{i,T}(T_{t-1,i})]$. ■

Lemma 18. *Equation* (109) *holds.*

Proof. **When $T_{t-1,i} > L_{i,T,2} = \frac{4c_2 B_1 K \log T}{\Delta_i^{\min}}$,**

we have $(109, i) \leq 4c_2 B_1 \frac{\log T}{T_{t-1,i}} - \frac{\Delta_{S_t}}{K} < \frac{\Delta_i^{\min}}{K} - \frac{\Delta_{S_t}}{K} \leq 0 = \kappa_{i,T}(T_{t-1,i})$.

When $T_{t-1,i} \leq L_{i,T,2}$,

We have $(109, i) \leq 4c_2 B_1 \frac{\log T}{T_{t-1,i}} - \frac{\Delta_{S_t}}{K} < \frac{4c_2 B_1 \log T}{T_{t-1,i}} = \kappa_{i,j_i^{S_t},T}(N_{t-1,i,j_i^{S_t}})$.

When $T_{t-1,i} \leq L_{i,T,1}$,

If there exists $i \in \tilde{S}_t$ so that $T_{t-1,i} \leq L_{i,T,1}$, then we know $\sum_{q \in \tilde{S}_t} \kappa_{i,T}(T_{t-1,q}) \geq \kappa_{i,T}(T_{t-1,i}) = \Delta_i^{\max} \geq \Delta_{S_t}$, which makes eq. (109) holds no matter what. This means we do not need to consider this case for good.

Combining all above cases, we have $\Delta_{S_t} \leq \mathbb{E}_t[\sum_{i \in \tau_t} \kappa_{i,T}(T_{t-1,i})]$. ■