Real-Time Context-Aware Reward Allocation in Mobile Loyalty Networks: Architecture, Budgeted Bandit Optimization, and Reproducible Evaluation

Naresh Alapati
Principal Cloud Architect
Walmart, Inc.
Bentonville, AR, USA
narencachy@gmail.com

Koteswararao Nallabothu
Lead Software Engineer
Lowe's, Inc.
Charlotte, NC, USA
nallabothukoteswar@gmail.com

Abstract—The paper proposes a closed-loop mobile loyalty architecture and a Spatial-Temporal Budgeted LinUCB (ST-BLinUCB) algorithm for budget-constrained allocation of promotions in streaming contexts. Unlike heuristic scoring, the policy defines the incentive arms, context vector, exploration with a linear upper-confidence, remaining budget state, and an adaptive budget pacing resource price. To evaluate the method without putting the learner in a self-generated data environment, an intentionally misspecified simulation is designed where the conversion rate depends nonlinearly on the observed and unobserved variables including latent loyalty, price sensitivity, competitor, and weather shock, plus stochastic mobility. In thirty paired simulations, one hundred thousand user profiles per simulation and thirty-three thousand sequential events per profile are generated to yield one million interactions per method. The proposed ST-BLinUCB policy yields 13.82% conversion rate and 0.93% incremental conversion rate with $5.78 incremental sales per dollar of discount. It outperforms significantly the vanilla LinUCB and the budgeted LinUCB without spatial context in terms of incremental conversion rate, but is not statistically different from geofencing in conversion rate, despite using 54.3% fewer offers than the latter method. Ablation and sensitivity experiments tease out the effect of spatial context, temporal context, budgeting, spatial decay, budget size of the campaign, and the incentive strength. The architecture and simulation codes are included in the paper, while the end-to-end latency in production is not part of the current experiment.

Index Terms—geofencing, loyalty systems, real-time stream processing, location-aware recommendation, contextual bandits, budget constraints.

I. Introduction

Loyalty program strategies are becoming ever more capable of exploiting information about location, clicks, and transactions at a velocity that allows decisions to be made before it becomes too late for the intent of the customer to be fulfilled. Nonetheless, many systems continue to use methods such as fixed-tier programs, static geofence coupons, or batch-revised rules which overlook the contextual aspects or give incentives to customers who are already likely to make purchases. What is needed is a technically feasible solution that will serve two goals at once: decide on what incentive should be given in the current context, and keep the accumulated cost under control in a given budget.

The exploration-exploitation trade-off problem can be handled with the help of the approach of contextual bandits [3]–[8] while finite resource availability can be described using the framework of knapsack-constrained bandits [9]–[11]. The role of the travel distance and geographic proximity in location-aware recommendation was demonstrated in [13] and [14]. Contextual bandits, resource-constrained online learning, and location-based recommendations are all established fields in their own right. As such, the contributions made in this paper do not involve the development of a new theory of contextual bandits or a new regret bound. The algorithmic contribution of this paper is the specific budgeted LinUCB implementation for the mobile loyalty problem, which differs from standard LinUCB in three ways: (i) action-specific feature mapping that combines the spatio-temporal context of the customer with the strength of the incentive offered; (ii) a hard feasibility constraint that ensures that incentives whose maximum allowable redemption amount exceeds the remaining campaign budget are never selected; and (iii) an adaptive resource price whose state is updated based on actual promotional consumption. Scientifically, the contribution includes a purposefully misspecified nonlinear evaluation environment, counterfactual experiments, several baseline algorithms, ablation tests, and parameter sensitivity analysis. As such, the novelty claim in this case pertains only to this particular application-specific algorithm and its evaluation.

The motivation behind the proposed research stems from the architecture which incorporates client telemetry, live data integration, space assessment, dynamic reward decisioning, and automation in delivery. Some aspects of the said architecture have been backed by pending patent in India as patent number IN202441089829 [17]. However, the current research is focused on developing and testing of the decision-making algorithm in a scientific manner.

A. Contributions

- We set up real-time loyalty reward allocation as a contextual bandit decision problem with four monetary reward arms, spatial-temporal customer context, and a well-defined finite promotional budget. Our approach improves the standard LinUCB algorithm by adding a budget constraint and a resource pricing term to guide action selection.
- We use a common action context representation wherein normalized monetary incentive interacts with a nine-dimensional context vector. The above approach facilitates the transfer of relevant information across four reward magnitudes and action-specific prediction through an online ridge estimator.

- We implement a simulation setting where the decision-making algorithm does not have access to any hidden customer loyalty, price sensitivity, competitive actions, weather shocks, heterogeneous responses of individuals, and outcome noise. Thus, the learner algorithm is isolated from the data-generating process instead of assuming the model inside the simulator.
- All policies are compared over 30 pairs of simulations, regenerated customer data, common random samples, 95% confidence intervals, paired statistical test, baseline comparisons, spatial and temporal ablation, and parameter sensitivity analysis on campaign budget, spatial decay constant, and reward magnitudes.
- The event-based reference architecture distinguishes reward decision making process from telemetry ingestion, business rules, offers delivery, campaign accounting and feedback processing processes.

II. Related work and research gap

A. Context-Aware Recommendation and Bandit Learning

The importance of context as an important aspect of recommendation beyond static user-item preference data has been noted for quite some time in recommender-systems literature [1, 2]. The notion of personalized news selection as a contextual bandit problem was considered by Li et al., who also proposed LinUCB-style online adjustment method [3]. Confidence bound-based techniques were pioneered by Chu et al. and Abbasi-Yadkori et al. in the setting of linear contextual bandits and stochastic bandits, respectively [4, 5]; Thompson sampling can be used as an alternate approach to exploration based on Bayesian sampling with extensive theoretical and empirical evidence [6], [7]. An oracle-efficient contextual-bandit algorithm for general policy classes was suggested by Agarwal et al. [8].

B. Budget-Constrained Sequential Decision Making

"Bandits with Knapsacks" is the mathematical formulation of reward maximization in an online setting, taking into account resource consumption per action [9]. The more recent work "Resourceful Contextual Bandits" extends this formulation to a context-dependent policy [10]; and Agrawal, Devanur, and Li further refine the formulation, proposing a computationally efficient contextual-bandit model with global knapsack constraints [11]. It is the theoretical advances in these papers that inspire the explicit consideration of the budget state and resource price in the current algorithm. This implementation emphasizes ease of computation over generality in the most theoretically advanced algorithms; it combines one linear UCB estimator and one promotional budget state.

In recent years, resource-constrained contextual learning has been extended to go beyond the bandits-with-knapsacks setting. Slivkins et al. introduce a new contextual-bandit algorithm that includes general packing and covering constraints using a modular Lagrangian approach, showing how resource consumption can be included in the process of online policy optimization. In another study, Zhao and Yang solve the problem of budget-limited recommendations by formulating it using a constrained contextual-bandit algorithm that uses LinUCB reward prediction and linear programming. All these works show that modeling resource consumption is crucial when exploration and recommendation come at a cost. This work is more focused, since it uses a light LinUCB controller for a specific budget limitation on promotions [19], [20].

In recent promotion-recommendation studies, there has been even more clarity about distinguishing between total response and incremental response. According to Sun et al., the process of recommending incentives under budget constraints using uplift modeling focuses on how one is trying to increase the effect of the incentive as opposed to the overall chance of conversion. An important insight gained from this literature that applies in our case includes the fact that ST-BLinUCB models total response, whereas increment conversion is just a measure for performance evaluation [21].

The delay in feedback represents an established problem in online recommendation and marketing. In Vernade et al., linear bandits are studied for the case of stochastic delayed feedback, even in situations where purchases take place after recommendations have been made. The study by Blanchet, Xu, and Zhou applies this concept to the context of generalized linear contextual bandits. Meanwhile, the Open Bandit Dataset and Pipeline provide a dataset and tools for reproducible off-policy evaluation using real-world bandit logging data. Altogether, these papers identify two significant extensions of our current work, namely the explicit handling of delayed conversion feedback and evaluation of policy candidates with real-world interaction logging data [22]-[24].

C. Location-Aware Recommendation and Streaming Systems

Models for LARS take account of travel penalties in location-aware recommendations [13], and Ye et al. prove that the geographical aspect improves POI recommendations [14]. All this makes one argue for an inclusion of distance as such, not only the binary indication of being close or far away. In regard to system design, Kafka has been proven to be capable of distributed event ingestion at scale [15], and D-Streams can be used for fault-tolerant stream processing [16]. In this sense, our solution takes into account both approaches mentioned above.

D. Research Gap

Previous literature provides the groundwork for contexts in exploration, constrained online learning, budgeted recommendations, incentives using uplift models, delayed feedback, and off-policy evaluation. In this way, the gap that our work fills is specifically focused. We explore the implementation of a mobile-loyalty decision policy wherein the size of financial incentives, the spatial-temporal context of use, budget of the campaign, feedback, and event-driven computing architecture are analyzed together through a repeatable counterfactual simulation.

This new policy should not be considered a replacement of theoretical frameworks concerning general constrained bandits and uplift learning, but an applied variant of LinUCB policy that accounts for the limited budget. The goal is to find out if there are tangible gains achieved through a combination of state

information and budget pacing as compared to a baseline of static tiers and geofencing.

III. Closed-loop system architecture

The architecture consists of six interoperable components. The mobile component generates events related to location and interactions that are consented to, and shows eligible offers. The integration layer takes care of normalization of asynchronous messages about app events, location, and point-of-sale events. The spatial evaluation component manages geofences and calculates the proximity to stores. The Dynamic Reward Engine creates context information, runs ST-BLinUCB and produces an action selection. The delivery layer ensures the administrative eligibility of actions and starts offers delivery. Lastly, feedback from conversions and redemptions modifies the campaign book and the online learner. Business analytics is not part of the decision-making process but used for reports and policy changes.

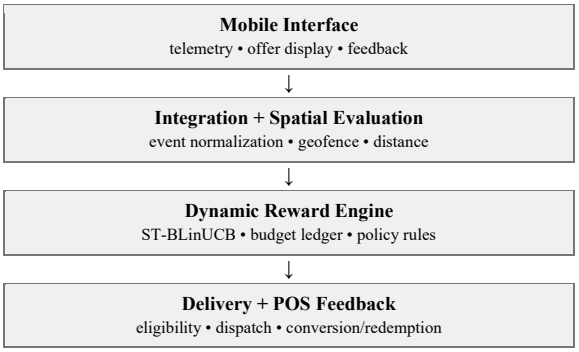


Fig. 1. Closed-loop event flow from mobile telemetry to budgeted reward selection and verified feedback.

Due to privacy and operational safety concerns, the learner does not have to store the raw location traces in high frequency. The decision service can take advantage of the proximity and dwell features that have been extracted from the traces, but the PII and the coordinates continue to be managed by the enterprise data layer.

IV. Spatial-temporal budgeted linucb

A. Actions, Context, and Budget State

At time t, the policy looks at the current context and picks one of four micro-incentive arms, $\mathcal{A}$ = {0, 2, 5, 8} USD. Arm 0 indicates that there is no offer. Fixed-dollar amounts make the minimum budget requirement of the campaign clear before the campaign execution starts. Let B0 be the total campaign budget and Bt the remaining budget before event t.

The context vector g at time t includes nine elements: the constant term; the normalized spend velocity; the normalized number of visits; product category preference; the decayed proximity function exp(-λ dt); whether the visit is done off-peak; dwell intention; low spend * proximity interaction; and product category preference * off-peak interaction. Here dt is distance in normalized units of 500 m radius and λ controls the decay rate.

$$g_t = \left[1,\ s_t,\ v_t,\ f_t,\ e^{-\lambda d_t},\ \tau_t,\ h_t,\ (1-s_t)e^{-\lambda d_t},\ f_t\tau_t\right]^\top \quad (1)$$

In the case of $\mathcal{A}$, the value of q_a is computed as $a/a_m ax$, and the common linear function combines the context, the weighted context using the action, and the quadratic component of the action. Thus, for nine features of the context, we have a 19-dimentional vector.

$$\phi_t(a) = [g_t^\top,\ q_a g_t^\top,\ q_a^2]^\top, \qquad q_a = \frac{a}{a_{\max}} \quad (2)$$

The use of one 19-dimensional action-context representation is by choice. The features $q_a g_t$ and q_a^2 ensure that data obtained with one level of incentives is useful for estimating parameters for the others, rather than needing four completely separate models. This allows for fewer state components per arm to learn and sharing of statistics across arms at the beginning of the campaign. The drawback is a more limited response surface – the representation assumes that effects of the incentive can be modeled with shared linear context interaction together with a quadratic action feature. If action counts are highly unbalanced, this parameterization may have less expressivity than per-arm estimators. The current experiment tests only the shared representation, hence making statements regarding its comparison to separate per-arm models is an interesting future direction.

B. UCB Action Selection with Resource Pricing

The learner keeps a ridge regularized precision matrix, denoted V_t, along with its corresponding response vector, b_t. With initialization $\mathrm{V}_1 = \mathrm{I}_{19}$ and $\mathrm{b}_1 = 0$, the current parameter estimates are given as

$$\hat{\theta}_t = \mathrm{V}_t^{-1}\mathrm{b}_t.$$

The exploration bonus for LinUCB algorithm is defined as

$$\alpha\sqrt{\phi_t(a)^\top \mathbf{V}_t^{-1}\phi_t(a)}.$$

$$U_t(a) = \hat{\boldsymbol{\theta}}_t^\top \phi_t(a) + \alpha\sqrt{\phi_t(a)^\top \mathbf{V}_t^{-1}\phi_t(a)} \quad (3)$$

$$S_t(a) = U_t(a) - \mu_t\left(\frac{a}{a_{\max}}\right) \quad (4)$$

Actions that have their maximum redeemable liability below the budget remaining become possible actions. Feasible action set at period t is defined as

$$\mathcal{F}_t = \{a \in \mathcal{A} : a \le B_t\}$$

Here, the policy makes its selection by choosing the highest ranking action from $\mathcal{F}_t$. And then the policy picks the highest-scoring arm among the eligible set. The conservative filtering process guarantees that $B_t \ge 0$ because the actual discounted cost $c_t = a_t y_t$ can never be more than a_t.

$$a_t = \arg\max_{a \in \mathcal{F}_t} S_t(a) \quad (5)$$

C. Learning and Budget Updates

After the action a_t is executed, the system observes the change $y_t \in \{0,1\}$. Only the chosen action will be used to update the algorithm online; no information about other actions' outcomes will be revealed to the learner. The common ridge state, the remaining budget of the campaign, and the adaptive resource price state are updated accordingly, where η is the step size of resource price adaptation and $\rho = B_0/T$ is the target average promotional cost per event.

$$\mathbf{V}_{t+1} = \mathbf{V}_t + \phi_t(a_t)\phi_t(a_t)^\top, \qquad \mathbf{b}_{t+1} = \mathbf{b}_t + y_t\phi_t(a_t) \quad (6)$$

$$B_{t+1} = B_t - a_t y_t \quad (7)$$

$$\mu_{t+1} = \max\left\{0,\ \mu_t + \eta\frac{a_t y_t - \rho}{a_{\max}}\right\} \quad (8)$$

For each outcome observed, the policy modifies the selected action context observation in the joint ridge state $(\mathbf{V}_t, \mathbf{b}_t)$, residual campaign budget B_t , and adaptive resource price state μ_t . Interpretation of the resource-price controller: The resource-price term in equation (4) serves as an adaptive budget pacing controller rather than as the exact Lagrangian for the expected-cost constrained contextual bandit problem. In the action selection process, the score penalizes the normalized nominal incentive size $a/a_{\max}$, which is known before the offer is accepted and is used as a conservative estimate of the possible liability from promotion. On the other hand, the update in equation (8) depends on the actual redeemed cost $a_t y_t$ because the campaign expense is paid out only in case of incentive acceptance after conversion in the model being evaluated. Hence, the two values play different roles: the nominal incentive value discourages actions with potential high costs, whereas the actual redeemed value updates the pacing state. Hard constraints on the budget are imposed independently of the resource-price controller through the feasible-action filter. A direct pricing of the expected redemption cost would give rise to a more natural Lagrangian resource constraint and is left as future work.

D. Decision objective and Incremental-effect Interpretation

The online learner maximizes an upper confidence bound on the total conversion response that depends on the observed action-context representation. The online learner does not estimate the incremental effect of an incentive over the zero-offer action. In contrast, the Incremental Conversion Rate (ICR) is a metric that can be computed in the synthetic environment since the pair of potential outcomes $y_t(a)$ and $y_t(0)$ can be drawn from the same latent event realization. Therefore, the proposed ST-BLinUCB policy cannot be considered an uplift-maximizing policy [21].

There is an important practical implication here. The total response policy may favor customers who have a high baseline probability of conversion even though the incremental effect of an incentive is relatively small. The current evaluation sheds light on this aspect by reporting the conversion rate (CR), ICR, efficiency of promotional spend, offer rate, and budget consumption simultaneously, but not by evaluating the policy in terms of conversion rate only. An obvious extension is to learn response models specific to each treatment and select the incentives according to the estimated incremental response over the zero-offer baseline. Recent work on incentivizing via uplift modeling under budget constraint is relevant for this task.

Algorithm 1. ST-BLinUCB Online Decision Procedure	
Input	Input $\mathcal{A}$={0,2,5,8}, B_0, exploration α, resource-price adaptation step η, and spatial decay λ.
1	Initialize $\mathrm{V}_1 = \mathrm{I}_{19}$, $\mathrm{b}_1 = 0$, $\mu_1 = 0$, and $B_1 = B_0$.
2	For event $t = 1, \ldots, T$, observe g_t and construct $\phi_t(a)$ for every $a \in A$.
3	Form $\mathcal{F}_t = \{a \in \mathcal{A}: a \le B_t\}$; compute $\hat{\theta}_t = \mathrm{V}_t^{-1}\mathrm{b}_t$.
4	For each $a \in \mathcal{F}_t$, compute $U_t(a)$ from (3) and $S_t(a)$ from (4).
5	Select $a_t = \arg\max_{a \in \mathcal{F}_t} S_t(a)$; dispatch only when $a_t > 0$.
6	Observe conversion y_t and redeemed cost $c_t = a_t y_t$.
7	Update V_{t+1}, b_{t+1}, B_{t+1}, and resource-price state μ_{t+1} using (6)-(8).
8	Continue until T events or until no positive-cost arm is feasible.
Output	Learned policy state, remaining budget, and campaign metrics.

E. Budget Feasibility and Computational Complexity

Proposition 1 (budget feasibility under synchronous outcome feedback). Under the assumption that the reward and redeemed cost for event t are known before policy processing on event $t + 1$, let $B1 = B0 \ge 0$ be the budget of the ST-BLinUCB algorithm where each arm has $at \le Bt$. In the case where $ct = atyt$ with $yt = 0 or yt = 1$, the ST-BLinUCB algorithm maintains $Bt \ge 0$. This proposition applies to the synchronous feedback model used in the present simulation. It does not establish production budget safety when multiple previously issued offers remain unresolved while new decisions are made. Let $K = |\mathcal{A}|$ be the number of arms and d be the dimension of the feature space. Testing all the feasible arms involves making K confidence intervals. To avoid matrix inversion after each event, we keep $\mathbf{V}_t^{-1}$ up-to-date via a rank-one Sherman-Morrison update. The computational cost is $O(Kd^2 + d^2)$ per event with $O(d^2)$ of learner state, and is independent of the number of previous events. In the used configuration, K = 4 and d = 19, leading to a small decision state. Thus, policy update becomes computationally feasible; nonetheless, this should not be considered as an alternative to empirical latency benchmarking.

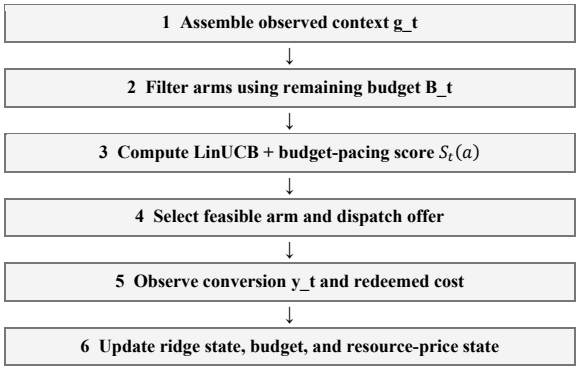


Fig. 2. Vector workflow for ST-BLinUCB. All objects are inline to avoid two-column overlap.

V. REPRODUCIBLE SIMULATION METHODOLOGY

A. Paired Trial Design

As there is no available public data source which allows for testing the combination of fine-grained geofence triggers, offer assignment, promotion cost, and conversion verification, the test is carried out within a transparent synthetic setting without claiming any causal effect in the real world. The main experiment consists of 30 paired runs with different seeds 1001–1030. For each run, 100,000 user profiles are created again and then 33,334 events are sampled in total to give 1,000,020 tested events for each method overall. Under a single seed, all methods share the same event sequence, latent shocks, basket value variance, and outcome draws.

For each run, the budget of the promotion campaign is equal to \$2,800 (or \$0.084 per event). For the configuration of the ST-BLinUCB algorithm, $\alpha = 0.35, \eta = 0.05, \lambda = 2.0$, unit ridge initialization is used, and arm set {0, 2, 5, 8} is applied. There is no hyperparameter tuning for individual baselines in the main experiment.

TABLE I. PRIMARY EXPERIMENTAL CONFIGURATION AND ST-BLINUCB HYPERPARAMETERS

Parameter	Primary value	Purpose
Paired trials	30	Statistical replication
Seeds	1001-1030	Exact reproducibility
Profiles per trial	100,000	Population diversity
Events per trial	33,334	Sequential decisions
Campaign budget	$2,800	Finite promotional resource
Incentive arms	$0, $2, $5, $8	Action space
UCB exploration α	0.35	Confidence-bound exploration
Resource-price step η	0.05	Budget pacing adaptation
Spatial decay λ	2.0	Proximity transformation
Context dimension	9	Observable customer-store state
Feature dimension	19	Shared action-context representation
Evaluation radius	500 m	Distance normalization

TABLE II. SYNTHETIC PROFILE AND EVENT DISTRIBUTIONS (below the table: "Hidden variables are never passed to the learning policy.")

Variable	Generation / role
Spend	LogNormal(3.4,0.65), log-transformed and clipped to [0,1].
Visits	NegativeBinomial(r=3,p=.45)/12, clipped to [0,1].
Affinity	Beta(2.4,2.0).
Price sensitivity	0.72(1-spend)+0.28 Beta(2.0,2.2); hidden from learner.
Loyalty	0.55 spend+0.45 Beta(2.3,1.9); hidden from learner.
Distance	68% near-store Beta(1.2,4.5), otherwise Beta(3.2,1.7).
Dwell	Nonlinear sigmoid of loyalty, affinity, distance, and N(0,.6).
External shocks	Competitor Bernoulli(.08), weather Bernoulli(.10), both hidden.
Outcome noise	Latent user N(0,.35) plus event N(0,.18).
Basket value	LogNormal around $48, clipped $12-$160, with lognormal event noise.

B. Non-Circular Conversion Generator

The primary validity check is the distinction between the learner and the data-generating process (DGP). ST-BLinUCB learns from spend, visits, affinity, transformed distance, time, and dwell features. It cannot learn latent loyalty, price sensitivity, competitor actions, weather, individual user response, and noise in the outcome. DGP shows nonlinear dependencies with respect to distance and incentive impact, while the learner uses a linear action-context approximation. Hence, the policy cannot just learn the generator by learning the same equation used for generating the outcome.

For a no-offer action, the latent conversion probability $z_t(0)$ is a combination of observable and unobservable variables. For an offer a, with $q = a/8$, the impact of the treatment is a function of unobservable price sensitivity p_t, distance, off-peak time, affinity, loyalty, and the squared cost due to the incentive:

$$z_t(0) = -2.85 + 0.55s + 0.40v + 0.72f + 0.50l + 0.45h - 1.10d^{1.35} - 0.22\tau_t + 0.32fl - 0.55c - 0.35w + \xi + \varepsilon_t \quad (9)$$

$$z_t(a) - z_t(0) = q[4.20p_t(1-d_t)^2 + 1.20p_t\tau_t + 0.40f_t - 0.70\ell_t] - 0.50q^2, \qquad q = \frac{a}{8} \quad (10)$$

Uniform(0,1) distribution U_t generates the potential outcome using $y_t(a) = 1U_t < sigmoid\big(z_t(a)\big)$. Using the same U_t for all the treatments during the trial guarantees counterfactual consistency for comparison. Note that the difference $y_t = y_t(a) - y_t(0)$, and not $max0, y_t$, because the effect produced by the intervention can be zero or even negative when the generating mechanism predicts a smaller probability of conversion due to that treatment.

C. Metrics, Baselines, and Statistical Tests

Conversion rate (CR) represents the fraction of events that culminate in a sale. Incremental conversion rate (ICR) is defined as the average of $y_t(a) - y_t(0)$. Promotional spending efficiency (PSE) is defined as incremental sales divided by the redeemed discount cost. The offer rate measures the level of notification exposure, while budget usage equals the realized cost divided by B0. For each policy, we provide means and a two-sided 95% t-interval across 30 runs. We compare the proposed policy pairwise to specified baselines via one-sided Wilcoxon signed-rank test due to the fact that all policies use the identical seed-level stream of events.

- Static tiering policy: an offer of $5 for spend over 0.72, an offer of $2 for spend above 0.45, and no offer otherwise; budget feasible.
- Heuristic geofencing: a fixed $5 offer within 125 meters from a 500-meter radius of the evaluation site; budget feasible.
- Vanilla LinUCB: This approach combines full spatiotemporal context with UCB exploration without using adaptive budget pacing resource price term; however, hard budget feasibility check is still active to avoid budget breaches.
- Budgeted LinUCB: Location and time blind adaptive budget-pacing control, used in order to evaluate the role of location and time in budget pacing.
- Ablations: The suggested policy proposal is offered without the context of geography, without the context of time, and without adaptive pricing of resources and budget constraints.

VI. Results and Ablation Analysis

Table III shows the main experiment, done through a pairing of experiments. While the primary finding from the results is not the supremacy of ST-BLinUCB algorithm over all heuristics in all metrics, it provides significant gains against learning algorithms where the pricing of resources and/or spatial-temporal context is absent, whereas a geofence strategy is still statistically indistinguishable from CR and PSE. However, the proposed policy achieves the same performance with significantly fewer offer exposures.

TABLE III. PRIMARY PAIRED EVALUATION OVER 30 TRIALS (below: "PSE is incremental sales per redeemed discount dollar. The no-budget ablation is intentionally infeasible and spends 24.1× the campaign budget.")

Method	CR % (95% CI)	ICR % (95% CI)	PSE $/$ (95% CI)	Offer %	Budget use %
Static tiers	13.33 [13.25, 13.41]	0.44 [0.43, 0.45]	2.70 [2.61, 2.78]	17.02	100.0
Fixed geofence	13.81 [13.73, 13.89]	0.91 [0.90, 0.92]	5.67 [5.58, 5.76]	4.90	100.0
LinUCB	13.56 [13.48, 13.64]	0.67 [0.66, 0.68]	4.19 [4.10, 4.27]	3.35	100.0
Budgeted LinUCB (no space/time)	13.72 [13.63, 13.81]	0.83 [0.80, 0.86]	5.11 [4.92, 5.30]	3.14	100.0
ST-BLinUCB (proposed)	**13.82 [13.74, 13.91]**	**0.93 [0.90, 0.96]**	**5.78 [5.56, 6.00]**	**2.24**	**100.0**
No-budget ablation	29.04 [28.94, 29.14]	16.15 [16.05, 16.25]	4.15 [4.13, 4.17]	80.64	2411.7

TABLE IV. PAIRED ONE-SIDED WILCOXON TESTS ON INCREMENTAL CONVERSION (below: "Statistical ties are reported as ties rather than framed as wins.")

Paired comparison	Δ ICR (pp)	ICR p	Operational interpretation
Proposed > LinUCB	+0.260	<1e-6	Significant
Proposed > Budgeted LinUCB	+0.105	0.0000	Significant
Proposed > Static tiers	+0.492	<1e-6	Significant
Proposed > Fixed geofence	+0.017	0.3606	Not significant
Proposed > No spatial	+0.056	0.0052	Ablation significant
Proposed > No temporal	+0.031	0.0266	Small / exploratory

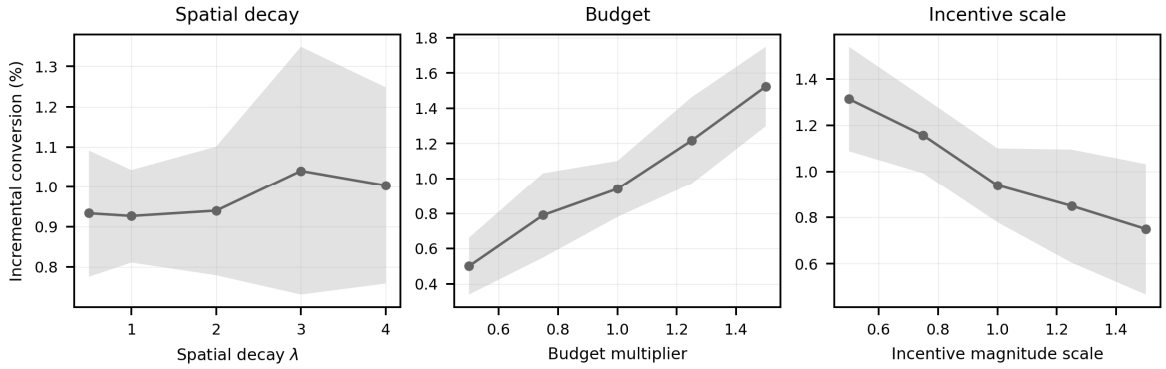


Fig. 3. Sensitivity of incremental conversion to spatial decay λ, campaign budget, and incentive magnitude. Curves show means with 95% confidence intervals across five paired seeds per setting.

A. Primary Comparison

In case of ST-BLinUCB, we obtain 13.824% (95% CI: 13.737–13.912) average conversion rate (CR) along with Incremental Conversion Rate (ICR) of 0.930% (0.897–0.964) and Promotional Spend Efficiency (PSE) amounting to $5.78 (interval: $5.56–$6.00). In comparison with regular LinUCB, ICR advantage is estimated at 0.260 percentage points ($p - value < 10^{-}6$), while PSE grows up to $1.59 for each discounted dollar ($p < 10^{-}9$). For Budgeted LinUCB without spatial/temporal variables, the difference in ICR amounts to 0.105 points ($p - value = 1.18 \times 10^{-}5$) and the growth of PSE is equal to $0.67 ($p - value = 4.16 \times 10^{-}7$). These results demonstrate that the benefits associated with our proposal come from the combination of resource pricing and contextual data, not just hard budget feasibility alone.

In regard to the fixed geofence, average CR is 13.807%, and average PSE equals $5.67; although the numbers are slightly lower, the paired difference is not statistically significant (ICR p-value=0.361; PSE p-value=0.516). The important difference here concerns exposure, namely, the rate of nonzero offers generated by the proposed approach is 2.24%, compared to 4.90% in case of geofencing, which implies a decrease of 54.3%.

B. Ablation Findings

The deletion of spatial attributes brings about a decrease in ICR from 0.930% to 0.874% and of PSE from $5.78 to $5.43; the difference between ICRs in paired situations is statistically significant (p=0.0052). Without temporal context, the ICR drops to 0.899% ($p = 0.0266 unadjusted$), while the difference in PSE is less prominent ($p = 0.062$). Hence, the findings show that the distance is the more robust factor in the presented simulation, while the temporal signal must be considered as a secondary one.

The "no-budget" ablation is intentionally impossible. Without the resource pricing and budget feasibility filters, the conversion rate of the campaign without budget will be quite higher but with promotional expenses that are beyond the set budget for the campaign. This example shows the reason why maximizing the conversions without any constraints cannot serve as an acceptable goal of promotion activity, as high conversion rates do not imply an admissible benchmark.

C. Sensitivity Analysis

In sensitivity analyses, one variable is altered while others are kept constant using five random seeds and 6,000 events per configuration. When λ is varied between 0.5 and 4.0, we observe overlapping confidence intervals and see that the mean Incremental Conversion Rate (ICR) is maximum for λ close to 3.0. An increased campaign budget increases ICR from 0.50% at 0.5× to 1.52% at 1.5× in a monotonic manner which agrees with our intuition because the higher the budget, the more interventions can be afforded. With incentive scaling analysis, we observe that for a fixed budget, the inverse efficiency relation holds as when we have a 0.5× incentive size, we have mean ICR of 1.31% and PSE of $8.40 but when we have a 1.5× incentive, we have mean ICR of 0.75% and PSE of $4.43 as the more incentives are used, the smaller budget remains. The sensitivity tests examine the robustness of the main results to changes in the campaign budget, the spatial decay rate, and the size of the incentive. The 30-trial base experiment measures the stochastic effects that are caused by independently generating different sets of customers and underlying shocks to the events. Therefore, this exercise provides evidence about sensitivity within a given model and does not imply robustness to any arbitrary customer response model or distribution. In the main experiment, the form and parameters of the nonlinear data generation process (DGP) are not changed. For testing robustness, different DGP formulations or even logged-policy datasets will need to be examined.

VII. Deployment, Reproducibility, and Threats to Validity

A. Deployment Considerations

A production rollout could do geofencing boundary detection either on the client or on the edge without affecting the reward policy and the campaign log on the decision service. A streaming bus could provide decoupling between mobile telemetry, point of sales feedback and business policy changes [15], while stateful processing could be implemented according to existing architectures of stream processing systems [16]. The eligibility check, consent, frequency caps and promotion rules should serve as deterministic guardrails outside of the learning policy in order to prevent exploration from breaking enterprise constraints.

Delayed redemption and outstanding liability: In a production mobile-loyalty system, offer dispatch and redemption at point of sale happen asynchronously. Thus, an extension to the production-safe approach should not only support outstanding promotional liability but also realized spend from the campaign budget. An overly conservative approach would be to hold the face value of each dispatched offer, consider available budget during decision time as the difference between campaign budget and realized spend and outstanding offers, and convert the offer on receiving redemption notification. The offer reservation helps in protecting the campaign budget from being oversubscribed due to outstanding offers. This is an extension to the system which has not been evaluated in any of the experiments done here [22].

The current research does not provide an estimate of production-level end-to-end latency. The previously stated sub-50 ms latency estimate is retracted since a synthetic simulation done in Python on one machine cannot serve as proof of mobile to stream to database to notification latency in production conditions. The correct way of measuring latency would involve a production deployment of Kafka/decision store/notification pipeline, controlled throughput levels, warm-up and repeated tests along with percentiles reporting failure and backpressure conditions [23].

B. Cold-Start and Operational Guardrails

The cold-start user lacks enough reliable data about their spending, visits, and affinity; therefore, the production environment should not assume that all the context data estimates are ready for action. One way to deal with this issue is to use campaign priors for the initialization of behavior variables, and limit initial offers to the no offer and low value arms until there is enough validated interaction data. Consent, customer eligibility, minimum margin requirement, campaign validity, frequency caps, and promotion laws are deterministic before and after decisions regardless of the confidence level in the model.

C. Reproducibility

The included Python code specifies all random distributions, seeds, hyperparameters, potential outcome construction, baselines, ablation, confidence intervals, and Wilcoxon test configurations. NumPy random generators are individually initialized at each trial. All the approaches use a common environment which is created for a specific seed value and there is no special data generator for an individual approach. Furthermore, the script generates the CSV files for each trial and sensitivity analysis reports which allow the creation of tables rather than copying from hand-crafted examples [18].

D. Threats to Validity

The main drawback is that all quantifiable results stem from synthetic data. Even though the simulator consciously disconnects the learner from the data generating process through hidden variables, nonlinear response functions, stochastic shocks, regeneration of the user population, and joint counterfactual outcome draws, this disconnection prevents circularity in the evaluation but does not ensure external validity. Numerical effect sizes, including CR, ICR, and PSE, are still contingent on the assumed synthetic population and response function.

Therefore, the results are better understood as an indicator of relative algorithm performance in a controlled stress test environment, rather than as an estimate of actual customer lift. To validate against real promotion logs, one would need proper randomization probabilities or logged propensities along with an off-policy evaluation technique. The Open Bandit Dataset shows how logged interaction data can be used for off-policy evaluation, but this dataset is specific to recommendation environments that are different from the geographical, monetary loyalty environment considered here [24].

Second, the policy makes decisions with one global promotion budget and fixed-dollar arms; actual promotions might have store-level, category-level, inventory-based, fairness, and frequency-based restrictions. Third, the UCB algorithm with linear responses is underfitting the non-linear response surface. Fourth, location-based sensing raises the issues of permissions, power consumption, indoor localization precision, and privacy that do not belong to the realm of the synthetic experiments. Finally, the tests are paired on 30 simulation seeds, not on independently observed real customers.

E. Intellectual Property Disclosure

A component of the system architecture and the integration process are covered by the pending Indian patent IN202441089829 “Dynamic Personalized Loyalty Program System with Real Time Data Integration and Location Based Promotions,” filed in November 2024 [17]. The equations, the design of replicable simulations, the baseline comparisons, and the statistical analysis used here make up the science validation component of the paper; patenting is not used as a proxy for peer review for proof of novelty.

VIII. Conclusion

In this paper, the authors develop and test an application-focused Spatial-Temporal Budgeted LinUCB formulation for reward-allocation. The learning policy comprises a shared action-context LinUCB estimator, which uses information from spatial-temporal customer states, filters hard budget-feasible actions, and implements a budget-pacing resource-price controller. In 30 paired experiments using synthetically generated data, the policy provides statistically significant gains in incremental conversion compared with plain LinUCB, reduced-budget LinUCB with less context, and tiering. Conversion of the learning policy is equivalent to the conversion of the fixed-geofence policy, whereas it shows significantly reduced offer exposure.

When interpreting results, one needs to account for the limitations of the study. Specifically, the optimization policy does not focus on uplift estimation but maximizes total conversions. The shared action-context framework implies a limited response structure of rewards. Proof of budget feasibility assumes synchronous outcome feedback, and the experiments use synthetically created data. Thus, the results show comparative behavior on the particular simulator but do not imply real-life lift or behavior. The future research might consider solving the following issues: estimating lift per arm, calculating redemption value of expected-resource, accounting for promotions of the past period, validating on randomized/propensity logged promotion data, distribution shifts test, and measuring production latency/scalability.

References

[1] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, 2005, doi: 10.1109/TKDE.2005.99.

[2] G. Adomavicius, B. Mobasher, F. Ricci, and A. Tuzhilin, “Context-aware recommender systems,” AI Magazine, vol. 32, no. 3, pp. 67–80, 2011, doi: 10.1609/aimag.v32i3.2364.

[3] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A contextual-bandit approach to personalized news article recommendation,” in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 661–670, doi: 10.1145/1772690.1772758.

[4] W. Chu, L. Li, L. Reyzin, and R. E. Schapire, “Contextual bandits with linear payoff functions,” in Proc. 14th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 15, 2011, pp. 208–214.

[5] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” in Advances in Neural Information Processing Systems 24, 2011, pp. 2312–2320.

[6] S. Agrawal and N. Goyal, “Thompson sampling for contextual bandits with linear payoffs,” in Proc. 30th Int. Conf. Machine Learning (ICML), PMLR, vol. 28, no. 3, 2013, pp. 127–135.

[7] O. Chapelle and L. Li, “An empirical evaluation of Thompson sampling,” in Advances in Neural Information Processing Systems 24, 2011, pp. 2249–2257.

[8] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire, “Taming the monster: A fast and simple algorithm for contextual bandits,” in Proc. 31st Int. Conf. Machine Learning (ICML), PMLR, vol. 32, no. 2, 2014, pp. 1638–1646.

[9] A. Badanidiyuru, R. Kleinberg, and A. Slivkins, “Bandits with knapsacks,” in Proc. IEEE 54th Annu. Symp. Foundations of Computer Science (FOCS), 2013, pp. 207–216.

[10] A. Badanidiyuru, J. Langford, and A. Slivkins, “Resourceful contextual bandits,” in Proc. 27th Conf. Learning Theory (COLT), PMLR, vol. 35, 2014, pp. 1109–1134.

[11] S. Agrawal, N. R. Devanur, and L. Li, “An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives,” in Proc. 29th Conf. Learning Theory (COLT), PMLR, vol. 49, 2016, pp. 4–18.

[12] L. Li, W. Chu, J. Langford, and X. Wang, “Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms,” in Proc. 4th ACM Int. Conf. Web Search and Data Mining (WSDM), 2011, pp. 297–306, doi: 10.1145/1935826.1935878.

[13] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel, “LARS: A location-aware recommender system,” in Proc. IEEE 28th Int. Conf. Data Engineering (ICDE), 2012, pp. 450–461, doi: 10.1109/ICDE.2012.54.

[14] M. Ye, P. Yin, W.-C. Lee, and D. L. Lee, “Exploiting geographical influence for collaborative point-of-interest recommendation,” in Proc. 34th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2011, pp. 325–334, doi: 10.1145/2009916.2009962.

[15] J. Kreps, N. Narkhede, and J. Rao, “Kafka: A distributed messaging system for log processing,” in Proc. NetDB Workshop, Athens, Greece, 2011.

[16] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica, “Discretized streams: Fault-tolerant streaming computation at scale,” in Proc. 24th ACM Symp. Operating Systems Principles (SOSP), 2013, pp. 423–438, doi: 10.1145/2517349.2522737.

[17] Indian Patent Office, “Dynamic Personalized Loyalty Program System with Real Time Data Integration and Location Based Promotions,” Patent IN202441089829, pending grant Nov. 2024.

[18] N. Alapati and K. Nallabothu, “Real-Time Context-Aware Reward Allocation in Mobile Loyalty Networks: Architecture, Budgeted Bandit Optimization, and Reproducible Evaluation”, Aug. 26, 2026, Zenodo. doi: 10.5281/zenodo.22104569.

[19] A. Slivkins, X. Zhou, K. A. Sankararaman, and D. J. Foster, “Contextual Bandits with Packing and Covering Constraints: A Modular Lagrangian Approach via Regression,” *Journal of Machine Learning Research*, vol. 25, no. 394, pp. 1–37, 2024.

[20] Y. Zhao and L. Yang, “Constrained contextual bandit algorithm for limited-budget recommendation system,” Engineering Applications of Artificial Intelligence, vol. 128, Art. no. 107558, 2024, doi: 10.1016/j.engappai.2023.107558.

[21] Z. Sun et al., “End-to-End Cost-Effective Incentive Recommendation under Budget Constraint with Uplift Modeling,” in *Proc. 18th ACM Conf. Recommender Systems (RecSys)*, 2024, pp. 560–569, doi: 10.1145/3640457.3688147.

[22] C. Vernade et al., “Linear Bandits with Stochastic Delayed Feedback,” in *Proc. ICML*, PMLR, vol. 119, pp. 9712–9721, 2020.

[23] J. Blanchet, R. Xu, and Z. Zhou, “Delay-Adaptive Learning in Generalized Linear Contextual Bandits,” *Mathematics of Operations Research*, vol. 49, no. 1, pp. 326–345, 2024.

[24] Y. Saito, S. Aihara, M. Matsutani, and Y. Narita, “Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation,” *NeurIPS Datasets and Benchmarks*, 2021.