Fostering Responsibility in Email Marketing: A Contextual Restless Bandit Framework

Ibtihal El Mimouni[1,2] (✉) and Konstantin Avrachenkov[1]

[1] INRIA Sophia Antipolis, Biot, France {ibtihal.el-mimouni, k.avrachenkov}@inria.fr
[2] Smartprofile, Valbonne, France

Abstract. Email marketing is increasingly criticized due to ethical concerns, as bulk email campaigns often result in spam, reduced engagement, and negative user experiences. In addition, there is increasing awareness of the environmental impact, as these large-scale campaigns contribute to carbon emissions. To address these issues, we introduce QWIC-Fair (Q-learning Whittle Index with Context and Fairness), an algorithm that operates within a Contextual Restless Multi-Armed Bandit framework. QWIC-Fair leverages implicit feedback to learn the dynamics of user interactions and thus target users with relevant content. In this model, each user represents an arm of the bandit, evolving as a Markov Decision Process that captures state transitions reflecting their interactions with email contents, while accounting for contextual information. The algorithm also incorporates a fairness constraint to ensure balanced selection and to avoid repetitive targeting of the same users. The experiments conducted, using synthetic and real-world data, show that QWIC-Fair outperforms existing email marketing approaches.

Keywords: Reinforcement learning · Restless bandits · Whittle index · Q-learning · Fairness · Recommender systems · Responsible email marketing

1 Introduction

In today's digital landscape, email marketing has become an essential tool for businesses to reach out to potential customers [12]. However, the volume of emails sent daily raises ethical and environmental issues, particularly with traditional marketing strategies that often deliver generic content to large market segments. This results in high spam rates [23], a negative user experience [22], and a tarnished domain reputation [44]. Moreover, these practices also contribute to digital carbon emissions: a typical email has a carbon footprint that ranges between 0.3g and 26g of CO_2, while an email with an attachment can reach up to 50g of CO_2 [5,36]. Individual emails may have a relatively small carbon footprint, but the cumulative effect of poorly targeted campaigns can be significant: in 2023, an estimated 347 billion emails were sent and received around the world [37].

These challenges are compounded by the fact that user engagement is not static: preferences and responsiveness shift over time due to factors such as seasonal trends or changes in personal preferences. As a result, traditional static or myopic approaches, which optimize only for immediate user response, often fail to capture these evolving patterns. This can lead to over-targeting of active users while neglecting others, resulting in disengagement and unfairness. Moreover, bulk campaigns frequently deliver irrelevant content, which not only contributes to user fatigue but also increases the carbon footprint by generating unnecessary emails. To address these issues, personalization is key. It allows for tailored offers that align with users' interests, enhancing engagement [28,41] and reducing the likelihood of emails being marked as spam. By targeting effectively, marketers can achieve better results with fewer emails, thereby minimizing the digital carbon footprint associated with mass campaigns.

Building on this principle of personalization, we propose to frame the problem as a sequential decision-making task using the Contextual Restless Multi-Armed Bandit framework. We model each user as an evolving arm, represented by a context-augmented Markov decision process that captures user's state transitions based on their email interactions. Within this framework, we introduce Q-learning Whittle Index with Context and Fairness (QWIC-Fair). The algorithm operates in an episodic manner. Each episode consists of L time steps, during which actions are taken using an epsilon-greedy strategy. This involves either randomly exploring users or exploiting by selecting users based on the highest Whittle indices, which are a measure of the value of activating a particular user. QWIC-Fair functions on two distinct timescales: on a fast timescale, it updates the Q-values by adjusting estimates of the expected cumulative reward for each state-action-context triplet. On a slow timescale, it updates the Whittle indices. At the end of each episode, the learning agent evaluates a fairness constraint to ensure balanced targeting among users. Specifically, it identifies under-selected users and prioritizes their selection in the following iterations. We showcase the effectiveness of QWIC-Fair by comparing it to baselines often used in email marketing, using two simulators: one built from real-world data and another from synthetic data.

Our contributions can be summarized as follows: (1) We introduce a Contextual Restless Multi-Armed Bandit framework designed for email Recommender Systems (RS), where we consider each user as an arm of the bandit. (2) We design a practically relevant algorithm called QWIC-Fair, which leverages context-aware Whittle index-based Q-learning, and incorporates a fairness constraint, thereby ensuring equitable selection of the users. (3) Our approach promotes ethical email marketing practices by guiding user selection to prevent spamming and to reduce the carbon footprint associated with bulk email campaigns. (4) Experiments, using both real and synthetic data, demonstrate the effectiveness of QWIC-Fair.

The paper is structured as follows: Section 2 reviews the related literature. Section 3 presents restless bandits and introduces the Whittle index policy. Section 4 formalizes the contextual restless bandit problem and explains the email

recommender application. Section 5 details the proposed algorithm. Section 6 outlines the experiments conducted and discusses the results.

2 Related Work

2.1 Bandits in Recommenders

Recommender Systems (RS) [40] have proven to be an effective tool [24] in guiding users through large pools of content, products, and services by suggesting the most pertinent items. RS are now broadly implemented across multiple industries [16,30,43], using various methods such as collaborative and content-based filtering [13,34]. Despite their popularity, these methods have some limitations, notably when favoring some popular items and limiting the exploration of potentially relevant ones. To address the exploration-exploitation dilemma, Multi-Armed Bandits (MAB) [10,25,50] have been explored in recommenders. These algorithms balance discovering new options, and exploiting previous knowledge about items that have been previously recommended. A more advanced type of bandits that incorporate contextual information (such as user demographics, time, or device) are the Contextual Multi-Armed Bandits (CMAB). LinUCB, proposed by Li et al. [27], is one the most popular CMAB algorithms. The authors solve the problem of personalized news recommendations at Yahoo!. However, even CMAB models may fall short in environments where user behavior keeps changing over time. To address this, Restless Multi-Armed Bandits (RMAB) emerged, allowing arms to evolve over time, even when not selected by the agent. Most bandit algorithms for recommenders consider each arm as the item to recommend [27,31]. In our work, we model each user as an arm.

2.2 Restless Bandits

Restless bandits have become very popular over the years, finding applications in different domains such as healthcare [6], web crawling [2], and communication networks [1]. For our RS use case, our study leverages seminal research on RMAB: Whittle [51] introduced the Whittle index policy, which allows for the activation of M arms on average by calculating an index for each arm and selecting top M arms with the highest indices. Building on Whittle's work, there has been a surge of interest in developing algorithms to calculate the Whittle index. For instance, Avrachenkov and Borkar [3] focused on the time-average criterion and developed a tabular algorithm that converges to the Whittle index. Various other approaches have been proposed to tackle the RMAB problem, including index policies [15,32,49] and Reinforcement Learning (RL) techniques [47,52]. Another line of research that is related to ours is the growing literature on fairness in RMAB. Some authors focused on quota-based fairness [19,35], while others studied satisfying the fairness among the allocated resources [7]. Other works explored soft fairness where fairness constraints are considered on average [26,46].

2.3 Contextual Restless Bandits

A particularly promising direction is the combination of contextual and restless bandits, coined as Contextual Restless Multi-Armed Bandits (CRMAB). While the incorporation of context has been explored in contextual MAB [27,9], research on CRMAB remains limited. To the best of our knowledge, the only works that address CRMAB are by Chen and Hou [11], who proposed a model-based online learning algorithm that combines index policies with a dual decomposition framework, allowing for simultaneous learning of the arm models and decision-making. They applied their algorithm to smart grid optimization. On the other hand, Liang et al [29], developed a Bayesian CRMAB approach tailored to public health interventions. They used Thompson sampling and relied on informative priors to model arm behavior. In contrast, our method is model-free, using Q-learning to learn arm dynamics from observed interactions. Moreover, in addition to reward maximization, our approach ensures balanced exposure across arms by incorporating a fairness constraint.

3 Preliminaries

3.1 Restless Bandits

The RMAB problem is a generalization of the MAB framework. Unlike the classical bandit problems, restless bandits model a more realistic scenario where arms evolve over time regardless of whether they are selected. As a result, computing optimal policies becomes PSPACE-hard [33], and the exploration-exploitation trade-off is further complicated by the need to balance learning both active and passive arms.

Let us consider a RMAB with N arms, each evolving according to a Markov Decision Process (MDP). At each time step t, the agent selects a subset of M arms to activate, where $M < N$. The state of arm i at time t is denoted by $s_i(t) \in \mathcal{S}$, where $\mathcal{S}$ is the finite state space of the arm.

The state of each arm is controlled by the action chosen. Specifically, let $a_i(t)$ be the action taken on arm i at time t, where $a_i(t) = 1$ if arm i is activated, and $a_i(t) = 0$ if left passive. The state transition probabilities are defined as:

$$P(s_i(t+1) = s' | s_i(t) = s, a_i(t) = a) = P^{i,a}_{s,s'}, \tag{1}$$

where $P^{i,a}_{s,s'}$ is the probability of transitioning from current state s to next state s' for arm i under action a.

Each arm generates a reward depending on its state and the action taken. Let $r_i(s_i(t), a_i(t))$ denote the reward obtained from arm i at time t, at state $s_i(t)$, when action $a_i(t)$ is taken. The goal is to determine a policy π, that specifies which arms to activate at each time step, to maximize the expected cumulative reward. Under the total discounted criterion ($\gamma \in (0, 1)$ being the discount factor) and infinite horizon, the objective is to solve the following:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{i=1}^{N} \gamma^t \; r_i(s_i(t), a_i(t))\right], \tag{2}$$

subject to the constraint that no more than M arms can be active simultaneously:

$$\sum_{i=1}^{N} a_i(t) \leq M, \quad \forall t \geq 0. \tag{3}$$

RMAB problems are computationally expensive [33]. Consequently, researchers have proposed approximation techniques to make these problems more tractable. One particularly significant advancement in this area is the Whittle index heuristic [51].

3.2 Whittle Index

Whittle's index policy utilizes the concept of Lagrangian relaxation to address the complexities of the RMAB problem. Whittle proposed to relax the constraint (3) to apply on average, rather than strictly, by incorporating a Lagrange multiplier, denoted $\tilde{\lambda}$. The objective function becomes the following:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{i=1}^{N} \gamma^t \left(r_i(s_i(t), a_i(t)) + \tilde{\lambda}(1 - a_i(t))\right)\right]. \tag{4}$$

This relaxation allows the RMAB problem to be decoupled into N independent subproblems, each solved by the associated Bellman equation for the state value function:

$$\begin{aligned} V_i(s) = \max_{a\in\{0,1\}} \Big[& a\big(r_i(s,1) + \gamma \sum_j p_i(j|s,1) V_i(j)\big) \\ & + (1-a)\big(r_i(s,0) + \tilde{\lambda} + \gamma \sum_j p_i(j|s,0) V_i(j)\big)\Big]. \end{aligned} \tag{5}$$

We rewrite Equation (5) as a function of the state-action pair, which determines the Q-values that reflect the expected future rewards of taking an action in a given state:

$$Q_i(s,a) = \begin{cases} r_i(s,1) + \gamma \sum_j p_i(j|s,1) V_i(j), & \text{if } a = 1, \\ r_i(s,0) + \tilde{\lambda} + \gamma \sum_j p_i(j|s,0) V_i(j), & \text{if } a = 0. \end{cases} \tag{6}$$

Whittle interpreted the Lagrange multiplier $\tilde{\lambda}$ as a subsidy for passivity. Accordingly, the Whittle index λ is defined as the smallest subsidy that makes the agent indifferent between choosing an arm i ($a_i = 1$) or not choosing it ($a_i = 0$). The RMAB problem is (Whittle) indexable if the set of states for which it is optimal to activate the arm increases monotonically with λ. For a given state k, $\lambda(k)$ is determined such that the expected reward from activating and not activating the arm are equal:

$$Q(k,1) = Q(k,0). \tag{7}$$

The objective is to learn the Whittle indices λ by solving (7).

4 Problem Formulation

4.1 An Email Recommender System

In RS, ethically collected user feedback [18], that respects user privacy and consent, is crucial for enhancing personalization and improving the performance of these systems. It is categorized into: *Explicit feedback* [4], which involves direct inputs from users such as ratings, reviews, likes, or explicitly stated preferences. This type of feedback offers precise insights into user preferences because it reflects their evaluations. However, explicit feedback is not always available as it relies on users actively providing it. On the other hand, *Implicit feedback* [20] is gathered passively through users' interactions with the system. It includes data such as browsing history, purchase frequency, shares, time spent on certain items, etc. This form of feedback tends to be more abundant providing a rich source of data for inferring user preferences.

An email RS benefits from a substantial amount of implicit feedback including actions like opening emails, clicking on call-to-action buttons, unsubscribing, etc.

In our work, we model an email RS as a sequential decision-making problem, specifically as a contextual restless bandit. Our goal is to maximize user engagement while minimizing unnecessary emails. This not only enhances user satisfaction and reduces the risk of email fatigue but also helps address the environmental challenges discussed in Section 1.

4.2 A Contextual Restless Bandit

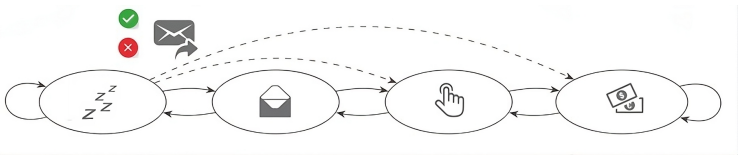


Fig. 1. Users are modeled as Markov decision processes, with state transitions reflecting engagement levels (idle, open, click, purchase).

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be the finite collection of N heterogeneous users, where each user u_i is an arm of the CRMAB. Each user behaves according to a Contextual Markov Decision Process (CMDP) [17]. CMDP extends traditional MDP by incorporating contextual information that influences both the dynamics of the environment and the rewards associated with different actions.

We define a CMDP by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{C}, P, r)$, where the state space $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ represents the users' implicit feedback. We consider fully observable states corresponding to four levels of engagement, as shown in Figure 1. *Open* is when the user opens an email, *click* is when the user clicks on a link

within the email, *purchase* refers to buying a product, *idle* indicates no interaction.

The context space $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ captures side information that could include the special features of the campaign (e.g. seasonal promotions), user's features (e.g. age, location, segment). Contexts enrich the data from which the system learns.

The actions correspond to the type of emails to send to users, each with a different content, such as offering promotions, showcasing product features, and inviting feedback, etc. This multi-action setup adds another layer of complexity, which we will address in an extended version of the current work. To simplify, for now, we only consider the two-action CRMAB framework. Thus, the action space is $\mathcal{A} = \{0, 1\}$, where $a = 1$ is the active action of sending a promotional email, and $a = 0$ is the passive action of not sending it.

The probability that user u transitions from state s to state s', given action a, and context c is: $P_{s,s'}^{u,a,c}$ following the notation in Equation (1).

The reward is a function of the current state, context, action, and next state, denoted as $r : \mathcal{S} \times \mathcal{C} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}$. It is designed by assigning a reward value for an email open, a higher reward for a click, and a highest reward for a purchase, while taking into account the contexts and the action taken. For instance, suppose we have two users: u_1, a young professional living in an urban area, and u_2, a retired individual living in a suburban area. Without adding context, the system might treat both users the same way and send them identical offers, potentially leading to lower engagement. Incorporating context into the reward function allows the recommender to learn the types of actions (promotional offers) that yield the best outcomes (user engagement) in different contexts.

The goal is to maximize the discounted cumulative reward over infinite time horizon, subject to the constraint (3) that, at time step t, no more than M out of N users can be chosen:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty}\sum_{i=1}^{N} \gamma^t \; r_{u_i}(s_{u_i}(t), c_{u_i}, a_{u_i}(t), s_{u_i}(t+1))\right]. \tag{8}$$

For ease of notation, we will refer to the reward as r.

5 Algorithm: QWIC-Fair

QWIC-Fair leverages Q-learning, a model-free RL technique [48], alongside the Whittle index policy [51]. It uses a two-timescale stochastic approximation scheme [8] where both control and parametric optimization occur simultaneously. The algorithm incorporates contextual information to guide informed decision-making based on user interaction dynamics. It also constrains selection with fairness criteria to ensure balanced targeting, and to prevent repeatedly activating the same users.

Algorithm 1 Q-learning Whittle Index with Context & Fairness (QWIC-Fair)

Initialize Q_table and λ_table, fairness threshold η, episodes $E_{episodes}$, time steps L, exploration parameter ϵ, discount factor γ, set of under-selected arms $\tilde{\mathcal{U}} = \emptyset$, $selection\_count$ to track arms' selections, $average\_reward$ to track arms' average reward across episodes, engagement threshold τ, interval to check inactive arms H
for $e = 1$ to $E_{episodes}$ **do**
  for $t = 1$ to L **do**
    /* *Exploration-Exploitation** /
    if Uni[0,1] $< \epsilon$ **then**
      Randomly select M arms from the set $\tilde{\mathcal{U}}$
      Fill remaining slots (if any) by selecting arms from the complement of $\tilde{\mathcal{U}}$
    else
      Select top M arms from the set $\tilde{\mathcal{U}}$ with the highest Whittle indices
      Fill remaining slots (if any) by selecting arms from the complement of $\tilde{\mathcal{U}}$
    end if
    /* *Update selection counters and* $\tilde{\mathcal{U}}$ */
    for each $selected\_arm$ **do**
      $selection\_count\ [selected\_arm] \leftarrow selection\_count\ [selected\_arm]+1$
      Remove $selected\_arm$ from the set $\tilde{\mathcal{U}}$ (if $selected\_arm \in \tilde{\mathcal{U}}$)
    end for
    Take actions $a(t)$, observe next states $s(t+1)$, contexts c, and rewards $r(t)$
    /* *On a faster timescale, update the Q-values* */
    for k in S **do**

$$Q(s(t),a(t),k,c) \leftarrow Q(s(t),a(t),k,c) + \alpha(t)\Big[(1-a(t))(r(t)+\lambda(k,c)) + a(t)r(t) + \gamma \max_{a' \in \{0,1\}} Q(s(t+1),a',k,c) - Q(s(t),a(t),k,c)\Big]$$

    end for
    /* *On a slower timescale, update the Whittle indices* */
    for k in $\mathcal{S}$ **do**
      for c in $\mathcal{C}$ **do**
        $\lambda(k,c) \leftarrow \lambda(k,c) + \beta(t)\Big(Q(k,1,k,c) - Q(k,0,k,c)\Big)$
      end for
    end for
  end for
  /* *Fairness constraint* */
  $\tilde{\mathcal{U}} = \{arm \mid selection\_count[arm] < \eta\ L\}$
  $selection\_count \leftarrow \{arm : 0\}$ for all arms
  /* *Put non-engaging arms to sleep* */
  if $e \equiv 0 \mod H$ **then**
    $\mathcal{Z} = \{arm \mid average\_reward[arm] < \tau\}$
    $\tilde{\mathcal{U}} = \{arm \mid selection\_count\ [arm] < \eta\ L$ and $arm \notin \mathcal{Z}\}$
  end if
end for

QWIC-Fair operates in an episodic manner, where each episode is of length L. The entire time horizon is denoted by T. Let $E_{episodes}$ be the total number

of episodes until time T, hence we have: $T = L\,E_{episodes}$. The episodic structure allows to periodically check the fairness constraint defined as follows:

Definition 1 (*Fairness constraint*). *Let $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ be the set of N arms, and let L be the length of an episode. The fairness constraint is satisfied if, in each episode, we have:*

$$selection\_count[u_i] \geq \eta L, \quad \forall u_i \in \mathcal{U}, \tag{9}$$

where $selection\_count[u_i]$ is the number of times arm u_i is selected during an episode e, and $\eta \in [0, 1]$ is a fairness threshold. The fairness constraint ensures that, at the end of each episode, each arm $u_i \in \mathcal{U}$ was selected at least ηL times. If $selection\_count[u_i] < \eta L$, the arm u_i is added to the set of under-selected arms $\tilde{\mathcal{U}}$ [1], and is prioritized in the subsequent episode.

In our QWIC-Fair algorithm, each arm u_i corresponds to an individual user. When clear from the context, we use the terms *arms* and *users* interchangeably. Thus, in an episode e, the set of under-selected arms $\tilde{\mathcal{U}}$ contains users who were targeted less frequently than others during that episode.

Each episode consists of L time steps. At each time step t, the algorithm employs an epsilon-greedy strategy to balance exploration and exploitation (see lines $4 \rightarrow 11$ of Algorithm 1):

- With probability ϵ, it explores the state-action-context space by randomly selecting arms from the set of under-selected users $\tilde{\mathcal{U}}$. If $|\tilde{\mathcal{U}}| < M$, the algorithm fills the remaining slots by selecting additional arms from the set: $\mathcal{U} \setminus \tilde{\mathcal{U}} = \{u_i \in \mathcal{U} \mid u_i \notin \tilde{\mathcal{U}}\}$.
- With probability $1 - \epsilon$, the algorithm exploits its current knowledge by selecting M users with the highest Whittle indices. Again, the algorithm first considers under-selected users by selecting the top M from $\tilde{\mathcal{U}}$ to ensure that fairness constraints are met. If $|\tilde{\mathcal{U}}| < M$, the remaining slots are filled by selecting arms with the highest Whittle indices from the complement of $\tilde{\mathcal{U}}$. This way the Whittle index policy respects the constraint (3).

For every selected user, the algorithm increments their selection counter, which tracks how often they are targeted during the episode. If the selected user $\in \tilde{\mathcal{U}}$, they are removed from this set for the remainder of the episode. This ensures that they are no longer prioritized as under-selected users for the rest of the current episode (see lines $12 \rightarrow 16$ of Algorithm 1).

Once the users are selected, the algorithm executes the actions $a(t)$, and observes the resulting next states $s(t+1)$, contexts c, and rewards $r(t)$. Next, it updates the Q-values and Whittle indices, following a two-timescale stochastic approximation approach with asynchronous iterates (see lines $17 \rightarrow 26$ of

[1] The maximum size of $\tilde{U}$ is $N - M$. This occurs when the algorithm repeatedly selects the same subset of M users. To ensure that all users in $\tilde{U}$ are covered throughout the episode, the minimum number of time steps required is: $L \geq \lceil \frac{N-M}{M} \rceil$

Algorithm 1). Specifically, on a fast timescale, it updates the Q-values:

$$\begin{aligned} Q(s(t),a(t),k,c) \leftarrow Q(s(t),a(t),k,c) + \alpha(t)\Big[&(1-a(t))\big(r(t)+\lambda(k,c)\big) \\ &+ a(t)r(t) + \gamma \max_{a'\in\{0,1\}} Q(s(t+1),a',k,c) - Q(s(t),a(t),k,c)\Big], \end{aligned} \tag{10}$$

and, on a slow timescale, it updates the Whittle indices:

$$\lambda(k,c) \leftarrow \lambda(k,c) + \beta(t)\left(Q(k,1,k,c) - Q(k,0,k,c)\right). \tag{11}$$

In the above, $\alpha(t)$ and $\beta(t)$ are the learning rates for the Q-values and the Whittle indices, respectively, with $\beta(t) = o(\alpha(t))$, as the Whittle index estimates need to be updated less frequently. We fix:

$$\alpha(t) = \frac{C}{\left\lceil \frac{t}{5000} \right\rceil}, \quad \beta(t) = \frac{C'}{1+\left\lceil \frac{t\log t}{5000} \right\rceil}\mathbf{I}\{t \mod N \equiv 0\}, \tag{12}$$

This ensures that:

- $\sum_{t=0}^{\infty}\alpha(t) = \infty$ and $\sum_{t=0}^{\infty}\alpha(t)^2 < \infty$: $\alpha(t)$ must decrease sufficiently slowly to ensure that the algorithm can explore the environment over time, while also decreasing quickly enough to guarantee convergence to the optimal Q-values.
- $\sum_{t=0}^{\infty}\beta(t) = \infty$ and $\sum_{t=0}^{\infty}\beta(t)^2 < \infty$: $\beta(t)$ must decrease gradually enough to allow continuous learning of Whittle index estimates, but also quickly enough to ensure the estimates eventually converge to their true values.

The theoretical guarantees from [3] include convergence of the Q-values and of the Whittle index estimates for each state. For our tabular setting, for time-independent context, the learning process inherits these convergence guarantees. Concerning time-dependant context, one may choose α and β to be small constant values with the condition that $\beta << \alpha$.

At the end of each episode, the algorithm checks the fairness criterion by reviewing the selection count for each user. Users who were under-selected are added to the set $\tilde{U}$. The elements of this set are prioritized in the subsequent episode, ensuring a more balanced targeting approach over time. The algorithm resets the selection counter for all users, to prepare for the next episode (see lines 28 $\rightarrow$ 30 of Algorithm 1). While fairness is important to ensure equitable selection among users, strictly adhering to the fairness constraint can sometimes lead to suboptimal outcomes. For instance, some users might consistently remain in the *idle* state, showing little to no engagement. Continuing to target these users can reduce system performance. To address this, after every H episodes, the algorithm identifies a set $\mathcal{Z}$ of users who exhibit low engagement by calculating the average reward of each user across previous episodes. If a user's average reward falls below a threshold τ, they are classified as non-engaging and added to $\mathcal{Z}$. We say that these users are *put to sleep*.

The threshold τ can either be set as a constant, or it can be a dynamic threshold that changes throughout episodes. For example, if we set τ to the 20th percentile of user rewards, it would be the value below which 20% of the rewards fall. Meaning that, in an episode e, if the 20th percentile of user rewards is 0.5, then τ would be 0.5. Users with rewards below this 0.5 threshold would be considered non-engaged and added to $\mathcal{Z}$. By periodically putting inactive users to sleep, the algorithm shifts its focus to users that are more likely to engage: users in $\mathcal{Z}$ are excluded from $\tilde{\mathcal{U}}$, so even though they would be flagged as under-selected, they are deprioritized in the following episode due to their low engagement (see lines 31 $\rightarrow$ 35 of Algorithm 1).

The episodic setting aligns well with the email RS, where campaigns are typically sent at regular intervals. An episode could represent a week or a month, with each time step corresponding to an email campaign or a decision to interact with a user.

6 Experiments and Results

6.1 Baselines

In our experiments, we compare QWIC-Fair against the following baselines:

Table 1. Baseline policies used for comparison with QWIC-Fair[2].

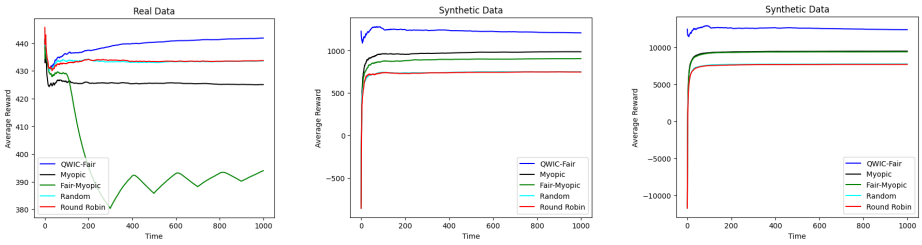
Policy	Definition
Random	Selects users randomly without considering engagement information, which is fair in expectation.
Myopic	Selects users who are most likely to result in conversions, based on immediate rewards.
Fair-Myopic	Selects users based on immediate rewards, while incorporating the fairness constraint.
Round-robin	Selects users in a cyclic order. This is by nature a fair policy because it guarantees equal distribution of email sends across all users.

We choose to compare with these baselines because they are currently used by Smartprofile [42], our partner company specializing in B2B digital marketing, and are often adopted by marketers in emailing. The objective is to demonstrate the practical improvements of our proposed method over existing emailing industry standards.

6.2 Real Dataset

We use a real-world dataset provided by Smartprofile. The dataset is a sample from one of their clients' data, which was gathered with user consent and adheres

[2] Code available at: https://github.com/cloud-commits/QWIC-Fair.



(a) N=100 / M=10 (b) N=1,000 / M=100 (c) N=10,000 / M=1,000

Fig. 2. Average rewards of different policies over a time horizon of $T = 1{,}000$. (a) shows results with real data; (b) and (c) show results with synthetic data. N is the total number of arms, and M[3] is the number of selected arms. The fairness threshold is $\eta = 10\%$, the exploration parameter is $\epsilon = 0.3$, and the discount factor is $\gamma = 0.9$.

to GDPR [14] regulations. The sample includes 10,000 distinct users, which is representative of small to medium-sized email marketing businesses. However, the proposed algorithm can scale to larger datasets involving millions of users. In fact, once Whittle indices are inferred from the modest-sized dataset, they can be immediately transferred to a dataset of million users: the application of Whittle indices is just a sorting procedure with complexity of $O(n \log n)$.

Analyzing user logs revealed that more than 60% of users exhibit low engagement. This indicates that most users are likely to remain in an *idle* state, with low probabilities of transitioning to more active states such as *clicking*, or *purchasing*. This results in data sparsity, which makes it challenging to have accurate model predictions about user behavior. To overcome this limitation, we developed a simulator that draws on the transitions observed in the real dataset. Despite initially having data from over 10,000 users, we could only reliably construct irreducible MDPs for 100 users due to sparse interactions and limited contextual features in the initial dataset. We used location as the primary time-independent contextual feature in this model.

We also explored a publicly available dataset, "messages-demo", provided by the REES46 Customer Data Platform project [38]. This dataset, accessible on Kaggle [39], contains messaging campaigns from a medium-sized retail company, delivered through various channels, including email, web and mobile push notifications, and SMS. Each message in the dataset is associated with detailed statistics, such as delivery, open, click, purchase events, and negative feedback (unsubscribes, spam complaints, and bounces). While we initially considered this dataset for our experiments, we encountered challenges when focusing solely on email interactions. In fact, after filtering the data to include only email-related events, we found that the number of interactions per user was insufficient to construct representative transition matrices. This sparsity posed a challenge for

[3] In a typical setup of Smartprofile, M represents 10% of N available users. We also use this ratio in our experiments.

our modeling framework, which requires adequate user interaction data to train effectively.

The REES46 data structure and content are very similar to the data provided by Smartprofile, which we used in our experiments. We encourage readers to explore this publicly available dataset as it is a useful resource for experimenting with similar algorithms and studying user interactions across multiple communication channels.

6.3 Synthetic Dataset

The purpose of the synthetic simulation is to enhance the modeling of user behavior by generating a richer dataset. To achieve this, we collaborated with our partner company Smartprofile, leveraging their expertise to accurately reflect email industry standards in the simulator's design. To provide a comprehensive representation of varied user behaviors, we categorized users into four distinct engagement levels: low, medium, high, and very high; and designed corresponding transition matrices for each class. We incorporated time-independent contextual features such as user location, age, and marital status, derived from distributions provided by INSEE [21]. Since the model is based on CMDP, having well-calibrated transition matrices is essential. We tested various setups. Plots (b) and (c) of Figure 2 illustrate the scenario where low user engagement is prominent, just like in the real dataset.

6.4 Results and Discussion

Performance plots in Figure 2 show that QWIC-Fair exceeds the policies, presented in Table 1, in terms of average rewards, for both synthetic and real-world data. In plot (a) of Figure 2, which uses real data, QWIC-Fair shows significant improvement over the other policies, particularly in the initial time steps, where it quickly converges to a higher average reward. Plots (b) and (c) of Figure 2, using synthetic data, also show that QWIC-Fair leads to higher rewards. The gap between QWIC-Fair and the other policies widens as the data size increases.

If fairness is omitted, standard Whittle-index-based Q-learning (QWIC) focuses solely on maximizing cumulative reward and does not inherently prevent repeated selection of high-reward arms. This can lead to fairness issues such as over-targeting some users while neglecting others, especially in domains where equitable exposure and long-term user engagement are important. While tuning parameters like ϵ in the ϵ-greedy strategy may help broaden exploration, this does not guarantee balanced exposure across users. QWIC-Fair addresses this by enforcing an episodic fairness constraint that ensures under-selected users are prioritized in subsequent episodes. This introduces a natural trade-off between fairness and immediate reward: in some real-world scenarios, user engagement follows a heavy-tailed Pareto distribution, where a small fraction of users drive most conversions. Allocating recommendations to less engaged users can reduce short-term reward, especially in early episodes when the system is still exploring.

The relevance of this trade-off depends on application goals. In email marketing, fairness is operationally preferred as over-targeting users risks unsubscribes and spam complaints, affecting both campaign effectiveness and sender reputation. Thus, fairness is not only an ethical consideration, but also a strategic requirement.

Smartprofile also provided data on carbon emissions from one of their email campaigns, estimating that a bulk campaign sent to around 197,000 users resulted in 789.16 kg of CO_2 emissions, whereas targeting only 17,885 users produced an estimated 71.54 kg of CO_2. This suggests that for a setting where $M \simeq 0.1\,N$, carbon emissions are reduced by approximately 90%. These estimates highlight that shifting to a more targeted approach can significantly reduce the environmental impact of email campaigns.

In this paper, we focused on a binary-action setting within the CRMAB framework, where the two actions are either to send a recommendation (active) or not send it (passive). However, the CRMAB framework can naturally extend to a multi-action setting, where different actions correspond to recommending various items. For instance, this is particularly relevant in sequential recommender systems [45]. Such systems are widely used in e-commerce and entertainment platforms, to suggest complementary or related products to users based on their interactions: when a user shows interest in a product by clicking on it (state S_1), the system may recommend the item (e.g., a smartphone). If the user purchases the item, they transition to a new state (S_2). This process repeats as the system dynamically recommends additional items, such as a smartwatch or headphones, based on the user's evolving engagement.

Beyond email marketing, this framework can be adapted for advertisements across other channels such as web push notifications, mobile app notifications, or even chatbot interactions. For example, in a mobile shopping app, the system could recommend various items via in-app notifications depending on the user's browsing behavior, purchase history, or demographic profile.

A key advantage of the CRMAB framework is its ability to model user behavior as a dynamic, evolving process. By capturing state transitions, it allows for personalized recommendations that adapt to changes in user preferences over time. This is particularly valuable in dynamic environments where user interests may shift rapidly, such as during seasonal sales or promotional campaigns.

7 Conclusion and Future Work

To our knowledge, we are the first to propose utilizing Whittle index-based Q-learning for CRMAB, and we are the first to propose an application of restless bandits for responsible email marketing. Our algorithm, QWIC-Fair, models implicit user feedback as state transitions in a context-augmented MDP to learn user interaction dynamics while ensuring equitable user selection through a fairness constraint. Experiments on both synthetic and real-world data showed that QWIC-Fair outperforms common email marketing approaches.

Our solution leverages context information by incorporating it into the Q-learning process, allowing the algorithm to adjust its actions based on both user states and contextual factors. While we used a tabular Q-learning method for initial validation, this approach effectively demonstrates the integration of context in decision-making. For larger context spaces, our future work will explore advanced methods such as function approximation to handle increased complexity. Additionally, we aim to incorporate multiple actions and conduct A/B testing to evaluate the algorithm's impact on customer behavior.

Acknowledgments. This research was conducted in collaboration between INRIA and NSP / Smartprofile (www.smartp.com), with support from the ANRT (Association Nationale de la Recherche et de la Technologie). Special thanks to our colleague Hervé Baïle from NSP / Smartprofile for his help throughout this project.

References

1. Akbarzadeh, N., Mahajan, A.: Restless bandits with controlled restarts: Indexability and computation of whittle index. In: 58th IEEE Conference on Decision and Control, CDC 2019, Nice, France, December 11-13, 2019. pp. 7294–7300. IEEE (2019), https://doi.org/10.1109/CDC40024.2019.9029182
2. Avrachenkov, K.E., Borkar, V.S.: Whittle index policy for crawling ephemeral content. IEEE Trans. Control. Netw. Syst. **5**(1), 446–455 (2018), https://doi.org/10.1109/TCNS.2016.2619066
3. Avrachenkov, K.E., Borkar, V.S.: Whittle index based q-learning for restless bandits with average reward. Autom. **139**, 110186 (2022), https://doi.org/10.1016/j.automatica.2022.110186
4. Bennett, J., Lanning, S.: The netflix prize. In: Proceedings of KDD Cup and Workshop. vol. 2007, p. 35. New York (2007), https://api.semanticscholar.org/CorpusID:1978078
5. Berners-Lee, M.: How bad are bananas?: The carbon footprint of everything. Profile Books (2020)
6. Biswas, A., Aggarwal, G., Varakantham, P., Tambe, M.: Learn to intervene: an adaptive learning policy for restless bandits in application to preventive healthcare. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021. pp. 4039–4046. ijcai.org (2021), https://doi.org/10.24963/ijcai.2021/556
7. Biswas, A., Killian, J.A., Diaz, P.R., Ghosh, S., Tambe, M.: Fairness for workers who pull the arms: An index based policy for allocation of restless bandit tasks pp. 1321–1328 (2023), https://dl.acm.org/doi/10.5555/3545946.3598779
8. Borkar, V.S.: Stochastic approximation with two time scales. Systems & Control Letters **29**(5), 291–294 (1997), https://www.sciencedirect.com/science/article/pii/S0167691197900153
9. Bouneffouf, D., Rish, I., Aggarwal, C.: Survey on applications of multi-armed and contextual bandits. In: 2020 IEEE congress on evolutionary computation (CEC). pp. 1–8. IEEE (2020)
10. Chapelle, O., Li, L.: An empirical evaluation of thompson sampling. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.)

Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain. pp. 2249–2257 (2011), https://dl.acm.org/doi/10.5555/2986459.2986710
11. Chen, X., Hou, I.: Contextual restless multi-armed bandits with application to demand response decision-making. CoRR **abs/2403.15640** (2024), https://doi.org/10.48550/arXiv.2403.15640
12. Chittenden, L., Rettie, R.: An evaluation of e-mail marketing and factors affecting response. Journal of Targeting, Measurement and Analysis for Marketing **11**, 203–217 (2003), https://doi.org/10.1057/palgrave.jt.5740078
13. Ekstrand, M.D., Riedl, J., Konstan, J.A.: Collaborative filtering recommender systems. Found. Trends Hum. Comput. Interact. **4**(2), 175–243 (2011), https://doi.org/10.1561/1100000009
14. GDPR: General data protection regulation (2025), https://gdpr-info.eu/., accessed: 16 January 2025
15. Glazebrook, K.D., Mitchell, H., Ansell, P.: Index policies for the maintenance of a collection of machines by a set of repairmen. European Journal of Operational Research **165**(1), 267–284 (2005)
16. Gomez-Uribe, C.A., Hunt, N.: The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manag. Inf. Syst. **6**(4), 13:1–13:19 (2016), https://doi.org/10.1145/2843948
17. Hallak, A., Castro, D.D., Mannor, S.: Contextual markov decision processes. CoRR **abs/1502.02259** (2015), http://arxiv.org/abs/1502.02259
18. Hemker, S., Herrando, C., Constantinides, E.: The transformation of data marketing: how an ethical lens on consumer data collection shapes the future of marketing. Sustainability **13**(20), 11208 (2021)
19. Herlihy, C., Prins, A., Srinivasan, A., Dickerson, J.P.: Planning to fairly allocate: Probabilistic fairness in the restless bandit setting. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 732–740 (2023), https://dl.acm.org/doi/10.1145/3580305.3599467
20. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy. pp. 263–272. IEEE Computer Society (2008), https://doi.org/10.1109/ICDM.2008.22
21. INSEE: National institute of statistics and economic studies (2025), https://www.insee.fr/en/accueil, accessed: 16 January 2025
22. Jenkins, S.: The truth about email marketing. FT Press (2008)
23. (source: Kaspersky), S.: Global spam volume as percentage of total e-mail traffic from 2011 to 2023 (2025), https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/, accessed: 16 January 2025
24. Ko, H., Lee, S., Park, Y., Choi, A.: A survey of recommendation systems: recommendation models, techniques, and application fields. Electronics **11**(1), 141 (2022)
25. Lattimore, T., Szepesvári, C.: Bandit algorithms. Cambridge University Press (2020)
26. Li, D., Varakantham, P.: Avoiding starvation of arms in restless multi-armed bandits p. 1303–1311 (2023), https://dl.acm.org/doi/10.5555/3545946.3598777
27. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: Rappa, M., Jones, P., Freire, J., Chakrabarti, S. (eds.) Proceedings of the 19th International Conference on World

Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010. pp. 661–670. ACM (2010), https://doi.org/10.1145/1772690.1772758
28. Lian, Z., Nath, R.: A conceptual model for effective email marketing. In: 17th International Conference on Computer and Information Technology, ICCIT 2014. pp. 250–256. IEEE (2014), https://doi.org/10.1109/ICCITechn.2014.7073103
29. Liang, B., Xu, L., Taneja, A., Tambe, M., Janson, L.: A bayesian approach to online learning for contextual restless bandits with applications to public health. CoRR **abs/2402.04933** (2024), https://doi.org/10.48550/arXiv.2402.04933
30. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Rich, C., Yang, Q., Cavazza, M., Zhou, M.X. (eds.) Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010. pp. 31–40. ACM (2010), https://doi.org/10.1145/1719970.1719976
31. Meshram, R., Manjunath, D., Gopalan, A.: A restless bandit with no observable states for recommendation systems and communication link scheduling. In: 54th IEEE Conference on Decision and Control, CDC 2015, Osaka, Japan, December 15-18, 2015. pp. 7820–7825. IEEE (2015), https://doi.org/10.1109/CDC.2015.7403456
32. Niño-Mora, J.: Restless bandits, partial conservation laws and indexability. Advances in Applied Probability **33**(1), 76–98 (2001), http://www.jstor.org/stable/1428442
33. Papadimitriou, C.H., Tsitsiklis, J.N.: The complexity of optimal queueing network control. In: Proceedings of the Ninth Annual Structure in Complexity Theory Conference, Amsterdam, The Netherlands, June 28 - July 1, 1994. pp. 318–322. IEEE Computer Society (1994), https://doi.org/10.1109/SCT.1994.315792
34. Pazzani, M.J., Billsus, D.: Content-based recommendation systems. In: The adaptive web: methods and strategies of web personalization, pp. 325–341. Springer (2007)
35. Prins, A., Mate, A., Killian, J.A., Abebe, R., Tambe, M.: Incorporating healthcare motivated constraints in restless bandit based resource allocation. NeurIPS 2020 Workshops: Challenges of Real World Reinforcement Learning, Machine Learning in Public Health (Best Lightning Paper), Machine Learning for Health (Best on Theme), Machine Learning for the Developing World (2020)
36. Project, T.C.L.: The carbon cost of an email (2022), https://carbonliteracy.com/the-carbon-cost-of-an-email/, accessed: 16 January 2025
37. (source: Radicati), S.: Number of sent and received e-mails per day worldwide from 2017 to 2026 (2022), https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/, accessed: 16 January 2025
38. REES46: Rees46 cdp for ecommerce (2025), https://rees46.com/en/cdp, accessed: 22 January 2025
39. REES46-Datasets: Direct messaging campaigns dataset overview (2025), https://www.kaggle.com/code/mkechinov/direct-messaging-campaigns-dataset-overview/input?select=messages-demo.csv, accessed: 22 January 2025
40. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM **40**(3), 56–58 (1997)
41. Singh, G., Singh, H., Shriwastav, S.: Improving email marketing campaign success rate using personalization. Advances in Analytics and Applications pp. 77–83 (2019)
42. Smartprofile: Smartprofile (2025), https://smartp.com/, accessed: 10 June 2025

43. Smith, B., Linden, G.: Two decades of recommender systems at amazon.com. IEEE Internet Comput. **21**(3), 12–18 (2017), https://doi.org/10.1109/MIC.2017.72
44. Taylor, B.: Sender reputation in a large webmail service. In: CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA (2006), http://www.ceas.cc/2006/listabs.html#19.pdf
45. Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q.Z., Orgun, M.A.: Sequential recommender systems: Challenges, progress and prospects. In: Kraus, S. (ed.) Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. pp. 6332–6338. ijcai.org (2019), https://doi.org/10.24963/ijcai.2019/883
46. Wang, S., Xiong, G., Li, J.: Online restless multi-armed bandits with long-term fairness constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15616–15624 (2024), https://dl.acm.org/doi/10.1609/aaai.v38i14.29489
47. Wang, S., Huang, L., Lui, J.C.S.: Restless-ucb, an efficient and low-complexity algorithm for online restless bandits (2020), https://proceedings.neurips.cc/paper/2020/hash/89ae0fe22c47d374bc9350ef99e01685-Abstract.html
48. Watkins, C.J.C.H., Dayan, P.: Technical note q-learning. Mach. Learn. **8**, 279–292 (1992), https://doi.org/10.1007/BF00992698
49. Weber, R.R., Weiss, G.: On an index policy for restless bandits. Journal of Applied Probability **27**(3), 637–648 (1990), http://www.jstor.org/stable/3214547
50. White, J.: Bandit algorithms for website optimization. O'Reilly (2013)
51. Whittle, P.: Restless bandits: Activity allocation in a changing world. Journal of applied probability **25**(A), 287–298 (1988)
52. Xiong, G., Wang, S., Yan, G., Li, J.: Reinforcement learning for dynamic dimensioning of cloud caches: A restless bandit approach. IEEE/ACM Transactions on Networking (2023)