IEEE INFOCOM 2024 - IEEE Conference on Computer Communications | 979-8-3503-8350-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/INFOCOM52122.2024.10621257

Learning Context-Aware Probabilistic Maximum Coverage Bandits: A Variance-Adaptive Approach

Xutong Liu[1], Jinhang Zuo[2,3], Junkai Wang[4], Zhiyong Wang[1], Yuedong Xu*[4], John C.S. Lui[1]
[1]The Chinese University of Hong Kong, [2]University of Massachusetts Amherst,
[3]California Institute of Technology, [4] Fudan University
Email: {liuxt, zywang21, cslui}@cse.cuhk.edu.hk, jhzuo@caltech.edu, {jkwang18,ydxu}@fudan.edu.cn

***Abstract*—Probabilistic maximum coverage (PMC) is an important framework that can model many network applications, including mobile crowdsensing, content delivery, and task replication. In PMC, an operator chooses nodes in a graph that can probabilistically cover other nodes, aiming to maximize the total rewards from the covered nodes. To tackle the challenge of unknown parameters in network environments, PMC are studied under the online learning context, i.e., the PMC bandit. However, existing PMC bandits lack context-awareness and fail to exploit valuable contextual information, limiting their efficiency and adaptability in dynamic environments. To address this limitation, we propose a novel context-aware PMC bandit model (C-PMC). C-PMC employs a linear structure to model the mean outcome of each arm, effectively incorporating contextual information and enhancing its applicability to large-scale network systems. Then we design a variance-adaptive contextual combinatorial upper confidence bound algorithm (VAC^{2}UCB), which utilizes second-order statistics, specifically variance, to re-weight feedback data and estimate unknown parameters. Our theoretical analysis shows that C-PMC achieves a regret of $\tilde{O}(d\sqrt{|\mathcal{V}|T})$, independent of the number of edges $|\mathcal{E}|$ and action size K. Finally, we conduct experiments on synthetic and real-world datasets, showing the superior performance of VAC^{2}UCB in context-aware mobile crowdsensing and user-targeted content delivery applications.**

I. Introduction

The probabilistic maximum coverage (PMC) problem is a simple yet powerful model which can be applied to many network scenarios, including network content delivery [1], [2], mobile crowdsensing [3]–[5], and vehicular network task replication [6], [7]. PMC is represented by a bipartite graph $\mathcal{G}(\mathcal{U}, \mathcal{V}, \mathcal{E})$ as input, where $\mathcal{U}$ are the nodes to be selected, $\mathcal{V}$ are the nodes to be covered, and $\mathcal{E}$ are the edges between $\mathcal{U}$ and $\mathcal{V}$. Each edge (u, v) in $\mathcal{E}$ is associated with a probability $p(u, v)$, while each target node v in $\mathcal{V}$ is assigned a weight $w(v)$. The probability $p(u, v)$ indicates the likelihood that node u from $\mathcal{U}$ can cover a target node v in $\mathcal{V}$ independently, and the weight $w(v)$ corresponds to the reward contributed by a successfully covered node v. The objective of the decision maker is to select at most K nodes in $\mathcal{U}$ to maximize the cumulative rewards obtained from the covered nodes in $\mathcal{V}$.

In content delivery networks (CDN), as depicted in Fig. 1, the PMC problem can be employed to strategically choose a subset of servers to enhance user experience. In this scenario,

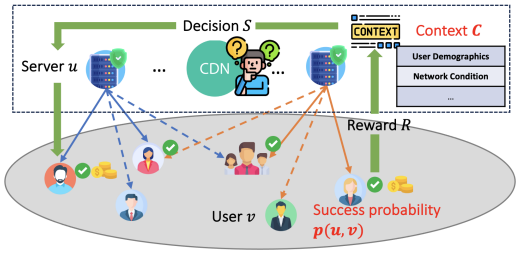


Fig. 1: C-PMC for content delivery: the decision maker chooses servers based on contextual information, successfully covers users (check marks) via edges (solid lines), and gains rewards.

servers can cache contents (e.g., pictures, videos), allowing end users to access them swiftly from the nearby servers [1]. The PMC model can cover this application by using $\mathcal{U}$ to denote the set of candidate servers responsible for content delivery and $\mathcal{V}$ to represent the set of users who consumes the contents. The probability $p(u, v)$ for each edge (u, v) captures the likelihood of successfully delivering content from server u to user v in a timely and high-quality manner, and the weight $w(v)$ represents the probability that user v ultimately consumes the content. The objective of the PMC problem in this context is to select some server nodes in $\mathcal{U}$, so as to maximize the number of users who successfully consume the content, thereby optimizing the user experience in the CDN.

For the PMC problem, it is crucial to set accurate parametric values (e.g., edge probabilities, node weights) to make optimal decisions. Previous studies have assumed that these parameters are known in advance [8]–[10]. However, in practical network applications, these parameters are *unknown* and must be estimated on the fly. This motivates the study of PMC in the online learning context, i.e., PMC bandit [2].

In PMC bandit, Chen et al. [2] associate each edge (u, v) and node v with an arm, where each arm has an unknown probability $p(u, v)$ or weight $w(v)$ to be learned in T consecutive decision rounds. In each round $t \in [T]$, the learning agent selects a set of arms as *action*, and the outcomes of these selected arms are observed as feedback. This feedback, known as *semi-bandit feedback*, allows the agents to gradually learn the unknown parameters and optimize their decisions to maximize their rewards. The agent's objective is to maximize the expected

*Yuedong Xu is the corresponding author. This work was supported in part by the Natural Science Foundation of China under Grant Grant 62072117, the Shanghai Natural Science Foundation under the Grant 22ZR1407000.

total rewards, or equivalently, minimize the expected *regret*, which quantifies the difference between the total expected rewards obtained by always playing the best action and playing according to the agent's own policy [11].

Recently, Liu et al. [12] propose a new variant of PMC bandit, called PMC-G bandit, which extends the deterministic semi-bandit feedback to general probabilistic feedback for a wider range of applications. For PMC-G bandit, a novel variance-adaptive algorithm, VACUCB, is proposed, which achieves a regret of $\tilde{O}(\sqrt{|\mathcal{E}||\mathcal{V}|T})$ regret (where $\tilde{O}$ hides logarithmic factors), improving upon the CUCB algorithm from Chen et al. [2] by a factor of $O(K)$.

Despite the success of PMC-G and VACUCB in handling general feedback and achieving improved regrets, there still exist limitations that can be significantly improved. Firstly, PMC-G does not utilize contextual information (e.g., geospatial information, network states, hardware parameters, user demographic, human activities, etc.), which are often present in network systems and can serve as valuable features to quantify network performances and user behaviors [13]. Without leveraging the contextual information, PMC-G needs to independently learn all $O(|\mathcal{E}|)$ success probabilities $p(u, v)$ (and weights $w(v)$), leading to regret that grows with an unsatisfying $O(\sqrt{|\mathcal{E}|})$ factor. This is highly inefficient and non-scalable for applications like CDN with thousands of candidate servers $\mathcal{U}$ and hundreds of thousands of users $\mathcal{V}$. Moreover, PMC-G cannot adapt to dynamic environments, where they assume $p(u, v)$ (and $w(v)$) are fixed for all T rounds. In realistic network applications, however, they may change over time. Take CDN as an example, different users $\mathcal{U}$ may arrive randomly at different times with varying preferences and locations, causing success probability $p(u, v)$ and weights $w(v)$ change dynamically. For the above setting that violates PMC-G's assumption, both its VACUCB algorithm and its performance guarantee will no longer be meaningful.

A. Our Contributions

To address the aforementioned limitations, this paper makes four key contributions as follows.

(1) Model Formulation: We propose a new variant of PMC bandit called the context-aware PMC bandit (C-PMC), which incorporates contextual information to enhance the PMC-G model. For each arm, we use a general time-varying feature map to leverage contextual information at each round t. We adopt a simple yet effective linear structure to model the mean outcome of each arm, where the mean is represented as a linear product of a d-dimensional feature and an unknown d-dimensional parameter. This formulation enables C-PMC to retain context awareness, scalability, and adaptability to time-varying environments, while benefiting from PMC-G's rich feedback models. We demonstrate the effectiveness of C-PMC by applying it to two representative network applications: context-aware mobile crowdsensing and user-targeted content delivery, each incorporating various contextual information.

(2) Algorithm Design: We propose a novel variance-adaptive contextual combinatorial upper confidence bound algorithm (VAC^{2}UCB) for C-PMC. Unlike traditional contextual bandit algorithms, VAC^{2}UCB leverages *second-order statistics*, specifically variance, as weights to re-weight each feedback data and construct a variance-adaptive least-squared estimator for learning the unknown parameters. The main technical difficulty is that the variance itself is unknown. To handle this challenge, we use an optimistic variance as a proxy for the true variance, which depends on our estimator and the uncertainty level of each arm. By combining these techniques, we achieve regret bounds that do not depend on the number of edges $|\mathcal{E}|$ and the action size K.

(3) Theoretical Analysis: We prove that our algorithm achieves a regret bound of $\tilde{O}(d\sqrt{|\mathcal{V}|T})$, matching the lower bound up to a factor of $\tilde{O}(\sqrt{d})$, where d is typically small in real applications (around 20). Importantly, our regret bound is independent of the number of edges $|\mathcal{E}|$ and the action size K, improving upon the state-of-the-art VACUCB and C^{3}UCB algorithms by factors of $\tilde{O}(\sqrt{|\mathcal{E}|/d})$ and $\tilde{O}(\sqrt{K})$, respectively. Our analysis tackles several technical challenges, such as bounding the uncertainty of our estimator in the face of unknown variance with time-varying contextual information, and bounding the total regret which is highly nonlinear with respect to estimation error and observation probability. The key contributions of our analysis are a tight concentration bound for our new estimator and a sensitivity lemma that relates the estimation error to the final regret, which may of independent interest for other related works.

(4) Performance Evaluation: We conduct comprehensive experiments on two representative applications mentioned earlier, using synthetic and real-world datasets, to validate our theoretical results. The experimental results demonstrate that our proposed algorithms achieve at least 11% and 77% less regrets compared to benchmark algorithms.

B. Related Works

The field of online learning problems under the multi-armed bandit (MAB) model has been extensively studied. The MAB model was first introduced by the seminal work [14] and has been expanded upon by many other researchers (cf. [11], [15], [16]). Linear contextual bandit is a notable extension that incorporates contextual information and assumes an effective linear structure to improve scalability [17]–[19]. Our work is inspired by the consideration of contextual information in these studies, however, C-PMC allows a combination of arms to be pulled in each round, which requires special treatment to handle the combinatorial explosion of the exploration space. To handle the combinatorial structure, contextual combinatorial MAB is proposed by Qin et al. [20] under semi-bandit feedback, then studied by Li et al. [21] under cascading feedback. However, these two works only achieve sub-optimal regret with an additional $O(K)$ factor, as they lack variance-adaptive algorithms and treat each feedback data equally. In contrast, our work leverages variance-adaptive algorithms, leading to superior performance in both theory and experiments, as shown in Section IV and Section V.

The Probabilistic Maximum Coverage (PMC) problem is first presented in [8], and finds applications in various computer science domains, particularly in the field of network optimization. Alongside the three applications discussed in this paper, PMC has relevance in wireless sensor placement [22] and social network advertising [23], [24]. The online learning variant of the PMC problem, known as PMC bandit, is initially proposed by Chen et al. [25] and has since been studied by Chen et al. [2], Merlis et al. [26], and other researchers. Recently, Liu et al. [12] proposes a PMC-G model, which is the closest to our work. PMC-G generalizes PMC with semi-bandit feedback for applications involving general probabilistic feedback. In contrast to non-contextual PMC-G, our work is context-aware, more scalable, and more adaptive to time-varying environments by leveraging the valuable contextual information.

II. System Model

In this paper, we use $[n]$ to represent set $\{1, ..., n\}$ for $n \in \mathbb{Z}^+$. 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 $\boldsymbol{x}$. We use $\boldsymbol{e}_i$ to denote the one-hot vector with 1 at the i-th entry and 0 elsewhere. For any event $\mathcal{E}$, we use $\mathbb{I}\{\mathcal{E}\}$ to denote the indicator function, where $\mathbb{I}\{\mathcal{E}\} = 1$ if $\mathcal{E}$ holds and $\mathbb{I}\{\mathcal{E}\} = 0$ otherwise. For any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we use $\boldsymbol{x} \odot \boldsymbol{y} \in \mathbb{R}^d$ to denote element wise product $(x_i y_i)_{i\in[d]}$. 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 $\boldsymbol{M}$.

The system model of context-aware PMC bandit (C-PMC), can be represented by a tuple $(\mathcal{G}, [m], \mathcal{S}, \mathcal{C}, \Phi, \Theta, D_{\text{obs}}, R)$ as follows. $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ denotes the underlying bipartite graph, where $\mathcal{U}$ is the set of candidate nodes, $\mathcal{V}$ is the target nodes to be covered by $\mathcal{U}$, and $\mathcal{E}$ is the set of edges connecting $\mathcal{U}$ and $\mathcal{V}$; $[m] = \{1, 2, ..., m\}$ denotes the set of base arms (or arms), where each base arm is associated with an unknown parameter that needs to be learned. Depending on different application scenarios in Section V, the base arms for C-PMC could refer to the edge set $\mathcal{E}$, or the edge and target node sets $\mathcal{E} \cup \mathcal{V}$. Thus, we use $[m]$ to cover both cases; $\mathcal{S}$ represents the set of eligible actions, where $S \in \mathcal{S}$ denotes an action. Similar to $[m]$, $\mathcal{S}$ is application-dependent and can be either a collection of subsets of $[m]$, or subsets of $\mathcal{U}$; $\mathcal{C}$ denotes the set of possible contexts; Φ denotes the set of possible feature maps, where any feature map $\boldsymbol{\phi} \in \Phi$ is a function $\mathcal{C} \times [m] \to \mathbb{R}^d$, and $\boldsymbol{\phi}(c, i)$ maps an arm i to a d-dimensional feature vector given context c. Here, we assume feature vectors are normalized, such that $\|\boldsymbol{\phi}(c, i)\|_2 \leq 1$. D_{obs} is the observation function used to model the general feedback, similar to that of [12], and R is the reward function, whose definition will be provided shortly.

In C-PMC, a learning game is played between a learning agent and the unknown environment in a sequential manner, according to the following procedure. Before the game starts, the environment chooses a parameter $\boldsymbol{\theta}^* \in \Theta$ unknown to the agent (without loss of generality, we assume $\|\boldsymbol{\theta}^*\|_2 \leq 1$). At the beginning of round $t = 1, 2, ..., T$, the agent receives a context $c_t \in \mathcal{C}$ and a feature map $\boldsymbol{\phi}_t \in \Phi$. Note that $c_t, \boldsymbol{\phi}_t$ could be time-varying, or stochastically generated. Then for each arm i, it is associated with a feature vector $\boldsymbol{\phi}_t(c_t, i) \in \mathbb{R}^d$ which encodes the contextual information c_t to a d-dimensional vector.[1] Based on the feature vectors $\boldsymbol{\phi}_t(i)$, the agent selects an action $S_t \in \mathcal{S}$. Subsequently, the environment generates Bernoulli outcomes $\boldsymbol{X}_t = (X_{t,1}, ...X_{t,m}) \in \{0, 1\}^m$ for base arms, with mean outcome $\mathbb{E}[\boldsymbol{X}_{t,i}|\mathcal{H}_t] = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i) \rangle$. Here, $\mathcal{H}_t$ denotes the history before the agent's selection of action S_t, which will be defined shortly. Note that the outcomes $\boldsymbol{X}_t$ are assumed to be conditional independent across arms given history $\mathcal{H}_t$, consistent with prior works [20], [21], [27].

When the action S_t is taken, the agent will receive a non-negative reward $R(S_t, \boldsymbol{X}_t)$. For C-PMC, the reward at round t is the total rewards obtained from the covered nodes,

$$R(S_t, \boldsymbol{X}_t) = \textstyle\sum_{v\in\mathcal{V}} X_{t,v} \cdot \mathbb{I}\{\exists\, u \in S_t \text{ s.t. } X_{t,(u,v)} = 1\}. \quad (1)$$

Essentially, for an edge $(u, v) \in \mathcal{E}$, $X_{t,(u,v)} = 1$ indicates that the target node $v \in \mathcal{V}$ is covered when $u \in \mathcal{U}$ is selected, and for $v \in \mathcal{V}$, $X_{t,v} = 1$ means that covering the target node v yields one unit of reward. Let $\boldsymbol{\mu}_t \triangleq (\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t(i) \rangle)_{i\in[m]}$ denote the mean vector of base arms' outcomes at round t, which are unknown to the agent. Under the assumption of independence, the expected reward $r(S; \boldsymbol{\mu}_t) \triangleq \mathbb{E}[R(S, \boldsymbol{X}_t)]$ is

$$r(S; \boldsymbol{\mu}_t) = \textstyle\sum_{v\in\mathcal{V}} \mu_{t,v} \left(1 - \prod_{u\in S}(1 - \mu_{t,(u,v)})\right). \quad (2)$$

It is worth noting that this expected reward function is *highly non-linear* with respect to $\boldsymbol{\mu}_t$, and finding the optimal solution $S_t^* = \text{argmax}_{S\in\mathcal{S}}\, r(S; \boldsymbol{\mu}_t)$ is generally NP-hard [8]. Fortunately, ultilizing submodular set function maximization technique, one can achieve $\alpha = (1 - 1/e)$-approximate solutions [8].

At the end of round t, the agent can observe some of the arm outcomes as feedback, which is critical for improving future decisions. Consistent with [12], we assume that the outcomes of base arms in a random set $\tau_t \sim D_{\text{obs}}(S_t, \boldsymbol{X}_t)$ are observed, meaning that the outcomes of arms in τ_t, i.e. $(X_t)_{t\in\tau_t}$ are revealed as the feedback to the agent. Here the function D_{obs} is used to model the general feedback and is referred as the *general feedback function*. For notational convenience, we define *observation probability* $p_i^{\boldsymbol{\mu}, D_{\text{obs}}, S}$ as the probability that base arm i is observed when the action is S, the mean vector is $\boldsymbol{\mu}$, and the feedback function is D_{obs}. Since D_{obs} is fixed in a given application, we ignore it in the notation for simplicity, and use $p_i^{\boldsymbol{\mu}, S}$ henceforth. To this end, we can give the formal definition of the history $\mathcal{H}_t = (c_s, \boldsymbol{\phi}_s, S_s, \tau_s, (X_{s,i})_{i\in\tau_s})_{s<t} \bigcup (c_t, \boldsymbol{\phi}_t)$, which encompasses all information prior to round t and includes the contextual information c_t and $\boldsymbol{\phi}_t$ at round t. For convenience, we define $\mathcal{M} \triangleq \{\langle \boldsymbol{\theta}, \boldsymbol{\phi}(c, i) \rangle_{i\in[m]} : c \in \mathcal{C}, \boldsymbol{\phi} \in \Phi, \boldsymbol{\theta} \in \Theta\}$ as the set of all possible mean vectors generated by $\mathcal{C}$, Φ and Θ.

It is important to emphasize that the introduction of D_{obs} enhances the modeling capabilities of previous PMC bandit [25]. It not only models deterministic semi-bandit feedback but

[1] For notational simplicity, we will use $\boldsymbol{\phi}_t(i)$ to denote (time-varying) feature vector $\boldsymbol{\phi}_t(c_t, i)$ at round t.

Algorithm 1 VAC^{2}UCB: Variance-Adaptive Contextual Combinatorial Upper Confidence Bound Algorithm for C-PMC

1: **Input:** Base arms $[m]$, dimension d, regularizer γ, failure probability δ, α-approximation oracle ORACLE.
2: **Initialize:** set gram matrix $\boldsymbol{G}_0$, regressand $\boldsymbol{b}_0$, optimistic variance $\bar{\boldsymbol{V}}_0$ to $\mathbf{0}$.
3: **for** $t = 1, ..., T$ **do**
4: $\quad \boldsymbol{G}_t = \gamma \boldsymbol{I} + \sum_{s<t} \sum_{i \in \tau_s} \bar{V}_{s,i}^{-1} \boldsymbol{\phi}_s(i) \boldsymbol{\phi}_s(i)^\top$.
5: $\quad \boldsymbol{b}_t = \sum_{s<t} \sum_{i \in \tau_s} \bar{V}_{s,i}^{-1} \boldsymbol{\phi}_s(i) X_{s,i}$. ▷*Compute statistics using historical data re-weighted by optimistic variance*
6: $\quad \hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1} \boldsymbol{b}_t$. ▷*Parameter estimation*
7: $\quad$ **for** $i \in [m]$ **do**
8: $\quad\quad \bar{\mu}_{t,i} = \langle \boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t \rangle + 2\beta \|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$. ▷*UCB value.*
9: $\quad\quad \underline{\mu}_{t,i} = \langle \boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t \rangle - 2\beta \|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$. ▷*LCB value.*
10: $\quad$ **end for**
11: $\quad S_t = \text{ORACLE}(\bar{\mu}_{t,1}, ..., \bar{\mu}_{t,m})$. ▷*Oracle using UCBs.*
12: $\quad$ Play S_t and observe arms τ_t with outcomes $(X_{t,i})_{i \in \tau_t}$.
13: $\quad \bar{V}_{t,i} = \max_{\mu \in [\underline{\mu}_{t,i}, \bar{\mu}_{t,i}]} \mu(1-\mu)$ for $i \in [m]$. ▷*Compute optimistic variance.*
14: **end for**

also allows for probabilistic feedback when τ_t is randomly determined. Furthermore, it can handle partial feedback that depends on specific stopping criteria [21]. It is also worth noting that C-PMC generalizes PMC-G by allowing a probably time-varying feature map $\boldsymbol{\phi}_t$. Specifically, if we set $\boldsymbol{\theta}^* = (\mu_1, ..., \mu_m)$ and fix $\boldsymbol{\phi}_t(i) = \boldsymbol{e}_i$, where $\boldsymbol{e}_i \in \mathbb{R}^m$ denotes the one-hot vector with 1 at the i-th entry and 0 elsewhere, we can easily reproduce the PMC-G setting described in [12].

The goal of C-PMC is to accumulate as much reward as possible over T rounds by learning the underlying parameter $\boldsymbol{\theta}^*$. The performance of an online learning algorithm A is evaluated based on its *regret*, defined as the difference between the expected cumulative reward obtained by always playing the best action $S_t^* \triangleq \text{argmax}_{S \in \mathcal{S}} r(S; \boldsymbol{\mu}_t)$ at each round t, and the expected cumulative reward by playing actions chosen by algorithm A. As mentioned before, it could be NP-hard to compute the exact S_t^* even when $\boldsymbol{\mu}_t$ is known, so similar to [21], [28]–[30], we assume that the algorithm A has access to an offline α-approximation oracle. This oracle takes a mean vector $\boldsymbol{\mu}$ as input and outputs an action S such that $r(S; \boldsymbol{\mu}) \geq \alpha \cdot r(S^*; \boldsymbol{\mu})$. Given the α-approximation oracle, the T-round α-approximate regret is defined as

$$\text{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\alpha \cdot r(S_t^*; \boldsymbol{\mu}_t) - r(S_t; \boldsymbol{\mu}_t)\right)\right], \quad (3)$$

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

III. Algorithm Design

In this section, we introduce the **V**ariance-**A**daptive **C**ontextual **C**ombinatorial **U**pper **C**onfidence **B**ound Algorithm for the C-PMC problem (VAC^{2}UCB), provided in Algorithm 1. Different from the CUCB algorithm [28] that neglects the contextual information and directly estimates the mean $\boldsymbol{\mu}_{t,i}$ for each arm, contextual bandit algorithms typically estimate the underlying parameter $\boldsymbol{\theta}^*$, which allows for efficient learning and computation for large-scale applications [19]. Specifically, $\boldsymbol{\theta}^*$ is estimated by solving the following ℓ_2-regularized least-square problem with regularization parameter $\gamma > 0$:

$$\hat{\boldsymbol{\theta}}_t = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmin}} \sum_{s<t} \sum_{i \in \tau_s} (\langle \boldsymbol{\theta}, \boldsymbol{\phi}_s(i) \rangle - X_{s,i})^2 + \gamma \|\boldsymbol{\theta}\|_2^2. \quad (4)$$

The closed form solution to this problem is precisely given by $\hat{\boldsymbol{\theta}}_t = \boldsymbol{G}_t^{-1} \boldsymbol{b}_t$, where the Gram matrix $\boldsymbol{G}_t = \sum_{s<t} \sum_{i \in \tau_s} \boldsymbol{\phi}_s(i) \boldsymbol{\phi}_s(i)^\top$ and the b-vector $\boldsymbol{b}_t = \sum_{s<t} \sum_{i \in \tau_s} \boldsymbol{\phi}_s(i) X_{s,i}$. However, directly using this parameter estimation will yield suboptimal regret $\tilde{O}(d\sqrt{KT})$ with an additional factor of $O(\sqrt{K})$ [21], where K is the maximum number of nodes in $\mathcal{U}$ that could be selected. The main issue lies in the lack of variance adaptivity and the equal treatment of each data point. To address this limitation, VAC^{2}UCB leverages the second-order statistics, specifically the variance, to re-weight each data point and obtain a more accurate estimator. To gain the intuition, we begin by assuming 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 employs the *weighted least-squared estimation* to learn the parameter $\boldsymbol{\theta}^*$:

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

where the first term is inversely *weighted* by the true variance $V_{s,i}$. The closed-form solution of this 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 $\boldsymbol{b}_t = \sum_{s<t} \sum_{i \in \tau_s} V_{s,i}^{-1} \boldsymbol{\phi}_s(i) X_{s,i}$. Notably the form of the solution is similar to that in VAC^{2}UCB, but the weights used for updating $\boldsymbol{G}_t$ and $\boldsymbol{b}_t$ differ (lines 4-5).

The intuition behind using the inverse of $V_{s,i}$ to re-weight the feedback data is as follows: when the variance is smaller, the observation $(\boldsymbol{\phi}_t(i), X_{t,i})$ is more accurate, making this data more important for the agent to learn unknown $\boldsymbol{\theta}^*$. By doing so, VAC^{2}UCB achieves a faster learning speed compared to the suboptimal approach that treats each data point equally (Li et al., 2016).

For our C-PMC setting, a new challenge arises since the variance $V_{s,i} = \mu_{s,i}(1 - \mu_{s,i})$ is not known a priori. To address this challenge, we construct an optimistic estimation $\bar{V}_{s,i}$ (line 13) 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 $\max_{\mu \in [\underline{\mu}_{t,i}, \bar{\mu}_{t,i}]} \mu(1-\mu)$, 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 become more accurate, the optimistic variance $V_{t,i}$ is also approaching the true variance $V_{t,i}$.

In order to theoretically guarantee $\hat{\boldsymbol{\theta}}_t$ is a good estimator, we prove a lemma (Lemma 1) to guarantee the concentration bound of $\boldsymbol{\theta}_t$ in the presence of unknown variance. The details will be presented in a later section. Building on this lemma, we construct an optimistic estimation, referred to as upper confidence bound (UCB) for each arm's mean $\bar{\mu}_{t,i}$ (line 8). The UCB values is computed using the empirical mean $\langle \boldsymbol{\phi}_t(i), \hat{\boldsymbol{\theta}}_t \rangle$ and a confidence interval $2\beta \|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$ along the direction $\boldsymbol{\phi}_t(i)$. Similarly, we construct $\underline{\mu}_{t,i}$ to provide lower bound

(LCB) of $\boldsymbol{\mu}_{t,i}$ (line 9). The size of the confidence interval influences the exploration for each arm, where larger intervals provide more incentive for the agent to explore arm i, thereby reducing uncertainty in the direction $\boldsymbol{\phi}_t(i)$. As a convention, we clip $\bar{\mu}_{t,i}, \underline{\mu}_{t,i}$ into $[0,1]$ if they are above 1 or below 0.

Once the UCB values $\bar{\boldsymbol{\mu}}_t$ are computed, the agent selects action S_t using an offline oracle with $\bar{\boldsymbol{\mu}}_t$ as input (line 11). After the agent plays S_t, the base arms in τ_t are revealed, and the agent receives $(X_{t,i})_{i\in\tau_t}$ as feedback. These observations (reweighted by optimistic variance $\bar{V}_{t,i}$) are used to update $\boldsymbol{G}_t$ and $\boldsymbol{b}_t$ to improve decisions in future rounds (lines 4-5).

IV. Theoretical Analysis

Here, we present our main theoretical result and analysis.

A. Main Result

Theorem 1. *For a C-PMC problem instance $(\mathcal{G}, [m], \mathcal{S}, \mathcal{C}, \Phi, \Theta, D_{obs}, R)$, the α-approximate regret of VAC^{2}UCB (Algorithm 1) is upper bounded by $\tilde{O}(d\sqrt{|\mathcal{V}|T})$, where $\tilde{O}$ hides polylogarithmic factors regarding T.*

Discussions. To state the regret bound, we first denote the minimum observation probability as $p_{\min} = \min_{i\in[m],\boldsymbol{\mu}\in\mathcal{M},S\in\mathcal{S},p_i^{\boldsymbol{\mu},S}>0} p_i^{\boldsymbol{\mu},S}$. Looking at Theorem 1, the regret does not have any dependence on action size K, number of base arms $\mathcal{E}$, and the triggering probability $p_{\min}$. For C-PMC bandit, [21] is the closest work to ours, and following their C^3-UCB algorithm can only give $\tilde{O}(d\sqrt{K|\mathcal{V}|T}/p_{\min})$. Our regret is *strictly better* than theirs by a factor of $\tilde{O}(\sqrt{K}/p_{\min})$. When the contexts remain unchanged at each round, Liu et al. [12] also propose a variance-adaptive algorithm VACUCB by directly estimating the unknown mean vector $\boldsymbol{\mu}$, which gives a $\tilde{O}(\sqrt{|\mathcal{E}||\mathcal{V}|T})$ regret bound. Our regret improves theirs by a factor of $O(\sqrt{|\mathcal{E}|}/d)$ by leveraging the context information via weighted least-squared estimator. For the classical PMC bandit with semi-bandit feedback, [31] recently gives a lower bound of $\Omega(\sqrt{dT})$, which means our regret bound matches the lower bound up to a factor of $\tilde{O}(\sqrt{d})$.

B. Regret Analysis

To prove our regret bound, we first introduce a series of key lemmas as follows. The first lemma gives a concentration bound of our weighted least-squared estimator $\boldsymbol{\theta}_t$ in the presence of the unknown variance, which ensures that $\hat{\boldsymbol{\theta}}_t$ is a good estimator.

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

Proof. See Appendix VII-A. ∎

Similar results exist in the literature for unweighted least-squared estimators [19], [21] and for weighted least-squared estimators where a single arm is selected/observed in each round. In our case, however, the weighted gram matrix $\boldsymbol{G}_t$ may be significantly larger than an unweighted version or the weighted version with single arm selection. To address this, we carefully use the Freedman's version of the Bernstein's inequality and apply recursion to prove a tight confidence radius β, which is only a factor of $O(\log K)$ larger compared to the unweighted least-squared estimator.

Building on this lemma, we claim that the UCB (resp. LCB) values are accurate optimistic (resp. pessimistic) estimations.

Lemma 2 (Arm-level over/under-estimation)**.** *With probability at least $1-\delta$, we have $\mu_{t,i} \le \bar{\mu}_{t,i} \le \mu_{t,i} + 3\beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$, and $\mu_{t,i} \ge \underline{\mu}_{t,i} \ge \mu_{t,i} - 3\beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$ for all $i \in [m]$.*

Proof. See Appendix VII-B ∎

Given the arm-level over/under-estimation, we proceed to relate the arm-level error to the total regret for C-PMC.

Lemma 3 (Reward and observation sensitivity)**.** *For C-PMC with semi-bandit or probabilistic feedback, and for any parameters $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathcal{M}$, let $\boldsymbol{\Delta} = (\mu_i' - \mu_i)_{i\in[m]}$, $\boldsymbol{V} = (1/\sqrt{(1-\mu_i)\mu_i})_{i\in[m]}$, and $\boldsymbol{p}^{\boldsymbol{\mu},S} = (p_i^{\boldsymbol{\mu},S})_{i\in[m]}$, the **reward sensitivity** $r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})$ satisfies $|r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})| \le \sqrt{|\mathcal{V}|}\|\boldsymbol{p}^{\boldsymbol{\mu},S} \odot \boldsymbol{V} \odot \boldsymbol{\Delta}\|_2$, and the **observation sensitivity** $p_i^{\boldsymbol{\mu},S} - p_i^{\boldsymbol{\mu}',S}$ satisfies $|p_i^{\boldsymbol{\mu}',S} - p_i^{\boldsymbol{\mu},S}| \le \|\boldsymbol{p}^{\boldsymbol{\mu},S} \odot \boldsymbol{\Delta}\|_1$,*

Proof. See Appendix VII-C ∎

Intuitively, the reward sensitivity bounds the reward difference by ℓ_2 norm of each arm's over/under-estimation $\boldsymbol{\Delta}$ given by our VAC^{2}UCB algorithm, and $\sqrt{|\mathcal{V}|}$ is to bound the non-linearity of $r(S;\boldsymbol{\mu})$. Similarly, the observation sensitivity bounds the observation probability difference by ℓ_1 norm of each arm's over/under-estimation $\boldsymbol{\Delta}$. Notice that $\boldsymbol{\Delta}$ is point-wise weighted by $\boldsymbol{p}^{\boldsymbol{\mu},S}$ in both reward and observation sensitivity, which reduces the regret contribution from unlikely observed arms to save a $1/p_{\min}$ factor. For observation sensitivity, $\boldsymbol{\Delta}$ is further point-wise weighted by a variance-related vector $\boldsymbol{V}$, which coincides with our variance-adaptive algorithm to improve a factor of $O(K)$.

Finally, we can relate the total regret to the cumulative contextual information bounded by the following lemma.

Lemma 4 (Weighted Ellipsoidal Potential)**.** *With probability at least $1-\delta$, $\sum_{t=1}^{T}\sum_{i\in\tau_t}\|\boldsymbol{\phi}_t(i)\|^2_{\boldsymbol{G}_t^{-1}}/\bar{V}_{t,i} \le O(d\log T)$.*

Proof. See Appendix VII-D. ∎

Equipped with the above lemmas, we are ready to show the analysis of Theorem 1 as follows.

Proof of Theorem 1. Let $\tilde{\boldsymbol{\mu}}_t$ be the vector whose i-th entry is the maximizer that achieves $\bar{V}_{t,i}$, i.e., $\tilde{\mu}_{t,i} = \operatorname{argmax}_{\mu\in[\underline{\mu}_{t,i},\bar{\mu}_{t,i}]}\mu(1-\mu)$, we bound the regret $\mathrm{Reg}(T)$ by $\mathrm{Reg}(T) = \mathbb{E}[\sum_{t=1}^{T}\alpha r(S_t^*;\boldsymbol{\mu}_t) - r(S_t;\boldsymbol{\mu}_t)] \le \mathbb{E}[\sum_{t=1}^{T}\alpha r(S_t^*;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)] \le \mathbb{E}[\sum_{t=1}^{T} r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\boldsymbol{\mu}_t)] \le \mathbb{E}[\sum_{t=1}^{T}|r(S_t;\bar{\boldsymbol{\mu}}_t) - r(S_t;\tilde{\boldsymbol{\mu}}_t)| + |r(S_t;\boldsymbol{\mu}_t) - r(S_t;\tilde{\boldsymbol{\mu}}_t)|]$, where the first inequality follows from $\bar{\boldsymbol{\mu}}_{t,i} \ge \boldsymbol{\mu}_{t,i}$ by Lemma 2 and the fact that $r(S;\boldsymbol{\mu})$ is monotone regarding $\boldsymbol{\mu}$, the second inequality is by the definition of S_t, the last inequality is by the triangle inequality.

Define $\tilde{S}$ to be the set of arms that can be observed, i.e., $\{i \in [m] : p_i^{\boldsymbol{\mu},S} > 0, \text{ for any } \boldsymbol{\mu} \in \mathcal{M}\}$. Now Let $\tilde{\boldsymbol{x}}_t = (|\bar{\mu}_{t,i} - \tilde{\mu}_{t,i}| / \sqrt{V_{t,i}})_{i \in \tilde{S}_t}$, $\boldsymbol{x}_t = (|\mu_{t,i} - \tilde{\mu}_{t,i}| / \sqrt{V_{t,i}})_{i \in \tilde{S}_t}$, $\bar{\boldsymbol{x}}_t = (|\bar{\mu}_{t,i} - \underline{\mu}_{t,i}| / \sqrt{\bar{V}_{t,i}})_{i \in \tilde{S}_t}$ we can bound term $\mathbb{E}[\sum_{t \in [T]} |r(S_t; \bar{\boldsymbol{\mu}}_t) - r(S_t; \tilde{\boldsymbol{\mu}}_t)|] \leq \sqrt{|\mathcal{V}|}\mathbb{E}[\sum_{t=1}^T ||\boldsymbol{p}^{\tilde{\boldsymbol{\mu}}_t, S_t} \odot \tilde{\boldsymbol{x}}_t||_2 \leq \sqrt{|\mathcal{V}|}\mathbb{E}[\sum_{t=1}^T ||\boldsymbol{p}^{\boldsymbol{\mu}_t, S_t} \odot \tilde{\boldsymbol{x}}_t||_2 + ||(\boldsymbol{p}^{\tilde{\boldsymbol{\mu}}_t, S_t} - \boldsymbol{p}^{\boldsymbol{\mu}_t, S_t}) \odot \tilde{\boldsymbol{x}}_t||_2] \leq \sqrt{|\mathcal{V}|}\mathbb{E}[\sum_{t=1}^T ||\boldsymbol{p}^{\boldsymbol{\mu}_t, S_t} \odot \tilde{\boldsymbol{x}}_t||_2 + ||\boldsymbol{p}^{\boldsymbol{\mu}_t, S_t} \odot \boldsymbol{x}_t||_1 ||\tilde{\boldsymbol{x}}_t||_2] \leq \sqrt{|\mathcal{V}|}\mathbb{E}[(T \sum_{t=1}^T \sum_{i \in \tilde{S}_t} p_i^{\boldsymbol{\mu}_t, S_t} \bar{x}_{t,i}^2)^{1/2} + ((K+1)|\mathcal{V}|/p_{\min})^{1/2} \sum_{t=1}^T \sum_{i \in \tilde{S}_t} p_i^{\boldsymbol{\mu}_t, S_t} \bar{x}_{t,i}^2] \leq (|\mathcal{V}| T \mathbb{E}[\sum_{t=1}^T \sum_{i \in \tau_t} \bar{x}_{t,i}^2])^{1/2} + C_0 \mathbb{E}[\sum_{t=1}^T \sum_{i \in \tau_t} \bar{x}_{t,i}^2] \leq O(d\sqrt{|\mathcal{V}|T} \log T + C_0 d^2 \log^2 T) = \tilde{O}(d\sqrt{|\mathcal{V}|T})$, where the first inequality follows from reward sensitivity in Lemma 3, the second inequality is by triangle inequality, the third inequality is by applying observation sensitivity in Lemma 3, fourth inequality follows from Cauchy-Schwarz inequality for the first term over T, and Cauchy-Schwarz inequality for the second term over i, incorporating the facts that $p_i^{\tilde{\boldsymbol{\mu}}_t, S_t} \geq p_{\min}$, $\tilde{x}_{t,i} \leq \bar{x}_{t,i}$, and $|\tilde{S}_t| \leq (K+1)|\mathcal{V}|$, the fifth inequality follows from Jensen's inequality, $\mathbb{E}[p_i^{\boldsymbol{\mu}_t, S_t} \mid \mathcal{H}_t] = \mathbb{E}[\mathbb{I}\{i \in \tau_t\} \mid \mathcal{H}_t]$, and letting constant $C_0 = ((K+1)|\mathcal{V}|^2/p_{\min})^{1/2}$, and the last inequality follows from $\sum_{t=1}^T \sum_{i \in \tau_t} \bar{x}_{t,i}^2 \leq \sum_{t=1}^T \sum_{i \in \tau_t} 9\beta^2 \|\boldsymbol{\phi}_t(i)\|^2_{\boldsymbol{G}_t^{-1}} / \bar{V}_{t,i} \leq O(d^2 \log^2 T)$ by Lemma 2 and Lemma 4. For term $\mathbb{E}[\sum_{t=1}^T |r(S_t; \boldsymbol{\mu}_t) - r(S_t; \tilde{\boldsymbol{\mu}}_t)|]$, we can bound it similarly by $\tilde{O}(d\sqrt{|\mathcal{V}|T})$ as above, which concludes the proof. ■

V. Applications for C-PMC

We consider two representative applications to validate the effectiveness of our proposed method: 1) context-aware mobile crowdsensing and 2) context-aware and user-targeted content delivery. We compare the regret of our VAC^{2}UCB algorithm with three baselines: VACUCB [12], the state-of-the-art variance-adaptive algorithm for PMC-G bandits; C^{3}UCB [21], the state-of-the-art contextual combinatorial bandit algorithm that is not variance-adaptive; and ϵ-greedy, which chooses a random action with fixed probability ϵ for exploration and otherwise greedily chooses the empirically optimal action.

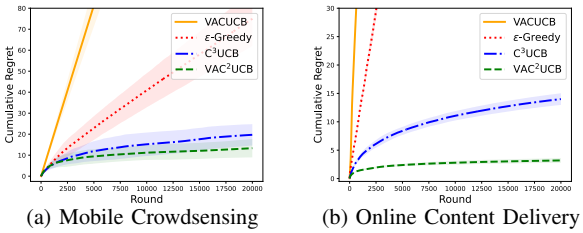
A. Context-Aware Mobile Crowdsensing

1) Problem Description: In recent years, the rapid growth of mobile devices such as smartphones and wearable devices, which are equipped with powerful built-in sensors like cameras, microphones, accelerometers, has given rise to the concept of mobile crowdsensing (MCS) [3]–[5]. MCS enables the collection and analysis of sensing data from physical environments with the active participation of mobile users. This paradigm offers both opportunities and challenges. On one hand, MCS facilitates large-scale sensing projects by recruiting a large group of individuals who can collectively utilize their devices to cover various locations as they move through an area [3]. On the other hand, the quality of the collected data can vary across different participants and locations due to differences in participants' movement trajectories and variations in device manufacturing quality [32].

A recent study by Liu et al. [5] highlights the close relationship between the MCS data quality and various sensing contexts, including hardware contexts (e.g., phone brand, sensor models, sensor calibration level), human behavior contexts (e.g., smartphone holding position, human movement during sensing), and geometric contexts (e.g., location, indoor/outdoor). Motivated by this finding, we propose the context-aware mobile crowdsensing (C-MCS) problem: taking into account the available user and geometric contexts, how should the task organizer strategically select a group of individuals to maximize the collection of high-quality data from different locations in a city, .

The C-MCS problem can be formulated within the C-PMC model. We consider a bipartite graph $G(\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{U}$ represents the set of candidate participants, $\mathcal{V}$ denotes the set of locations in the city, and $\mathcal{E}$ models the data collection process. At each time t, the task organizer can observe contextual information c_t, which captures the hardware and human behavior contexts for each candidate participant, as well as the geometric contexts for each target location. The objective of C-MCS is to select at most K participants (with K determined based on a recruitment budget) to perform the sensing task. Each selected participant $u \in S_t$ independently uploads sensor data at location $v \in V$. This upload is modeled as a Bernoulli random variable $X_{t,(u,v)} \in 0, 1$, with a probability $\mu_{t,(u,v)}$ that the data collected by participant u at location v is valid and can be used to cover location v. In this formulation, the arms correspond to the edges in $\mathcal{E}$, and given the feature vector $\boldsymbol{\phi}_t((u,v))$ of each user-location pair (u,v), the probability is modeled as $\mu_{t,(u,v)} = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t((u,v)) \rangle$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ represents the unknown parameter. The task organizer receives *semi-bandit feedback*, observing whether the uploaded data is valid or not for the pair (u,v) given that $u \in S_t$. The observation probability is denoted as $p_{(u,v)}^{\mu_t, S_t} = 1$ if $u \in S_t$, and 0 otherwise. The reward is defined as the weighted total number of locations covered with valid information: $r(S; \boldsymbol{\mu}_t) = \sum_{v \in V} \mu_{t,v} \left(1 - \prod_{u \in S}(1 - \mu_{t,(u,v)})\right)$, where the known weight $\mu_{t,v}$ represents the importance of covering location v in the crowdsensing task. Busy areas, for instance, may have higher sensing importance as their environmental conditions impact more people.

2) Performance Evaluation: We simulate the C-MCS problem using complete bipartite graphs with $|\mathcal{U}| = 40$ candidate nodes (participants) and $|\mathcal{V}| = 10$ target nodes (locations in the city). We first choose $K = 5$ (number of chosen participants) with dimension $d = 10$, and generate $\mu_{t,(u,v)} = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}_t((u,v)) \rangle$, where each dimension of the feature vector $\boldsymbol{\phi}_t((u,v))_i$ and θ_i^* are sampled from uniform distribution $U(0, 0.15)$ and $U(0, 0.5)$ for $i \in [d]$, respectively. The importance weights $\mu_{t,v}$ are sampled from the uniform distribution $U(0, 0.5)$ and known to the task organizer. Fig. 2a shows the cumulative regrets of algorithms for 20000 rounds, VAC^{2}UCB achieves 96%, 82% and 25% less regret than the VACUCB, ϵ-greedy (where we choose $\epsilon = 0.2$ in all experiments) and C^{3}UCB algorithms. To verify how parameters $K, |\mathcal{V}|, d$, and feature distribution affect the regret, we conduct



(a) Mobile Crowdsensing (b) Online Content Delivery

Fig. 2: Cumulative regrets in different applications. We show the average regrets with standard deviations over 5 experiments.

a parameter study on the base case $(d, K, |\mathcal{V}|, U(0, x)) = (10, 5, 40, U(0, 0.15))$, where other parameters follow the setting of Fig. 2a, and change the corresponding parameter within $d \in \{10, 30, 50\}$, $K \in \{10, 30, 50\}$, $|\mathcal{V}| \in \{10, 100, 1000\}$, $x \in \{0.05, 0.10, 0.15\}$, respectively. For parameter K in Fig. 3a, the regret of our algorithm changes mildly as K increases, consistent with our regret bound in Theorem 1 that is independent of K. For parameter $|\mathcal{V}|$ in Fig. 3b, the regret decreases as $|\mathcal{V}|$ increases, which contradicts our (worst-case) regret bound. The reason is that in the average case, it may be easier to find the optimal action with a large number of target nodes $|\mathcal{V}|$, causing less regret. For parameter d in Fig. 3c, the regret increases almost linearly, consistent with our theory. Fig. 3d shows with the decrease of x, the improvement over C^{3}UCB increases from 25% to 51%. The reason is that with small x, the variance is smaller, and each data is re-weighted with larger weights, making our variance-adaptive algorithm learn faster and incur much less regret. Compared with the most competitive algorithm C^{3}UCB, our VAC^{2}UCB consistently outperforms C^{3}UCB algorithm by at least 11%, 25%, 12%, and 25% less regret for varying $K, |\mathcal{V}|, d, x$, respectively.

B. Context-Aware and User-Targeted Content Delivery

1) Problem Description: Content delivery network (CDN) is a network of distributed servers strategically placed across various locations, which widely appears in web services such as video streaming, software downloading, and web loading [1], [33]. Unlike traditional methods that rely on a central server, CDNs replicate and cache content on multiple servers, enabling users to access data from servers that are physically closer to them. This approach improves delivery speed and reliability, resulting in a better user experience.

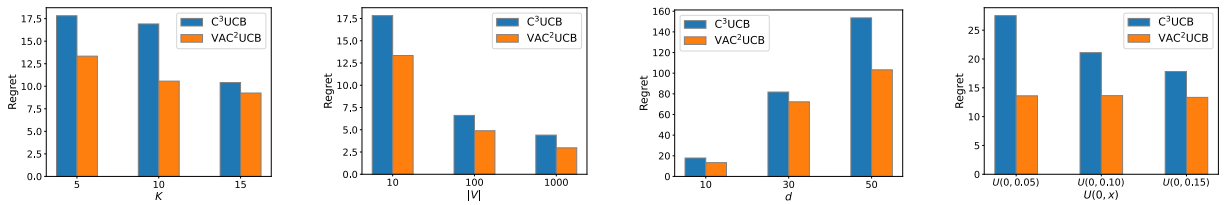
Context-aware and user-targeted CDN takes a step further by prioritizing service based on user preferences and dynamically adapting the content delivery based on users' network capabilities [34], [35]. Our framework aims to help content owners (e.g., media companies or e-commerce vendors) to select a set of servers to enhance the user experience by delivering content more effectively, reducing latency, and providing content that is most relevant and engaging for each user.

The above application scenario fits naturally into our C-PMC problem with a bipartite graph $\mathcal{G}(\mathcal{U}, \mathcal{V}, \mathcal{E})$, where $\mathcal{U}$ models the set of candidate servers, $\mathcal{V}$ are the end users, and $\mathcal{E}$ models the user-server interactions as follows. At each time slot t, the agent (or the content owner) aims to select $S_t \subseteq \mathcal{U}$ servers that can cache the t-th content and deliver the content to users through the CDN network. The number of selected servers at each round is constrained by K, as the maintenance costs of each server and the content owner's budget are considered. Before making the decision, the content owner collects contextual information c_t, which includes users' demographic information (gender, age, location, etc.) and network conditions (network delay, bandwidth, jitter, etc.) between each server and user. The selected servers $u \in S_t$ then independently deliver contents for each user $v \in V$ with unknown success probability $\mu_{t,(u,v)}$, depending on varying network condition [36]. By "success," we mean the content is delivered in a timely and high-quality manner, which can be modeled by a Bernoulli random variable $X_{t,(u,v)} \in \{0, 1\}$, with the mean $\mu_{t,(u,v)} = \langle \boldsymbol{\theta}_1^*, \boldsymbol{\phi}_t((u, v)) \rangle$, where $\boldsymbol{\theta}_1^*$ is the unknown parameter and $\boldsymbol{\phi}_t$ is the feature mapping function given context c_t. For user-targeted CDN, we use the quality of service (QoS) scores as features. In particular, for each $(u, v) \in \mathcal{E}$, the feature is a $d_1 = 4$ dimensional score vector $\boldsymbol{\phi}_t((u, v)) = [L, J, D, B]$, where L, J, D, B are scores determined by packet loss, jitter, packet delay, and bandwidth between server u and user v. Specifically, $\langle \boldsymbol{\theta}_1^*, \boldsymbol{\phi}_t((u, v)) \rangle$ is a weighted combination of above four QoS scores.

We assume that each user v attempts to preload content from the selected servers to their device [37]. We use a Bernoulli random variable with unknown mean $\mu_{t,v}$ to represent whether the preloaded content is ultimately consumed (e.g., video is viewed) by the user. To model users' preference for the contents, we use $\boldsymbol{\phi}_t(v)$ to map a d_2-dimensional feature vector related to users' demographic information (gender, age, location, etc.). Similarly, the mean $\mu_{t,v}$ is represented as $\mu_{t,v} = \langle \boldsymbol{\theta}_2^*, \boldsymbol{\phi}_t(v) \rangle$, where $\boldsymbol{\theta}_2^* \in \mathbb{R}^{d_2}$ is the unknown preference vector modeling user-content interaction. In this formulation, we can see that arms correspond to the success probability $\mu_{t,(u,v)}$ for $(u, v) \in \mathcal{E}$ and the consuming probability $\mu_{t,v}$ for users $v \in \mathcal{V}$. The question is how to select K mirror servers to maximize the total number of users that consume the contents with unknown success rates and consuming probabilities. A good server selection policy should prioritize successful delivery to users more likely to consume the content. The underlying parameter is now $\boldsymbol{\theta}^* = [\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*] \in \mathbb{R}^{d_1+d_2}$.

As for the feedback, the content owner can observe whether the contents are successfully delivered from the selected servers, i.e., the values of $X_{t,(u,v)}$ for $u \in S_t$ and $v \in \mathcal{V}$. This feedback is known as semi-bandit feedback, and the observation probability $p_{(u,v)}^{\boldsymbol{\mu}_t, S_t}$ equals 1 if $u \in S_t$ and 0 otherwise. Additionally, if a user v successfully receives the content, the content owner can observe whether the user consumes the content, i.e., the value of $X_{t,v}$ is observed if $\exists v$ s.t. $X_{u,v} = 1$. This feedback is called *probabilistic feedback* because it depends on other random outcomes, and the observation probability is given by $p_v^{\boldsymbol{\mu}_t, S_t} = 1 - \prod_{u \in S_t}(1 - \mu_{t,(u,v)})$. The expected reward is essentially Eq. (2) and the agent's goal is to minimize the total regrets in Eq. (3).

2) Performance Evaluation: For the user-targeted online content delivery experiment, we consider 10 candidate server



(a) Varying K (b) Varying $|\mathcal{V}|$ (c) Varying d (d) Varying distribution $U(0, x)$

Fig. 3: Total regrets after 20000 rounds in different settings for context-aware mobile crowdsensing.

locations from the point-of-presence (POP) locations of Microsoft Azure CDN in North America[2]. We assume the users are distributed in 10 POP locations. For the feature vector of network condition $\boldsymbol{\phi}_t((u, v)) \in \mathbb{R}^{d_1}$, we extract a $d_1 = 4$ dimensional QoS score vector, regarding network delay, jitter, bandwidth, and package loss. The network delay and jitter are collected from real network testing results between user u's location and server v's location[3]. The bandwidth and package loss are sampled from uniform distribution $U(0.9\text{Mbps}, 3\text{Mbps})$ and $U(0\%, 1\%)$. To compute the score $\boldsymbol{\phi}_t(u, v)$, we use function $\max((200 - \text{delay})/200, 0)$ and $\max((10 - \text{jitter})/10, 0)$ to normalize delay and jitter, bandwidth and package loss are linearly normalized to $(0, 1)$ by dividing by the maximum value. For the parameter $\boldsymbol{\theta}_1^*$, we extract the relative importance degree of the four parameters above to model $\boldsymbol{\theta}_1^* = [0.149, 0.151, 0.111, 0.598]$ as suggested by [38]. To extract users' preference for the contents, we use the MovieLens-1M dataset, which contains 1 million ratings from 6000 users on 4000 movies.[4] Using a rank-d singular value decomposition (SVD) with $d = 10$, we learn a feature mapping $\boldsymbol{\phi}(v)$ from users' rating to the probability that a uniformly random movie is rated by the user v more than three stars (indicating the user v likes the movie). This gives a preference distribution of user v for a uniformly random movie, i.e., $\mu_v = \langle \boldsymbol{\theta}_2^*, \boldsymbol{\phi}(v) \rangle$. In each round t, since users randomly arrive to be served, both a content (movie) and 10 users at 10 user locations are sampled uniformly at random in our experiments. Fig.2b shows the cumulative regrets of different algorithms for 20000 rounds. VAC^{2}UCB achieves 99%, 99% and 77% less regret than VACUCB, ϵ-Greedy and C^{3}UCB.

VI. Conclusion

In this paper, we propose the first context-aware PMC bandit model which can incorporate time-varying contextual information. We devise a variance-adaptive online learning algorithm and conduct rigorous analysis to show strictly better regrets. Experiments validate that our algorithm can achieve at least 11% and 77% improvement on mobile crowdsensing and content delivery applications, respectively. For future directions, it would be interesting to generalize the linear structure by considering neural network structures [39] or model misspecifications [40].

[2]https://docs.microsoft.com/en-us/azure/cdn/cdn-pop-locations
[3]https://wondernetwork.com/pings
[4]https://grouplens.org/datasets/movielens/1m/

VII. Appendix

A. Proof of Lemma 1

Recall that for $t \geq 1$, $X_{t,i}$ is a Bernoulli random variable with mean $\mu_{t,i} = \langle \theta^*, \boldsymbol{\phi}_t(i) \rangle$. We can rewrite $X_{t,i} = \mu_{t,i} + \eta_{t,i}$, where noise $\eta_{t,i} \in [-1, 1]$, its mean $\mathbb{E}[\eta_{t,i} \mid \mathcal{F}_{t-1}] = 0$, and its variance $\text{Var}[\eta_{t,i} \mid \mathcal{F}_{t-1}] = \mu_{t,i}(1 - \mu_{t,i})$, where filtration $\mathcal{F}_{t-1} = \mathcal{H}_t \bigcup S_t$. Let us define $\boldsymbol{\zeta}_t = \sum_{s<t} \sum_{i \in \tau_s} \eta_{s,i} \boldsymbol{\phi}_t(i) / \bar{V}_{s,i}$.

Definition 1. *We define failure events $F_0 \subseteq F_1 \subseteq ... \subseteq F_T$, be a sequence of events by $F_t = \{\exists s \leq t$ such that $\|\boldsymbol{\zeta}_s\|_{\boldsymbol{G}_s} + \sqrt{\gamma} \geq \beta\}$, where $\beta = 1 + \sqrt{\gamma} + 4\sqrt{\log\left((6TN/\delta)\log(3TN/\delta)\right)}$ and $N = (4d^2K^4T^4)^d$.*

If $\neg F_t$ holds, then by the definition of $\hat{\boldsymbol{\theta}}_t, \boldsymbol{G}_t, \boldsymbol{\zeta}_t$ and $X_{t,i}$, $\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{G}_t} = \|\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t - \gamma \boldsymbol{G}_t^{-1}\boldsymbol{\theta}^*\|_{\boldsymbol{G}_t} \leq \|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}} + \gamma\|\boldsymbol{\theta}^*\|_{\boldsymbol{G}_t^{-1}} \leq \beta - \sqrt{\gamma} + \sqrt{\gamma} = \beta$, which concludes Lemma 1.

To prove Lemma 1, we only need to bound the probability of the failure events F_t, which for any $t = 1, ..., T$,

Lemma 5. $\Pr\left[\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t} + \sqrt{\gamma} \geq \beta \text{ and } \neg F_{t-1}\right] \leq \delta/T$.

Before we prove Lemma 5, we need following lemmas.

Lemma 6. *For any $s < t$, $\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_t^{-1}}/\bar{V}_{s,i} \leq \|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i}$, and if $\neg F_{t-1}$ holds and $\bar{V}_{s,i} < 1/4$, $\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}/\bar{V}_{s,i} \leq 2/\beta \leq 1$ for any $i \in [m]$.*

Proof. The first inequality is by $\boldsymbol{G}_t \geq \boldsymbol{G}_s$. For the second inequality, when $\neg F_{t-1}$ holds, $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_s\|_{\boldsymbol{G}_s} \leq \beta$, and since $\bar{V}_{s,i} < 1/4$, it follows from the definition of $\bar{V}_{s,i}$ that at least one of the following is true: (a). $\bar{V}_{s,i} \geq (\langle \boldsymbol{\phi}_s(i), \hat{\boldsymbol{\theta}}_s + 2\beta\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}} \rangle)/2 \geq \beta\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}/2$, (b). $\bar{V}_{s,i} \geq (1 - \langle \boldsymbol{\phi}_s(i), \hat{\boldsymbol{\theta}}_s + 2\beta\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}} \rangle)/2 \geq \beta\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}/2$. ■

Lemma 7. *If $\neg F_t$, then (1). $\|\boldsymbol{\phi}_t(i)\|_2^2/\bar{V}_{t,i} \leq 4dKt$, (2). $\|\boldsymbol{\phi}_t(i)\|_1/\bar{V}_{t,i} \leq 4dKt$, (3). $\|\boldsymbol{\zeta}_{t+1}\|_1 \leq 2dK^2(t+1)^2$.*

Proof. For (1), if $\bar{V}_{t,i} = 1/4$, the inequality trivially holds since $\|\boldsymbol{\phi}_t(i)\| \leq 1$. Consider $\bar{V}_{t,i} < 1/4$, and $\lambda_{\max}$ be the maximum eigenvalue of $\boldsymbol{G}_t$. Then, it holds that $\|\boldsymbol{\phi}_t(i)\|_2^2/\bar{V}_{t,i} \leq \|\boldsymbol{\phi}_t(i)\|_2^2/(\beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}) \leq \|\boldsymbol{\phi}_t(i)\|_2/\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}} = \|\boldsymbol{G}_t^{1/2}\boldsymbol{G}_t^{-1/2}\boldsymbol{\phi}_t(i)\|_2/\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}} \leq \sqrt{\lambda_{\max}}$, where the first inequality follows from Lemma 6, the second inequality is by $\beta \geq 1, \|\boldsymbol{\phi}_t(i)\| \leq 1$.

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 $\|\boldsymbol{\phi}_t(i)\|_2^2/\bar{V}_{t,i} \leq \sqrt{\lambda_{\max}} \leq$

$\sqrt{\text{trace}(\boldsymbol{G}_t)} = \sqrt{\gamma d + \sum_{s=1}^{t-1}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_2^2/\bar{V}_{s,i}} \le \sqrt{Kd + \sum_{s=1}^{t-1} 4dK^2 s} \le \sqrt{d(K+2K^2t(t-1))} \le 4dKt$, 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.

For (2), $\|\boldsymbol{\phi}_t(i)\|_1/\bar{V}_{t,i} \le \sqrt{d}\|\boldsymbol{\phi}_t(i)\|_2/\bar{V}_{t,i} \le \sqrt{d}\|\boldsymbol{\phi}_t(i)\|_2/(\beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}) \le \|\boldsymbol{\phi}_t(i)\|_2/\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}} = \|\boldsymbol{G}_t^{1/2}\boldsymbol{G}_t^{-1/2}\boldsymbol{\phi}_t(i)\|_2/\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}} \le \sqrt{\lambda_{\max}} \le 4dKt$, where the first inequality uses Cauchy-Schwarz, the second inequality uses $\beta \ge \sqrt{d}$, and the rest follows from the proof of (1).

For (3), $\|\boldsymbol{\zeta}_t\|_1 = \|\sum_{s<t}\sum_{i\in\tau_s}\eta_{s,i}\boldsymbol{\phi}_s(i)/\bar{V}_{s,i}\|_1 \le \sum_{s<t}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)/\bar{V}_{s,i}\|_1 \le \sum_{s<t}\sum_{i\in\tau_s} 4dKt \le 2dK^2t^2$, where the first inequality follows from $\eta_{s,i}\in[-1,1]$, the second inequality follows from (1). ■

Proof of Lemma 5. Let $\boldsymbol{v} \in \mathbb{R}^d$ and define $V_{s,i,\boldsymbol{v}} = \text{Var}[\eta_{s,i} \mid \mathcal{F}_{s-1}]\langle\boldsymbol{\phi}_s(i),\boldsymbol{v}\rangle^2/\bar{V}_{s,i}^2$ if $\bar{V}_{s,i} < 1/4$, and $V_{s,i,\boldsymbol{v}} = \langle\boldsymbol{\phi}_s(i),\boldsymbol{v}\rangle^2/\bar{V}_{s,i}$ otherwise. Let $R_{\boldsymbol{v}} = \max_{s<t,i\in\tau_s}\{\langle\boldsymbol{\phi}_s(i),\boldsymbol{v}\rangle/\bar{V}_{s,i} : \bar{V}_{s,i} < 1/4\}$.

By applying the Theorem 9 in [41], which is essentially a modification of the Freedman's version of the Bernstein's inequality [42], [43], with probability at least $1-\delta/T$, it holds that $\langle\boldsymbol{\zeta}_t,\boldsymbol{v}\rangle = \sum_{s<t}\sum_{i\in\tau_s}\eta_{s,i}\langle\boldsymbol{\phi}_s(i),\boldsymbol{v}\rangle/\bar{V}_{s,i} \le 2(R_{\boldsymbol{v}}+1)/3\cdot\log(1/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}})\log(1/\delta_{\boldsymbol{v}})}$ where $\delta_{\boldsymbol{v}} = 3\delta/(T(1+R_{\boldsymbol{v}})^2(1+\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}})^2)$.

Since $\boldsymbol{v}$ could be a random variable, we use the covering argument trick (Chap.20, [11]) to handle $\boldsymbol{v}$. Specifically, we define the covering set $\Lambda = \{j\cdot\varepsilon : j = -C/\varepsilon, -C/\varepsilon+1, ..., C/\varepsilon-1, C/\varepsilon\}^d$, with size $N = |\Lambda| = (2C/\varepsilon)^d$ and parameters C,ε will be determined shortly after. By applying union bound on, we have with probability at least $1-\delta$ that $\langle\boldsymbol{\zeta}_t,\boldsymbol{v}\rangle \le 2(R_{\boldsymbol{v}}+1)/3\cdot\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}})\log(N/\delta_{\boldsymbol{v}})}$ for all $\boldsymbol{v}\in\Lambda$.

Now we can set $\boldsymbol{v} = \boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t$, and it follows from Lemma 7 that $\|\boldsymbol{v}\|_\infty \le \|\boldsymbol{\zeta}_t\|_1 \le 2dK^2t^2 \triangleq C$. 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 $\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2 = \langle\boldsymbol{\zeta}_t,\boldsymbol{v}\rangle \le \|\boldsymbol{\zeta}_t\|_1\varepsilon + \langle\boldsymbol{\zeta}_t,\boldsymbol{v}'\rangle \le \|\boldsymbol{\zeta}_t\|_1\varepsilon + 2(R_{\boldsymbol{v}}+1)/3\cdot\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}})\log(N/\delta_{\boldsymbol{v}})} \le \|\boldsymbol{\zeta}_t\|_1\varepsilon + 2(R_{\boldsymbol{v}}+1)/3\cdot\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2)\log(N/\delta_{\boldsymbol{v}})}$, where the second inequality uses the fact that $R_{\boldsymbol{v}'}\le R_{\boldsymbol{v}}, V_{s,i,\boldsymbol{v}'}\le V_{s,i,\boldsymbol{v}}, 1/\delta_{\boldsymbol{v}'}\le 1/\delta_{\boldsymbol{v}}$ for any $\boldsymbol{v}'\le\boldsymbol{v}$, the third inequality follows from the following derivation, $\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}} \le \sum_{s<t}\sum_{i\in\tau_s}\langle\boldsymbol{\phi}_s(i),\boldsymbol{v}\rangle^2/\bar{V}_{s,i} = \sum_{s<t}\sum_{i\in\tau_s}(\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t)^\top\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t/\bar{V}_{s,i} = (\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t)^\top(\sum_{s<t}\sum_{i\in\tau_s}\boldsymbol{\phi}_s(i)\boldsymbol{\phi}_s(i)^\top/\bar{V}_{s,i})\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t \le (\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t)^\top\boldsymbol{G}_t(\boldsymbol{G}_t^{-1}\boldsymbol{\zeta}_t) = \|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2$, where the first inequality follows from $\neg F_{s-1}$ which implies $\|\boldsymbol{\theta}^*-\hat{\boldsymbol{\theta}}_s\|_{\boldsymbol{G}_s}\le\beta$ for $s<t$ and thus $\bar{V}_{t,i}\ge\text{Var}[\eta_{s,i}\mid\mathcal{F}_{s-1}]$, the second inequality 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 $\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2 \le \|\boldsymbol{\zeta}_t\|_1\varepsilon + 2(R_{\boldsymbol{v}}+1)/3\cdot\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2)\log(N/\delta_{\boldsymbol{v}})} \le C\varepsilon + 2(2\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}/\beta + 1)/3\cdot\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2)\log(N/\delta_{\boldsymbol{v}})} \le 1 + 2\log(N/\delta_{\boldsymbol{v}}) + \sqrt{2(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2)\log(N/\delta_{\boldsymbol{v}})}$, where the second inequality is by $R_{\boldsymbol{v}}\le 2\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}/\beta$ under $\neg F_{t-1}$, the last inequality holds since β is an upper bound of $\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}$.

By rearranging and simplifying the above derivation, we have $\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}} + \sqrt{\gamma} \le 1 + \sqrt{\gamma} + 4\sqrt{\log(N/\delta_{\boldsymbol{v}})} \le 1 + \sqrt{\gamma} + 4\sqrt{\log(6TN/\delta(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2))}$, where the last inequality is because of $\delta_{\boldsymbol{v}} \ge (3\delta)(T(1+\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2))$ from the definition of $\delta_{\boldsymbol{v}}$, and $\sum_{s<t}\sum_{i\in\tau_s}V_{s,i,\boldsymbol{v}} \le \|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}}^2$. Finally, we solve the above equation and set $\beta = 1 + \sqrt{\gamma} + 4\sqrt{\log(6TN/\delta\cdot\log(3TN/\delta))}$, which completes reduction on t to show $\Pr[\|\boldsymbol{\zeta}_t\|_{\boldsymbol{G}_t^{-1}} + \sqrt{\gamma} \ge \beta] \ge 1-\delta/T$ under $\neg F_{t-1}$. ■

B. Proof of Lemma 2

Proof. For any $i\in[m], t\in[T]$, we have $\left|\langle\hat{\boldsymbol{\theta}}_t,\boldsymbol{\phi}_t(i)\rangle - \langle\boldsymbol{\theta}^*,\boldsymbol{\phi}_t(i)\rangle\right| = \left|\langle\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}^*,\boldsymbol{\phi}_t(i)\rangle\right| \le \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{G}_t}\cdot\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}} \le \beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$, where the first inequality by Cauchy-Schwartz, the second by Lemma 1. Now use the definition of $\mu_{t,i} = \langle\boldsymbol{\theta}^*,\boldsymbol{\phi}_t(i)\rangle$ and $\bar{\mu}_{t,i} = \langle\hat{\boldsymbol{\theta}},\boldsymbol{\phi}_t(i)\rangle + 2\beta\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}$ finises the proof. ■

C. Proof of Lemma 3

Proof. For the reward sensitivity, it follows from Lemma 1 in [12] by setting $\eta_i = 0, \zeta_i = \mu_i' - \mu_i$. For observation sensitivity, we consider two cases: semi-bandit feedback and probabilistic feedback. For semi-bandit feedback, it trivially holds. For probabilistic feedback, $p_v^{\boldsymbol{\mu},S} = 1 - \prod_{u\in S}(1-\mu_{(u,v)})$, without loss of generality, we assume $S = \{1,...,K\}$, $|p_v^{\boldsymbol{\mu}',S} - p_v^{\boldsymbol{\mu},S}| = \sum_{i\in[K]}|\mu'_{(u,v)} - \mu_{(u,v)}|\prod_{j=1}^{i-1}(1-\mu_{(i,v)})\prod_{j=i+1}^{K}\mu'_{(i,v)} \le \sum_{i\in[K]}|\mu'_{(u,v)} - \mu_{(u,v)}| \le \|\boldsymbol{p}^{\boldsymbol{\mu},S}\odot\boldsymbol{\Delta}\|_1$. ■

D. Proof of Lemma 4

Proof. We begin by recursively bound the determinant of $\boldsymbol{G}_{t+1}$, $\det(\boldsymbol{G}_{t+1}) = \det(\boldsymbol{G}_t + \sum_{i\in\tau_t}\boldsymbol{\phi}_t(i)\boldsymbol{\phi}_t(i)^\top/\bar{V}_{t,i}) = \det(\boldsymbol{G}_t)\cdot\det(\boldsymbol{I} + \sum_{i\in\tau_t}G_t^{-1/2}\boldsymbol{\phi}_t(i)(G_t^{-1/2}\boldsymbol{\phi}_t(i))^\top/\bar{V}_{t,i}) \ge \det(\boldsymbol{G}_t)\cdot(1+\sum_{i\in\tau_t}\|\boldsymbol{\phi}_t(i)\|_{\boldsymbol{G}_t^{-1}}^2/\bar{V}_{t,i}) \ge \det(\gamma\boldsymbol{I})\prod_{s=1}^t(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{t,i})$. If $\bar{V}_{s,i} = 1/4$, $\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{s,i} \le 4\|\boldsymbol{\phi}_s(i)\|^2/\lambda_{\min}(\boldsymbol{G}_s) \le 4/\gamma \le 1/K$, else if $\bar{V}_{s,i} < 1/4$, and since $\neg\mathcal{F}_T$, by Lemma 6, $\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{s,i} \le 1/(\beta\sqrt{\gamma}) \le 1/\gamma \le 1/(4K)$. Therefore, we have $\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2 \le 1$. Using the fact that $2\log(1+x)\ge x$ for any $[0,1]$, we have $\sum_{s\in t}\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{s,i} \le 2\sum_{s=1}^t\log(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{s,i}) = 2\log\prod_{s=1}^t(1+\sum_{i\in\tau_s}\|\boldsymbol{\phi}_s(i)\|_{\boldsymbol{G}_s^{-1}}^2/\bar{V}_{s,i}) \le 2\log(\det(\boldsymbol{G}_{t+1})/\det(\gamma\boldsymbol{I})) \le 2\log((\gamma+KT/d)^d/\gamma^d) = 2d\log(1+4dK^2T^2/(\gamma d)) \le 4d\log(KT)$, where the second last inequality follows from the determinant bound at the beginning of this section, the last inequality follows from Lemma 15 of [19] by setting $L = \|\boldsymbol{\phi}_s(i)\|^2/\bar{V}_{s,i} \le 4dKs$ (from Lemma 7). ■

REFERENCES

[1] M. Pathan, R. Buyya, and A. Vakali, "Content delivery networks: State of the art, insights, and imperatives," *Content Delivery Networks*, pp. 3–32, 2008.

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

[3] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[4] K. Han, C. Zhang, and J. Luo, "Taming the uncertainty: Budget limited robust crowdsensing through online learning," *Ieee/acm transactions on networking*, vol. 24, no. 3, pp. 1462–1475, 2015.

[5] S. Liu, Z. Zheng, F. Wu, S. Tang, and G. Chen, "Context-aware data quality estimation in mobile crowdsensing," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 2017, pp. 1–9.

[6] Y. Sun, S. Zhou, and Z. Niu, "Distributed task replication for vehicular edge computing: Performance analysis and learning-based algorithm," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1138–1151, 2020.

[7] Y. Zhou, H. Sun, Y. Jin, Y. Zhu, Y. Li, Z. Qian, S. Zhang, and S. Lu, "Inference replication at edges via combinatorial multi-armed bandit," *Journal of Systems Architecture*, vol. 129, p. 102636, 2022.

[8] D. S. Hochba, "Approximation algorithms for np-hard problems," *ACM Sigact News*, vol. 28, no. 2, pp. 40–52, 1997.

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

[10] A. Krause and C. Guestrin, "Near-optimal observation selection using submodular functions," in *AAAI*, vol. 7, 2007, pp. 1650–1654.

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

[12] X. Liu, J. Zuo, H. Xie, C. Joe-Wong, and J. C. Lui, "Variance-adaptive algorithm for probabilistic maximum coverage bandits with general feedback," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023.

[13] Ö. Yürür, C. H. Liu, Z. Sheng, V. C. Leung, W. Moreno, and K. K. Leung, "Context-awareness for mobile sensing: A survey and future directions," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 68–93, 2014.

[14] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

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

[16] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[17] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[18] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 208–214.

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

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

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

[22] M. Hefeeda and H. Ahmadi, "A probabilistic coverage protocol for wireless sensor networks," in *2007 IEEE International Conference on Network Protocols*. IEEE, 2007, pp. 41–50.

[23] J. Zuo, X. Liu, C. Joe-Wong, J. C. Lui, and W. Chen, "Online competitive influence maximization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 11 472–11 502.

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

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

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

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

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

[29] X. Liu, J. Zuo, S. Wang, C. Joe-Wong, J. Lui, and W. Chen, "Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms," *arXiv preprint arXiv:2208.14837*, 2022.

[30] X. Liu, J. Zuo, S. Wang, J. C. Lui, M. Hajiesmaili, A. Wierman, and W. Chen, "Contextual combinatorial bandits with probabilistically triggered arms," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 559–22 593.

[31] N. Merlis and S. Mannor, "Tight lower bounds for combinatorial multi-armed bandits," in *Conference on Learning Theory*. PMLR, 2020, pp. 2830–2857.

[32] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, "Automatically characterizing places with opportunistic crowdsensing using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 481–490.

[33] L. Chen, J. Xu, S. Ren, and P. Zhou, "Spatio–temporal edge service placement: A bandit learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8388–8401, 2018.

[34] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–9.

[35] F. Paganelli, M. Ulema, and B. Martini, "Context-aware service composition and delivery in ngsons over sdn," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 97–105, 2014.

[36] S. A. Bitaghsir, A. Dadlani, M. Borhani, and A. Khonsari, "Multi-armed bandit learning for cache content placement in vehicular social networks," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2321–2324, 2019.

[37] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria, "Take-away tv: Recharging work commutes with predictive preloading of catch-up tv content," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2091–2101, 2016.

[38] H. J. Kim, D. G. Yun, H.-S. Kim, K. S. Cho, and S. G. Choi, "Qoe assessment model for video streaming service using qos parameters in wired-wireless network," in *2012 14th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2012, pp. 459–464.

[39] C. Liu and Y.-X. Wang, "Global optimization with parametric function approximation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 113–22 136.

[40] Z. Wang, J. Xie, X. Liu, S. Li, and J. C. Lui, "Online clustering of bandits with misspecified user models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

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

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