Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential Decision-Making for Continuous Monitoring

Runzhe Wan*
Amazon
Seattle, USA
runzhe.wan@gmail.com

Yu Liu*
Amazon
Seattle, USA
liuyu0jlu@gmail.com

James McQueen
Amazon
Seattle, USA
jmcq@amazon.com

Doug Hains
Amazon
Seattle, USA
dhains@amazon.com

Rui Song
Amazon
Seattle, USA
ruisong@amazon.com

ABSTRACT

With the growing needs of online A/B testing to support the innovation in industry, the opportunity cost of running an experiment becomes non-negligible. Therefore, there is an increasing demand for an efficient continuous monitoring service that allows early stopping when appropriate. Classic statistical methods focus on hypothesis testing and are mostly developed for traditional high-stake problems such as clinical trials, while experiments at online service companies typically have very different features and focuses. Motivated by the real needs, in this paper, we introduce a novel framework that we developed in Amazon to maximize customer experience and control opportunity cost. We formulate the problem as a Bayesian optimal sequential decision making problem that has a unified utility function. We discuss extensively practical design choices and considerations. We further introduce how to solve the optimal decision rule via Reinforcement Learning and scale the solution. We show the effectiveness of this novel approach compared with existing methods via a large-scale meta-analysis on experiments in Amazon.

CCS CONCEPTS

• **General and reference** → **Experimentation**; • **Computing methodologies** → **Sequential decision making**; **Reinforcement learning**.

KEYWORDS

Sequential Decision Making, A/B testing, Reinforcement learning

ACM Reference Format:
Runzhe Wan, Yu Liu, James McQueen, Doug Hains, and Rui Song. 2023. Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential Decision-Making for Continuous Monitoring. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599818

*Equal contribution

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599818

1 INTRODUCTION

Online A/B testing has become a common practice in online service companies (e.g., Amazon, Google, Netflix, etc.) to evaluate the customer impact of new features (treatments) in comparison to the old ones (control) [16]. Unlike traditional randomized-controlled trials (RCT), in online A/B tests the experimental population is not known a priori and *random assignment* is often employed. In online experiments, customers interacts with the feature being experimented on (e.g. makes a search query) are randomly assigned to either the treatment or control and are part of the experiment. As a result, the sample size of an online experiment is a function of its duration and the longer an experiment is run the more samples are collected and the higher power we get.

A/B experiments are typically conducted using *fixed-horizon hypothesis testing*, where the duration of an experiment is predetermined in order to collect enough samples to achieve some desired statistical power (e.g., 80%) [25, 36]. However, running experiments has a non-zero cost. Longer experiments can slow down the innovation cycle and consume (potentially costly) hardware and human resources. The impact of the experiment on the customers is also a real concern. For example, a treatment with negative impact should be terminated early to reduce customer exposure to a negative experience. Similarly, positive treatments should be launched as soon as possible to maximize the customer benefit. In contrast to the fixed-horizon approach, methods like *continuous monitoring* (also known as *early termination*) allow us to update guidance on experiment duration while the experiment is still running.

Despite the extensive research on continuous monitoring (see Section 2 for a review), to the best of our knowledge, almost all of them are from the *hypothesis testing* perspective, which focuses on the decision accuracy. The primary objectives are typically the false positive rate (FDR), type-I error or type-II error. This is probably due to the reason that most of these methods are from high-stake applications such as clinical trials, and the hypothesis testing viewpoint is known as conservative. In contrast, in online e-commerce companies, tens of thousands of experiments are run per year and most of these experiments are not high-risk; meanwhile, the majority of them cannot achieve statistical significance with desired

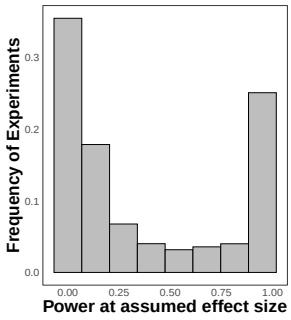

Figure 1: Histogram of the frequency of the power values (at a given assumed effect size) at the end of the experiments.

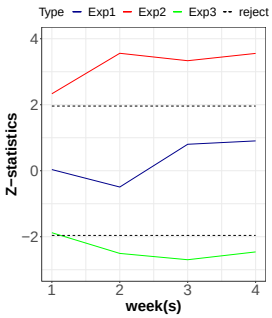

Figure 2: Trajectories of z-stats for three actual experiments. The dotted lines indicate the rejection boundary.

powers, since the noise level of responses is high (given that the whole system is very complicated with numerous moving parts). In addition, online experiment platform are designed to encourage exploring novel ideas, but on the flip side many of them will not yield positive or significant results. The opportunity cost of conducting the experiment is also high due to the sibling teams are waiting on the same resources. For example, when teams experiment on the same customer group, they may need to wait their turn to access the same resources. Therefore, finding an appropriate balance between customer benefits and opportunity costs is crucial for continuous monitoring A/B testing in online experimentation platforms.

Classic methods that focus on the type-I error and power turn out to be too conservative, as we show empirically using real experiments conducted in Amazon. Figure 1 displays the *ex-post* measured powers of a large number of experiments. The histogram shows that there is a large number of experiments with very low powers, which we can terminate earlier since they are not able to achieve the significance level (with high probability). There is also a large number of experiments with very high powers, for which we can terminate earlier as well, since the evidence becomes strong in a short time window. In Figure 2, we further zoom in by looking at the trajectories of z-statistics of a few representative experiments. Z-statistics are calculated using a fixed-horizon two-sided z-test with a significance level of 0.05 [50]. Experiment 2 shows consistently positive z-statistics (i.e. treatment is consistently better than control), experiment 3 shows consistently negative z-statistics, and the z-statistics of experiment 1 are close to zero. It is intuitive to consider terminating these experiments earlier. In conclusion, applying earlier termination can be beneficial a lot in online A/B testing, in terms of efficiency and resource allocation.

Motivated by these real needs, we are concerned with the following question:

How can we efficiently and appropriately early terminate an experiment, in a way that comprehensively considers and balances different sources of costs and benefits to maximize the customer utility?

Contribution. Motivated by the real needs we observed for online experiments, we propose a continuous monitoring service that can provide early termination recommendations when appropriate. We propose to formulate the problem as a Bayesian sequential decision making (SDM) problem [6]. We design a concrete yet general framework that appropriately formulates the online A/B testing into SDM. The objective function is the expected cumulative utility from an online A/B experiment, where the decision accuracy, the opportunity costs and the experimentation impacts are all taken into consideration. We discuss a few extensions to accommodate different real needs. We also discuss how to solve the problem via RL and how to scale our solution by leveraging the contextual RL technique. Finally, we present a large-scale meta-analysis on a large number of past experiments at Amazon. The results clearly demonstrate the advantage of RL, as an effective and practical solution for early stopping online experiments. To the best of our knowledge, this paper is the *first* systematic study of RL-based approach for continuous monitoring. The paper aims to provide a detailed guidance on how to apply these techniques in industrial settings.

Outline. We organize the paper as follows. In Section 2, we introduce the related works. Section 3 gives the problem setting and assumptions which are consistent with Amazon's practice. In Section 4, we provide some preliminaries on SDM. Section 5 provides the details of our approach, which is supported by our meta-analysis results in Section 6. Section 7 concludes this paper.

2 RELATED WORK

Early termination. In the statistics literature, there is a long history on studying how to detect the true treatment effect as quickly as possible or give up sooner if there is little hope. The alpha-spending function approach[10, 15] is proposed to control the overall type-I error across the interim tests. In Bayesian hypothesis testing, the sequential Bayes factor (SBF) procedure is widely used [21, 38], which is similar to the sequential testing proposed in [48]. [11] shows the validity of SBF and proposed a stopping rule based on it, which controls the FDR. Always valid inference [22, 23] is proposed as an user-friendly framework to control the type-I error, by converting any existing sequential testing procedure into a sequence of p-values. A few methods (such as the conditional power or the predictive power approach) that predict the outcome of the experiment so as to early terminate are also designed [26]. However, all of them focus on conventional statistical accuracy metrics, and hence the objectives are different from ours.

RL for experiment design. RL has been applied to many real domains including healthcare [34], robotics [3], epidemiology [49] and autonomous driving [37]. In recent years, RL has been also successfully applied to sequential experiment design. [33] proposes a few simulation-based methods for maximizing the expected utility, which however is not computationally feasible with continuous spaces. RL techniques such as approximate DP [19] and policy gradient [41] are later applied to obtain an approximate solution. This problem is recently extended to even more challenging cases, where an explicit model is not available but sampling from the distribution is feasible [13, 20]. [44] compared RL with simulation-based methods and discussed its potential applications in clinic trials. Overall, few attention is focused on the early stopping problem and it is only sometimes used as a toy example. To the best of our knowledge, this paper is the *first* systematic study of RL-based approach for early stopping. We advance both methodology development and industrial applications on this topic.

3 PROBLEM SETUP

In a typical online A/B test, customers may visit the site multiple times, and when customers visit can't be controlled. Every visit generates an outcome of interest. However, typically an online site is very complicated and a new feature is incremental. Therefore, it is common practice to only focus on those outcomes after a customer is *triggered* based on some experiment-specific trigger mechanism definition (e.g., after she sees the related features). This is when we call the customer has entered the experiment [36]. Besides, unlike traditional testing settings where there are a small number of i.i.d. units, online experiments may have millions of units and are non-stationarity. It is not practice to make a decision with every new data point. Therefore, we follow this tradition to aggregate users' responses on a weekly basis. Other choices of the time unit are also possible, depending on the applications.

Notations. Let T be the horizon length, which represents the maximum allowed duration (number of weeks) for an experiment. Denote $N_{k,g}$ as the number of customers who are triggered at week k in group $g \in \{Tr, C\}$. In this paper, Tr denotes the treatment group and C denotes the control group. Denote the number of customers and the set of those customers who are triggered between week t and l as $N_{t:l,g}$ and $\mathcal{I}_{t:l,g}$, respectively, where $N_{t:l,g} = \sum_{k=t}^{l} N_{k,g}$. For a customer i, denote $W_{i,k}$ as the (aggregated) outcome in week k, and $W_{i,t:l} = \sum_{k=t}^{l} W_{i,k}$ as the cumulative outcome between week t and l. A customer begins to generate outcome after being triggered.

Assumptions on stationary treatment effects. For ease of exposition of our main idea, we assume the treatment effects are stationary over time. Formally, we assume $E[W_{i,t}] = \mu_g, \forall i \in \mathcal{I}_{1:k,g}$ and $\forall t \in \{1, \ldots, T\}$, where μ_g is the population mean of weekly aggregated response for group $g \in \{Tr, C\}$. We will discuss how to relax this assumption. We further define the per-customer per-week treatment effect as $\delta = \mu_{Tr} - \mu_C$.

Earlier Termination. With a fixed-horizon procedure, we always wait until time T and use all data to make the decision. Since we observe data sequentially over time, we can consider terminating earlier when appropriate. Specifically, we can make a decision at each time point t, such as whether to continue the experiment, launch the new feature or not, based on all available data.

4 PRELIMINARIES

We first introduce a general formulation of the Sequential Decision Making (SDM) problem. We focus on the finite-horizon undiscounted non-stationary setup, which is most relevant to our problem. We start with the (potentially) non-Markovian setup in Section 4.1, which is most general and hence flexible. In Section 4.2, we discuss the Markovian variant, and introduce the belief states for the Bayesian perspective.

4.1 Sequential Decision Making

Let the horizon be T. A (potentially) *non-Markov decision process (NMDP)* is defined by a series of observation space $\mathcal{Y}_t$, action space $\mathcal{A}_t$, transition kernel $\mathcal{P}_t$, and reward kernel $\mathcal{R}_t$ for every $t = 1, \ldots, T$. Let $\{(Y_t, A_t, R_t)\}_{t \geq 1}$ denote a trajectory generated from the NMDP model, where (Y_t, A_t, R_t) denotes the observation-action-reward triplet at time t. Denote $\mathcal{H}_t = (Y_1, A_1, R_1, \ldots, Y_{t-1}, A_{t-1}, R_{t-1}, Y_t)$ as the historical information available until time t. We denote all unknown parameters of the system by $\boldsymbol{\theta}$. In the Bayesian setting, we assume a prior over $\boldsymbol{\theta}$ as $\rho(\boldsymbol{\theta})$. To simplify the presentation, we assume the probability spaces are continuous. The transition kernel $\mathcal{P}_t(Y_t|\mathcal{H}_{t-1}, A_{t-1}; \boldsymbol{\theta})$ gives the probability density of observing Y_t by taking action A_{t-1} given the history H_{t-1}, and similarly $\mathcal{R}_t$ generates the reward. A (deterministic) history-dependent *policy (a.k.a. decision rule)* $\pi = (\pi_1, \ldots, \pi_T)$ is a series of functions, where π_t map the history to an available action $A_t = \pi_t(\mathcal{H}_t) \in \mathcal{A}_t$.

A trajectory following policy π is generated as follows. The agent starts from an observation Y_1. At each time point $t \geq 1$, the agent selects an action $A_t = \pi_t(\mathcal{H}_t)$, then receives a random reward $R_t \sim \mathcal{R}_t(\cdot \mid \mathcal{H}_t, A_t; \boldsymbol{\theta})$, and finally observes the next observation $Y_{t+1} \sim \mathcal{P}_t(\cdot \mid \mathcal{H}_t, A_t; \boldsymbol{\theta})$. For a policy π, its history value function (V-function) and history-action value function (Q-function) [43] are defined as

$$V_{t'}^{\pi}(h) = \mathbb{E}^{\pi}\Big(\sum_{t=t'}^{T} R_t | \mathcal{H}_t = h\Big), Q_{t'}^{\pi}(a, h) = \mathbb{E}^{\pi}\Big(\sum_{t=t'}^{T} R_t | A_t = a, \mathcal{H}_t = h\Big),$$

where the expectation $\mathbb{E}^{\pi}$ is defined by assuming the system follows the policy π. More specifically, the expectation is taken over trajectories with $A_t = \pi_t(\mathcal{H}_t), \forall t$, and the other dynamics following the underlying MDP model.

The objective is to find an optimal policies $\pi^* = (\pi_1^*, \ldots, \pi_T^*)$ following which we can maximize the expected cumulative reward (i.e., the utility). It can be defined as $\pi^* \in \arg\max_{\pi} V_{t'}^{\pi}(h), \forall t', \forall h$. The optimal Q-function is defined as:

$$Q_T^*(a, h) = \mathbb{E}\Big(R_T | A_T = a, \mathcal{H}_T = h\Big)$$
$$Q_t^*(a, h) = \mathbb{E}\big[R_t + \max_{a' \in \mathcal{A}_{t+1}} Q_{t+1}^*(a', \mathcal{H}_{t+1}) | A_t = a, \mathcal{H}_t = h\big], \forall t < T$$

we have $\pi_t^*(h) = \arg\max_{a \in \mathcal{A}_t} Q_t^*(a, h)$, for any h and t.

4.2 Markov Decision Process

When additional structures are satisfied, we can consider a more efficient decision process model, the Markov Decision Process [MDP, 35]. The (Bayesian) MDP is particularly relevant when disucssing the model that we are working with (see Section 6.1).

Specifically, suppose we can construct some observable so-called state variable S_t at every time point t. The key assumption of MDPs is the Markov assumption (with slight overload of notations)

$$\mathbb{P}(S_{t+1} | \{Y_j, A_j, R_j\}_{1 \leq j \leq t}; \boldsymbol{\theta}) = \mathcal{P}_t(S_{t+1} \mid A_t, S_t; \boldsymbol{\theta}),$$
$$\mathbb{P}(R_t | S_t, A_t, \{Y_j, A_j, R_j\}_{1 \leq j < t}; \boldsymbol{\theta}) = \mathcal{R}_t(R_t \mid A_t, S_t; \boldsymbol{\theta}).$$

The assumption requires that there is no information useful for predicting the future being omitted from the state vector. In other words, roughly speaking, the state vector is a sufficient statistic of the history. Under an MDP, the optimal policy can only depend on the current state S_t instead of the whole history $\mathcal{H}_t$, so do the Q- and V-function for this class of policies.

Bayesian MDP and belief states. In the Bayesian setup, the unknown parameter $\boldsymbol{\theta}$ is a random variable, which becomes a confounding variable and typically invalidates the Markov property. The model becomes a so-called partially observable MDP (POMDP). Fortunately, one can adopt a classic technique to transform such

a Bayesian problem as an MDP, by constructing the so-called *belief state*. Specifically, we can include the posterior of $\boldsymbol{\theta}$, $P(\boldsymbol{\theta}|\mathcal{H}_t)$, as a component of the state S_t. Then it is easy to check that the problem becomes an MDP again [47]. When the posterior belongs to a parametric class, it is sufficient to include the parameters of the posterior (i.e., the sufficient statistics that capture all historical and prior information). One can also show that, the transition and reward function can be equivalently written as being based on the marginalized distribution over the posterior $P(\boldsymbol{\theta}|\mathcal{H}_t)$.

5 EARLY TERMINATION WITH RL

We present our proposed framework in this section. We first discuss the choice of the overall RL approach in Section 5.1. We then lay down the model formulation in Section 5.2, and next discuss this formulation and its extensions in Section 5.3, We intentionally make the formulation general, so that it can be adapted to different real needs. Finally, we introduce how to solve an optimal policy via RL and how to scale the solution in Section 5.4.

5.1 Policy Optimization with Model-based Bayesian RL

Reinforcement Learning [RL, 43] solves the optimal policy π by aggregating information from data (trajectories). There are multiple ways to classify RL methods.

First of all, RL methods can be dichotomized as *model-based* v.s. *model-free*. The model-based approach requires knowledge of (or assumption on) the transition and reward kernels, while the model-free approach does not. The former approach is typically more efficient, at the cost of needing knowledge (assumptions) on the model structure. In our problem, there are a lot of problem-specific structures we can utilize (e.g., the experiment terminates when we take the corresponding actions; see Section 5.2 for more details). Moreover, we observe that different experiments differ significantly. A model-free approach needs to learn one policy from them all (without model structure) and hence would be sub-optimal. Therefore, it is natural to adopt the model-based approach.

Second, RL approaches can also be divided as *online* and *offline*: the former continuously interacts with the real environment, while the latter does not. In our applications, it is not feasible to trial and error with every new experiment. Therefore, we utilize historical information to build a model for the new experiment, use the model as a simulator and solve it via online RL algorithms. The overall framework is offline.

Finally, we adopt a Bayesian approach, which allows us to utilize valuable prior information and also formulate an optimization problem in a natural way. Such a choice is classic and closely related to the vast literature on Bayesian decision theory [14] and Bayesian experiment design [9].

5.2 SDM Problem Formulation

In the following subsections, we define the components of our model. We exposit under the most general non-Markovian scenario. One example about our application of this framework at Amazon will be given in Section 6.1, where we consider a special Markovian case. Figure 3 provides an illustration. We consider a *finite-horizon* setup, where at every decision point $t = 1, \ldots, T$, we will make a decision. Note that, when the experiment has already been terminated, any action below will not make any difference. Alternatively, one can also consider an *indefinite-horizon* setup, based on which we observed similar performance. Recall that the t-th decision point is the *end* of the t-th time period.

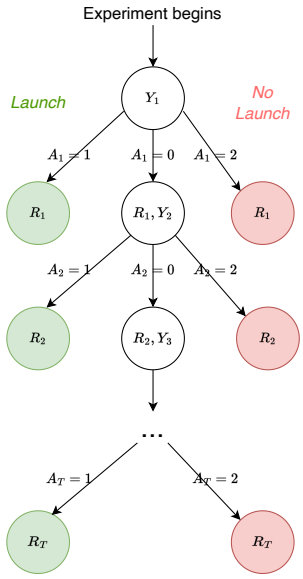

Figure 3: Sequential Decision Making Process for early terminating experiments. Experiment starts at week 0 and must continue for at least one week. At each decision point t (except for the last one), we can choose from "continue", "stop and launch", and "stop and no launch". A reward R_t will then be realized.

5.2.1 Unknown Parameters. The unknown parameter vector $\boldsymbol{\theta}$ contains at least the mean response of customers in the two groups, i.e., μ_{Tr} and μ_C. $\boldsymbol{\theta}$ can also contain other unknown quantities, such as the sample size distribution and the variance/covariance terms, etc., depending on design choice. In these cases, we only need to include the related data in the observation vector and modify the components to be introduced accordingly.

In practice, there is a rich literature on sample size prediction [2, 7], and with the model proposed in [36] we observed a high accuracy and low uncertainty. Therefore, to avoid unnecessary complexity and for ease of exposition, we assume the sample sizes in every week are known *a priori*. Similarly, we assume the variances of outcomes are known, which in practice are typically plugged in with the sample variances. The framework can be naturally extended by treating the uncertainty and estimation of these components as part of the transition dynamic. We assume we have a prior distribution $\rho(\boldsymbol{\theta})$. We use $\mathbb{E}_\rho$ to denote $\mathbb{E}_{\boldsymbol{\theta}\sim\rho(\boldsymbol{\theta})}$, i.e., taking expectation over this prior.

5.2.2 Observations. In week t, if the experiment remains continuing, the observation Y_t contains the outcomes of all customers triggered before the end of that week, i.e., $\{W_{i,t}\}_{i\in\mathcal{I}_{1:t,Tr}}$ and $\{W_{i,t}\}_{i\in\mathcal{I}_{1:t,C}}$. Y_t may also include other observed variables useful for estimating $\boldsymbol{\theta}$. If the experiment has been terminated, we denote $Y_t = \emptyset$. Because

in our case the observations will become part of the history, we discuss the observation generation model in Section 5.2.5, where we focus on the transition kernel.

5.2.3 Actions. From a high level, there are three categories of decisions we can make at every decision point, including (i) when to stop the experiment, (ii) if stopped, whether to launch the new feature, and (iii) what statistical statements to make (e.g., the p-value of a given metric). There may be many variants of the action space. In this paper, we will focus on the following three options that we are working with:

- a = 0: keep running;
- a = 1: stop and launch;
- a = 2: stop and no launch.

It is natural to set the action spaces as $\mathcal{A}_t = \{0, 1, 2\}, \forall t < T$, and $\mathcal{A}_T = \{1, 2\}$. Let I_t be the indicator for the event $\bigcup_{t'<t}\{A_{t'} \neq 0\}$, i.e., the experiment has been terminated before time t. We note that, when $I_t = 1$, any choice of actions does not make difference, as the experiment has terminated.

5.2.4 Objective and Reward. In this section, we give our optimization objective. Specifically, for any trajectory $\{(A_t, Y_t)\}_{t\leq T}$ and any value of the unknown parameter $\boldsymbol{\theta}$, we define an utility function and aim to solve a policy that maximizes the expected utility. The utility should include (at least) the following components:

(1) c: we assume there is a fixed and pre-specified *weekly cost for running the experiment*. This may include (with some potential overlap): The *personel and harware cost* to support this specific experiment, the *opportunity cost for the experimentation platform*, and the *opportunity cost for the experimenters*.
(2) c_h: The *hurdle cost*, which is the the cost needed to launch (e.g., the implementing in production) new feature.
(3) The *customers impact from making the launch recommendation* on weeks t. This term concerns the decision accuracy, i.e., we would like to launch those new features that indeed have positive customer impacts.
(4) The *customers impact on the treatment group from the experiment itself*. For example, if the treatment has clear negative impacts on customers, it should be terminated earlier.

We assume all these components share the same unit D, which can be monetary or something else that measures customer impacts.

Utility of an experiment. Notice that, since ATE is a relative value, our utility/reward definitions below are also relative to the outcomes under the control. The utility can then be defined as

$$\begin{aligned} &u(\{(A_t, Y_t)\}_{t\leq T}, \boldsymbol{\theta}) \\ &= \underbrace{-c \times \sum_{t=1}^{T} \mathbb{I}(I_t = 0, A_t = 0)}_{\text{Opportunity Cost}} + \underbrace{\sum_{t=1}^{T} \big(\delta \times N_{1:t+1,Tr}\big)\mathbb{I}(I_t = 0, A_t = 0)}_{\text{Impact during the experiment (relative to control)}} \\ &+ \underbrace{\sum_{t=1}^{T} (u_2(\delta, H, T, t) - c_h)\mathbb{I}(I_t = 0, A_t = 1)}_{\text{Launch impact (relative to control)}}, \end{aligned} \tag{1}$$

where $N_{1:l,g}$ denotes the cumulative number of triggered customers in group $g \in \{\text{Tr}, \text{C}\}$ and recall δ is the per-customer per-week average treatment effect defined in Section 3. The first term represents the opportunity cost incurred for conducting the experiment until termination. The second term represents the impact on the treatment group during the experiment. If the treatment has a clearly negative impact, the second term will encourage we to terminate the experiment earlier. In the third term, $u_2(\delta, H, T, t)$ represents the impact from launching the treatment feature on the whole population at week t, given a pre-specified time horizon H. Typically, we assume a one-year time horizon ($H = 52$ weeks) and linear extrapolation, with all customers triggered in the experiment period regarded as the target population. This gives an example of the launch impact as $u_2(\delta, H, T, t) = \delta \times (H + T - t) \times (N_{1:T,Tr} + N_{1:T,C})$. We note that, the third term has an decreasing relationship with the week we recommend launch. Thus, the third term encourages launching a good feature earlier.

We emphasize that, Equation (1) does not include the impact and cost in week 1, since they are constants that our policy does not have control on (any experiment needs to run for at least one week). Besides, the first two terms correspond to the impact on week $t + 1$ from the decision made at the end of week t (i.e., at the t-th decision point). These two terms are always 0 at $t = T$. This is because the experiment will end no later than T, so $A_T \neq 0$ (recall our definition of the action spaces).

Objective. Our objective is to solve a policy π^* that maximizes

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta}\sim\rho(\boldsymbol{\theta})} u\Big(\{(A_t, Y_t)\}_{t\leq T}, \boldsymbol{\theta} \mid Y_1 = y\Big) \tag{2}$$

for any y. Here, the expectation is taken over the trajectories following π. In other words, at every time point, given the prior, the observed data, and some other known components needed for modeling the transition (such as the sample size prediction model), we aim to determine the optimal policy (decision rule) that maximizes the expected utility, in the Bayesian sense.

Rewards. To apply RL algorithms, we decompose the objective into the per-round instant reward R_t in a way that the objective is equal to $\mathbb{E}^{\pi}_{\rho} \sum_{t=1}^{T} R_t$. There exist multiple equivalent ways. We consider a natural way as

$$R_t = \begin{cases} -c + \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_t\big] * N_{1:t+1,Tr}, & \text{when } A_t = 0 \text{ and } I_t = 0, \\ \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_t\big] * N_{1:T}(H + T - t) - c_h, & \text{when } A_t = 1 \text{ and } I_t = 0, \\ 0, & \text{when } A_t = 2 \text{ and } I_t = 0, \\ 0, & \text{when } I_t = 1, \end{cases} \tag{3}$$

where $N_{1:T} = \sum_{g\in\{Tr,C\}} N_{1:T,g}$. When defining the impact from the treatment (in the first two cases), we replaced δ in (1) by $\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_t\big]$. This is a common trick which gives an equivalent objective and reduces the randomness to make training easier.

5.2.5 Transition Kernel. The last piece of the formulation is the transition model $p(Y_{t+1} \mid A_t, \mathcal{H}_t)$, i.e., the distribution of our next-period observations given the history, our current action, and the prior. Based on the Bayes rule, we can first sample $\tilde{\boldsymbol{\theta}}$ from $p(\boldsymbol{\theta} \mid A_t, \mathcal{H}_t) = p(\boldsymbol{\theta} \mid \mathcal{H}_t)$ and then sample from $p(Y_{t+1} \mid A_t, \mathcal{H}_t, \tilde{\boldsymbol{\theta}})$. We will focus on discussing $p(Y_{t+1} \mid A_t, \mathcal{H}_t, \boldsymbol{\theta})$, because $p(\boldsymbol{\theta} \mid \mathcal{H}_t)$ is essentially an Bayesian estimation problem and should have been studied extensively in an experimentation platform.

By definition, we have $p(Y_{t+1} = \emptyset \mid A_t = a, \mathcal{H}_t = h, \boldsymbol{\theta}) = 1$ whenever $I_t = 1$ or $a \in \{1, 2\}$. In other words, the experiment has been terminated and there are no further observations. In all the other cases, this requires us to model the distribution of customer-level outcomes in the next time period, conditioned on all customer-level observations so far and the value of $\boldsymbol{\theta}$.

If the observations in different weeks are independent, then sampling based on $p(Y_{t+1} \mid A_t = 0, \mathcal{H}_t, \boldsymbol{\theta})$ would be straightforward: since Y_{t+1} are independent from $\mathcal{H}_t$ conditioned on $\boldsymbol{\theta}$ (i.e., $p(Y_{t+1} \mid A_t = 0, \mathcal{H}_t, \boldsymbol{\theta}) = p(Y_{t+1} \mid A_t = 0, \boldsymbol{\theta})$), we only need to sample i.i.d. data points from the corresponding environment model with known parameters $\boldsymbol{\theta}$. Furthermore, by the central limit theorem, we can typically avoid the need of specifying a parametric model for each data point and can focus on the average outcome (see Section 6.1 for more details). In some other applications where the observations could be dependent over different weeks (e.g., the outcomes from the same customer can be correlated), typically we apply a repeated measurement model (see, e.g., [29]).

The choice of the outcome model depends on the application, so we will not stick to a specific one here. As a case study, in Section 6.1, we present the model that we are working with at Amazon.

5.3 Discussion and Extensions of the Formulation

In this section, we list a few questions frequently asked in practice, to help elucidate the formulation and its natural extensions. We defer some more complicated extensions to the conclusion section.

Bandits. One common question is why do we formulate the problem as an NMDP (a multi-step model) instead of a bandit problem (a single-step model), which is more commonly considered in experimentation. This is because that our actions include "terminating an experiment", which will naturally affects time points after it and hence introduces long-term dependency.

Guardrails. Experiments may have certain safety guardrails on some metrics - once these guardrails are violated, the experiment will be terminated automatically. Such a mechanism can be naturally incorporated into our framework by modifying the transition function: once the constraints are violated, we set $I_t = 0$ automatically no matter what action the policy takes.

Statistical inference. Our formulation focuses on utility maximization, which is different from pure statistical inference and fits our applications better. However, in some cases, experimenters may still want some valid statistical statements along with the recommendation from the experimentation platform. We provide two approaches to add this important feature. First, we can modify the action space so as RL only determines whether or not we should terminate, and we slightly modify the transition and reward functions to *automatically* make the launch decision based on the always valid p-values (AVP) [24] when terminated. As such, our procedure satisfies the requirements in [24] and hence the launch decision has a valid type-I error control. Such a decision rule has the highest utility among all rules that satisfy the type-I error bound. Alternatively, we can formulate a multi-objective problem, where the objective is the sum of the original utility and an indicator of making the wrong decision (multiplied by a parameter λ). This is essentially a Lagrange transformation. We can tune λ until the error rate falls below a desired threshold, or share a list of policies under different values of λ to provide users with choices.

Multi-armed experiments. This extension is straightforward, by duplicating (and slightly modifying) the treatment-related components in the state and action definitions.

ATE estimators. The framework introduced in Section 5.2 is general and not restricted to a specific ATE estimator. For example, one can apply covariate adjustment [28] or model the heterogeneity [36] among customers via hierarchical modeling. One only needs to keep the related variables in the observation vector and modify the transition kernel accordingly.

Role of information gain. The weekly cost c clearly reflects the penalty for running long experiments. One natural question is how does the formulation encourage gaining information, which is central to an experiment. This may not be explicit, since our objective, as the *expected* utility, seems not related to the *uncertainty*.

The answer is, this component has been naturally but implicitly considered in our objective. To see this clearly, we present an illustrative example below to show how does our objective encourages gaining information. Note that a fixed-horizon Bayesian optimal decision rule is one that waits until a pre-specified time point t (i.e., $\mathcal{A}_l = \{0\}, \forall l < t$), and then launch when the posterior mean of the launch impact is higher than the hurdle cost (i.e., $\pi_{1,t}(\mathcal{H}_t) = 2 - \mathbb{I}\big[\mathbb{E}_\rho(\delta \mid \mathcal{H}_t) \times H \times N > c_h\big]$).

Lemma 1. *Assume two policies π_1 and π_2: π_1 waits until $t = t'$ to make a decision $A'_t \in \{1, 2\}$ following the fixed-horizon Bayesian optimal decision rule; while π_2 waits until $t = t' + 1$ to do so. To simplify the notation, we assume $c_h = 0$. For any history h at time t', the difference between the expected utilities of the two policies is*

$$
\begin{aligned}
&\mathbb{E}_\rho^{\pi_2}\Big[\sum_{t=t'}^{T} R_t \mid \mathcal{H}_{t'} = h\Big] - \mathbb{E}_\rho^{\pi_1}\Big[\sum_{t=t'}^{T} R_t \mid \mathcal{H}_{t'} = h\Big] \\
&= \Big\{\mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\
&\quad - \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times \mathbb{I}[\pi_{2,t}(h) = 1]\Big\} \times N_{1:T} \times (H + T - t') \\
&- \mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\
&+ \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times N_{1:t'+1,Tr} - c.
\end{aligned}
$$

Proof can be found in Appredix A.2. The first term represents the value of information. Specifically, we have

$$
\begin{aligned}
&\mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\
&\quad - \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times \mathbb{I}[\pi_{2,t}(h) = 1] \\
&= \mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}\big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] > 0\big] \mid \mathcal{H}_{t'} = h\Big] \\
&\quad - \mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times \mathbb{I}\big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] > 0\big] \\
&\geq 0,
\end{aligned}
$$

where the equality follows from the policy definitions. The inequality holds as long as $\mathbb{E}_\rho\Big\{\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \mid S_{t'}\Big\} = \mathbb{E}_\rho\big[\delta \mid S_{t'}\big]$, which is true since $\mathcal{H}_{t'+1}$ contains more information than $\mathcal{H}_{t'}$ does.

5.4 Solving the Optimal Policy

With the formulation above, the next step is to solve (2) to obtain the optimal policy, so that we can use it for optimal termination. In the RL literature, there exists many efficient algorithms that be can be applied. For completeness, in Section 5.4.1, we recap a classic one that we are working with. In Section 5.4.2, we further illustrate how to scale the solution so as to support real applications on major experimentation platforms.

5.4.1 Policy Learning with RL. With an environment model, one can apply various state-of-the-art RL algorithms to solve the problem, such as the value-based algorithms(e.g., DQN [32], Double DQN [46]), policy-based algorithms (e.g., TRPO [39], PPO [40]) which directly learn a policy function, or actor-critic algorithms (e.g., A2C [31], SAC [17], A3C [31]) which reduce the variance of policy-based methods by additionally learning the value function. Typically, the history can be summarized into a finite-dimensional vector. When not feasible, one can consider using models such as the long short-term memory network [4].

With a discrete action set in our problem, we are currently using a classic value-based approach, deep Q-network [DQN, 32]. For completeness, we briefly review its main idea here. DQN uses a deep neural network to parameterize the value function Q, and it is a Q-learning-type algorithm based on the fixed-point iteration principle. It is motivated by the fact that $Q^{\pi^*} = \{Q_t^{\pi^*}\}_{t=1}^T$ is the unique solution of the *Bellman optimality equation*

$$Q_t(a,h) = \mathbb{E}\Big(R_t + \gamma \arg\max_{a \in \mathcal{A}_{t+1}} Q_{t+1}(a, \mathcal{H}_{t+1}) | A_t = a, \mathcal{H}_t = h\Big), \forall t.$$

Regard the right-hand side as an operator on $Q = \{Q_t\}_{t=1}^T$. We can prove it is a contraction mapping and Q^* is its fixed point, based on which we can derive an iterative algorithm. Details of the MDP-version of this algorithm can be found in, e.g., Algorithm 1 in [43].

Remark 1 (Choice of the policy/value function class). *In our applications, we use neural networks as our value function class, due to their great representation power to approximate the true optimal policy (which is typically highly non-parametric and does not belong to a pre-defined function class). In use cases where the interpretability is desired, one can instead use, e.g., linear models.*

5.4.2 Contextual RL: Scalability and Deployment Feasibility. So far, we have described how to solve the optimal policy for one specific experiment. However, nowadays, an online experimentation platform in large companies may need to support tens of thousands of experiments per year. For different experiments, the NMDP models might be different (e.g., with different priors, different sample sizes, etc.) and hence the optimal policies differ. Standard RL algorithms are designed for a fixed MDP environment, and its limited generalizability is well known as a huge bottleneck [42].

Therefore, one practical issue is the scalability and tractability of maintaining such an RL service. If we need to re-run the RL algorithm for every experiment, then such a service in not practical, as it requires significant manual effort. Even if we can trigger the training automatically, the computational resource needed and the training instability are of concerns.

To address the scalability issue, we form the task as a contextual RL problem [5, 18], a direction that is attracting increasing attention. Suppose each NMDP $\mathcal{M}_i = \{\mathcal{Y}_t^i, \mathcal{A}_t^i, \mathcal{P}_t^i, \mathcal{R}_t^i\}_{t=1}^T$ is associated with a feature vector $\mathbf{x}_i$ (i.e., the so-called *context*) and the difference between all NMDPs can be fully captured by the context, i.e., we can rewrite as $\mathcal{M}_i = \{\mathcal{Y}_t^{\mathbf{x}_i}, \mathcal{A}_t^{\mathbf{x}_i}, \mathcal{P}_t^{\mathbf{x}_i}, \mathcal{R}_t^{\mathbf{x}_i}\}_{t=1}^T$. Then, as long as we include the context as part of the history to define an *augmented* history, we can construct one single new NMDP that unifies all these NMDPs (and generalize to those still unseen). For example, we can define the transition function $\mathcal{P}_t$ as $\mathcal{P}_t(h_t, \mathbf{x}_i) = \mathcal{P}_t^{\mathbf{x}_i}(h_t)$. Similar arguments apply to the other components.

After this transformation, all the policy learning algorithms discussed in Section 5.4.1 can be applied in the same way, except that we are going to sample trajectories by interacting with a set of MDPs in learn a generalizable policy. We can run training for only once to learn the policy, save it, and then apply it to any new experiments. Intuitively, this contextual policy learns how to act given both the history of an experiment and its features (the context). In Section 6.1, we present a concrete example of the context vector.

6 CASE STUDY ON HISTORICAL EXPERIMENTS IN AMAZON

So far, we lay down a concrete but general framework for utility-maximizing early termination. In this section, we present an example of how we apply it to one of Amazon's largest experimentation platforms. We first introduce a specific transition model in Section 6.1, and then present the meta-analysis results in Section 6.2.

6.1 SDM Model with Independent Outcomes

The SDM formulation and the RL algorithms in Section are general and not restricted to a specific outcome model. In this section, we provide one concrete form when the outcomes are independent across weeks. This model is what we are working with and the numerical results are also based on this model. Moreover, we can design appropriate state vector to make our NMDP model an MDP.

As discussed in Section 5.2.5, we will focus on discussing $p(Y_{t+1} \mid A_t = 0, \mathcal{H}_t, \boldsymbol{\theta})$. We assume the observations in different weeks are independent. More formally, we assume that, conditioned on $\boldsymbol{\theta}$ and $\{A_t\}_{t=1}^T$, $\{Y_t\}_{t=1}^T$ are mutually independent. For simplicity of notations, we also assume the outcomes follow Gaussian distributions. However, we emphasize that, by the central limit theorem [8], all derivations below still approximately hold with non-Gaussian outcomes, when the number of customers is large (typically the case). We assume a conjugate prior for μ_g, the mean of group g. Therefore, our full model is

$$\begin{aligned} \mu_g &\sim N(\mu_{0g}, \sigma_{0g}^2), \\ W_{i,t}|\mu_g &\sim N(\mu_g, \sigma_g^2), \text{ for customer } i \text{ in group } g \in \{\text{Tr, C}\}. \end{aligned} \quad (4)$$

Recall that $W_{i,t}$ are responses observed in week t for customer i. In practice, the prior parameters μ_{0g} and σ_{0g} are typically estimated via empirical Bayes [30] from historical experiments. Based on the derivations in Appendix A.1, following the Bayesian rule, for each

group $g \in \{\text{Tr}, \text{C}\}$, we have

$$\mu_g|\overline{W}_{l,g} \sim N\left(\left(\frac{1}{\sigma_{0g}^2}+\frac{a_g(l)^2}{\sigma_g^2 b_g(l)}\right)^{-1}\left(\frac{\mu_{0g}}{\sigma_{0g}^2}+\frac{a_g(l)\overline{W}_{l,g}}{\sigma_g^2 b_g(l)}\right), \left(\frac{1}{\sigma_{0g}^2}+\frac{a_g(l)^2}{\sigma_g^2 b_g(l)}\right)^{-1}\right),$$

$$\overline{W}_{l+1,g}|\mu_g, \overline{W}_{l,g} \sim N\left(\frac{\overline{W}_{l,g}N_{1:l,g}+\mu_g N_{1:l+1,g}}{N_{1:l+1,g}}, \frac{1}{N_{1:l+1,g}}\sigma_g^2\right).$$

Here, $\overline{W}_{l,g}$ is the average outcomes of $W_{i,1:l}$ among customers who participated in the experiment up to week l, $a_g(l) = c_g(l)/N_{1:l,g}$, $b_g(l) = c_g(l)/N_{1:l,g}^2$, and $c_g(l) = \sum_{t=1}^{l} N_{t,g} \times (l-t+1)$.

MDP formulation. With this model, we can design appropriate state vector such that our model is an MDP. Note that, besides the posteriors at time l, the posterior at time $l+1$ only depends on the sample size, the data variance, and the priors. All of these components are regarded as known in our approach. Therefore, the history can be fully represented by the parameters of the posteriors at time l, i.e., $(\overline{W}_{l,T}, \overline{W}_{l,C})$. Hence, all information in the history can be summarized in the state vector $S_t = (S_t^b, I_t)$, where $S_l^b = (\overline{W}_{l,Tr}, \overline{W}_{l,C})^T$ and recall that I_t is the terminal indicator (see Section 5.2.3). With the state vector defined, the full definition of the MDP then follows. The vector $(\mu_{0Tr}, \mu_{0C}, \sigma_{0Tr}, \sigma_{0C}, \sigma_{Tr}, \sigma_C, \{N_{t,Tr}\}, \{N_{t,C}\})$ contains all parameters needed to specify the environment and is the *context vector* that we used in contextual RL (see Section 5.4.2).

6.2 Meta-analysis

Dataset and setup. We collect all historical experiments that run on Amazon's largest experimentation platform in the past two years and have a length of at least 4 weeks. Since we do not know the ground truth of the treatment effect (and hence the correct decision and the impacts), we cannot directly run a real data analysis [1]. Therefore, We design a real data-calibrated simulation study, where we calibrate a distribution of problem instances and study the average performance over them. We also present a simulation study in Appendix B.2 to facilitate reproducibility, since we cannot share the real dataset and its more details due to confidentiality.

For each historical experiment, we keep most variables (sample sizes, sample variances, etc.) the same as in the real trajectory. Regarding the opportunity cost, we first estimate the total saving if we can reduce one day for every experiment (details omitted due to business confidentiality), and then decompose these savings to each experiment based on their sample sizes. We estimate the prior distributions using the empirical Bayes method from historical experiments. We set the maximum duration as $T = 4$ weeks for all experiments. We take the first week of observations from the real data, use the corresponding posterior to generate the ground truth of the treatment effect, and then simulate data based on the formulae in Section 6.1.

Baseline methods. We compare the proposed RL framework with classic statistical methods for early termination. As reviewed in Section 2, all these methods aim to control either the type-I error rate or FDR. For the alpha-spending approach (Frequentist) [10, 15], we use the O'Brien-Fleming spending function in the R package *ldbounds* [1] to control the type I error rate under 0.05. For the sequential bayesian testing procedures [11, 21, 38], we consider three variants: (i) we compute the exact Bayes factor for one-sided hypothesis testing (our analysis shows that the one-sided test outperforms the two-sided one on this dataset), which we refer to as BF; (ii) we use the Posterior Odds (POS) [11] in place of the Bayes factors, which also takes prior odds into consideration; (iii) we use the Jeffrey-Zellner-Siow (JZS) Bayes Factor implemented in the Python package *Pingouin* [45]. Following the guidelines [21, 38], we try three threshold values (3, 10 and 30) for the three variants above. Lastly, we compare with the always valid p-values (AVP) approach [22, 24] based on a mSPRT test with type-I error constraint 0.05. For the proposed RL method, we use a two-layer neural network with 128 hidden nodes as the function class, and use the DQN implementation in *Rllib*[27], an open-source RL package. We also compare with three fixed-horizon procedures, including the frequentist fixed horizon testing with level 0.05 (FFHT), the Bayesian fixed horizon testing (BFHT) which rejects the null when the posterior probability of having a positive effect is larger than 0.66, and the Bayesian fixed-horizon optimal decision rule (BFHOD) which recommends launch when the posterior mean of the gain is positive.

Metrics. We consider the following metrics for comparisons:

(1) The percentage of experiments terminated early and the average number of weeks run for each experiment.
(2) False discovery rate (FDR; #false positives / (#false positives+#true positives)), power (proportion of correctly detecting significantly positive/negative effect when the true effect is indeed positive/negative), and type-I error (false positive rate, i.e. the probability of mistakenly detecting significance when the true effect is not).
(3) The three components in Equation (1), including the opportunity cost, the impact during the experiment (relative to the control), and the launch impact (relative to the control).
(4) The average utility, which includes the three components above and is our main objective.

Results. We present results from 50 thousands trajectories in Table 1. As expected, RL outperforms other methods and generates the highest average cumulative reward, i.e., RL maximizes customer experience. It is not surprising that most baseline methods have low type-I error rate or FDR. However, as mentioned in the introduction, most experiments in our applications are flat, therefore focusing on these metrics turns out to be too conservative. These limitations are particularly clear for frequentist methods or Bayesian early termination methods. The Bayesian fixed-horizon methods (BFHT4 and BFHOD4), though have a great power and a high average reward, need too much opportunity costs since they cannot terminate experiments earlier. Finally, thanks to the contextual RL technique, the computational cost of making a decision for a new experiment is negligible (less than 0.005 second). In conclusion, through the meta-analysis, we found that the proposed approach achieves a desired balance between various metrics.

Policy behavior. Recall that the learned policy is a neural network, mapping from the vector of (experiment's observations, experiment's features, time index) to the recommended optimal

[1] In Appendix B.1, we run on real data with a heuristic way of defining the ground truths. However, the set of results is only used for reference and should not be used to directly compare methods.

Table 1: Meta-analysis results. The number after each method name indicates the tuning parameter being used. We omitted results with some tuning parameters that have very poor performance. Recall that all utility-related metrics share the same unit D, the meaning of which is omitted due to confidentiality.

Method	% Early Terminated Experiments	Type I	Power	FDR	Average Weeks	Average Opportunity Cost (D)	Average Launch Impact (D)	Average Experiment Impact (D)	**Average Cumulative Reward** (D)
FFHT	0.0%	0.0%	0.43	0.0%	4.0	2.63	4.31	-0.0	1.68(0.07)
alpha-spending	27.32%	0.0%	0.42	0.0%	3.64	2.48	4.24	-0.0	1.83(0.07)
BFHT	0.0%	0.03%	0.86	0.04%	4.0	2.63	6.28	-0.0	3.64(0.07)
BFHOD	0.0%	5.2%	0.95	5.4%	4.0	2.63	6.35	-0.002	3.72(0.06)
BF 3	89.91%	0.39%	0.52	0.74%	1.7	0.69	4.4	0.04	3.91(0.06)
BF 10	61.65%	0.02%	0.47	0.05%	2.69	1.85	4.52	0.05	2.86(0.07)
POS 3	89.94%	0.39%	0.52	0.74%	1.7	0.68	4.39	0.04	3.9(0.06)
JZS 3	11.17%	0.0%	0.17	0.0%	3.84	2.61	1.67	-0.0	-0.91(0.05)
AVP	16.15%	0.01%	0.41	0.02%	3.73	2.5	4.09	-0.05	1.64(0.07)
RL	88.48%	7.01%	0.95	5.24%	2.31	0.53	5.18	-0.01	**4.64**(0.05)

action. It is practically useful to provide interpretability of this policy to experimenters. One way to do that is by looking into how the recommended actions change with different inputs. To illustrate, we fix down one historical experiment at time $t = 2$, then vary its opportunity cost and the posterior mean of ATE around the corresponding observed values while keeping the other variables fixed, and see how does the recommended action change.

We present the results in Figure 4, where the findings are overall reasonable. Along the y-axis, when the absolute value of the ATE posterior mean is away from zero, it implies the launch decision is less uncertain, and hence we observe the policy is more inclined towards "stop". Along the x-axis, when the opportunity cost decreases, it implies the cost of keeping running the experiment decreases, hence we observe the policy is more inclined towards "keep running". Similar findings on the relationship with other variables such as the noise level are observed as well.

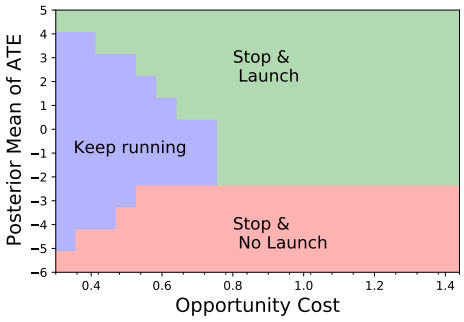

Figure 4: Trend of the recommended action with the opportunity cost and the posterior mean of the treatment effect.

7 CONCLUSION

In this paper, we propose an RL-based approach for continuously monitoring online A/B experiments. Unlike traditional statistical approaches, our method aims to maximize the expected utility that consider a few different factors. We introduce in detail how we formulate this problem at Amazon, and also discuss how to solve the policy by RL algorithms. With a large-scale meta-analysis using past experiments from a large experimentation platform in Amazon, we find that the proposed approach leads to a significant gain in the expected utility.

The task of making sequential decisions for online A/B experiments is challenging, as different experiments can vary a lot in their properties and it is not practical to tailor for tens of thousands of experiments. There are a few meaningful extensions of our framework. First, in some cases, the treatment effect is not homogeneous over time. We can model the uncertainty with a Bayesian time series model, from which we can solve a conservative and robust early termination rule. Second, it is practically meaningful to integrate statistical inference and optimal decision making. We propose a few approaches in Section 5.3, which we will investigate numerically as our next step. Third, we estimate the priors from historical experiments, which can guarantee the optimality on average. It is useful to study the impact of the prior specifications. Last, we use a simple contextual RL algorithm to solve the problem. Recently, there are more advanced algorithms being proposed (e.g., [12] and references therein) and can potentially further improve the policy learning performance.

REFERENCES

[1] Charlie Casper, Thomas Cook and Oscar A. Perez. 2022-02. *An R Package for Group Sequential Boundaries Using Alpha Spending Functions.* https://cran.r-project.org/web/packages/ldbounds/index.html.
[2] Linda Anderson. 2016. Library website visits and enrollment trends. *Evidence Based Library and Information Practice* 11, 1 (2016), 4–22.

[3] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research* 39, 1 (2020), 3–20.

[4] Bram Bakker. 2001. Reinforcement learning with long short-term memory. *Advances in neural information processing systems* 14 (2001).

[5] Carolin Benjamins, Theresa Eimer, Frederik Schubert, André Biedenkapp, Bodo Rosenhahn, Frank Hutter, and Marius Lindauer. 2021. CARL: A benchmark for contextual and adaptive reinforcement learning. *arXiv preprint arXiv:2110.02102* (2021).

[6] James O Berger. 2013. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

[7] Kenneth P Burnham and Walter Scott Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65, 3 (1978), 625–633.

[8] George Casella and Roger L Berger. 2021. *Statistical inference*. Cengage Learning.

[9] Kathryn Chaloner and Isabella Verdinelli. 1995. Bayesian experimental design: A review. *Statistical science* (1995), 273–304.

[10] David L Demets and KK Gordon Lan. 1994. Interim analysis: the alpha spending function approach. *Statistics in medicine* 13, 13-14 (1994), 1341–1352.

[11] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE, 243–252.

[12] Theresa Eimer, André Biedenkapp, Frank Hutter, and Marius Lindauer. [n. d.]. Towards Self-Paced Context Evaluation for Contextual Reinforcement Learning. ([n. d.]).

[13] Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. 2021. Deep adaptive design: Amortizing sequential bayesian experimental design. In *International Conference on Machine Learning*. PMLR, 3384–3395.

[14] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

[15] KK Gordon Lan and David L DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70, 3 (1983), 659–663.

[16] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[17] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.

[18] Assaf Hallak, Dotan Di Castro, and Shie Mannor. 2015. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259* (2015).

[19] Xun Huan and Youssef M Marzouk. 2016. Sequential Bayesian optimal experimental design via approximate dynamic programming. *arXiv preprint arXiv:1604.08320* (2016).

[20] Desi R Ivanova, Adam Foster, Steven Kleinegesse, Michael U Gutmann, and Thomas Rainforth. 2021. Implicit deep adaptive design: policy-based experimental design without likelihoods. *Advances in Neural Information Processing Systems* 34 (2021), 25785–25798.

[21] Harold Jeffreys. 1961. *The theory of probability*. Oxford University Press.

[22] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1517–1525.

[23] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2022. Always valid inference: Continuous monitoring of a/b tests. *Operations Research* 70, 3 (2022), 1806–1821.

[24] Ramesh Johari, Leo Pekelis, and David J Walsh. 2015. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922* (2015).

[25] Prashant Kadam and Supriya Bhalerao. 2010. Sample size calculation. *International journal of Ayurveda research* 1, 1 (2010), 55.

[26] Madan Gopal Kundu, Sandipan Samanta, and Shoubhik Mondal. 2021. Conditional power, predictive power and probability of success in clinical trials with continuous, binary and time-to-event endpoints. (2021).

[27] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. 2018. RLlib: Abstractions for distributed reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3053–3062.

[28] Winston Lin. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. (2013).

[29] James K Lindsey et al. 1999. Models for repeated measurements. *OUP Catalogue* (1999).

[30] Johannes S Maritz. 2018. *Empirical bayes methods*. Chapman and Hall/CRC.

[31] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[33] Peter Müller, Don A Berry, Andy P Grieve, Michael Smith, and Michael Krams. 2007. Simulation-based sequential Bayesian design. *Journal of statistical planning and inference* 137, 10 (2007), 3140–3150.

[34] Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. 2001. Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* 96, 456 (2001), 1410–1423.

[35] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[36] Thomas S Richardson, Yu Liu, James McQueen, and Doug Hains. 2022. A Bayesian Model for Online Activity Sample Sizes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1775–1785.

[37] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. 2017. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging* 2017, 19 (2017), 70–76.

[38] Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods* 22, 2 (2017), 322.

[39] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*. PMLR, 1889–1897.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[41] Wanggang Shen and Xun Huan. 2021. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *arXiv preprint arXiv:2110.15335* (2021).

[42] Shagun Sodhani, Amy Zhang, and Joelle Pineau. 2021. Multi-Task Reinforcement Learning with Context-based Representations. *arXiv preprint arXiv:2102.06177* (2021).

[43] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[44] Mauricio Tec, Yunshan Duan, and Peter Müller. 2022. A Comparative Tutorial of Bayesian Sequential Design and Reinforcement Learning. *arXiv preprint arXiv:2205.04023* (2022).

[45] Raphael Vallat. 2018. Pingouin: statistics in Python. *J. Open Source Softw.* 3, 31 (2018), 1026.

[46] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[47] Nikos Vlassis, Mohammad Ghavamzadeh, Shie Mannor, and Pascal Poupart. 2012. Bayesian reinforcement learning. *Reinforcement learning* (2012), 359–386.

[48] Abraham Wald and Jacob Wolfowitz. 1948. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics* (1948), 326–339.

[49] Runzhe Wan, Xinyu Zhang, and Rui Song. 2021. Multi-objective model-based reinforcement learning for infectious disease control. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1634–1644.

[50] Bernard L Welch. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* 34, 1-2 (1947), 28–35.

A DERIVATIONS

A.1 Predictive distribution for Normal-Normal model

By definition, the number of new customers enrolled in the experiment during week k and assigned to group $* \in \{Tr, C\}*$ is represented by $N_{k,*}$. The observed activities of customer i who first enrolled at week k are represented by $W_{i,k}, .., W_{i,T}$. Sample mean of aggregated responses among customer participated in the experiments up to week l is defined as:

$$\begin{aligned}\overline{W}_{l,*} &= \frac{\sum_{i\in\mathcal{I}_{1:l,*}} W_{i,1:l} + \sum_{i\in\mathcal{I}_{2:l,*}} W_{i,2:l} \cdots + \sum_{i\in\mathcal{I}_{l:l,*}} W_{i,l:l}}{\sum_{t=1}^{l} N_{t,*}} \\ &= \frac{\sum_{t=1}^{l}\sum_{i\in\mathcal{I}_{t:l,*}} W_{i,t:l,*}}{\sum_{t=1}^{l} N_{t,*}}\end{aligned}$$

Following equation (4), we had:

$$\overline{W}_{l,*}|\mu_* \sim N(\mu_* * a_*(l), \sigma_*^2 b_*(l))$$

where $a_*(l) = \frac{c_*(l)}{N_{1:l,*}}$, $b_*(l) = \frac{c_*(l)}{N^2_{1:l,*}}$, $c_*(l) = \sum_{t=1}^{l} N_{t,*} \times (l - t + 1)$. Given $\overline{W}_{l,*}$ (one data point each arm at week l), posterior distributions of μ_* are:

$$\begin{aligned}\mu_*|\overline{W}_{l,*} &\propto exp\left(-\frac{1}{2}\left(\frac{\mu_* - \mu_{0*}}{\sigma_{0*}}\right)^2\right) exp\left(-\frac{1}{2}\left(\frac{\overline{W}_{l,*} - \mu_* * a_*(l)}{\sigma_*\sqrt{b_*(l)}}\right)^2\right) \\ &\sim N\left(\left(\frac{1}{\sigma_{0*}^2} + \frac{a_*(l)^2}{\sigma_*^2 b_*(l)}\right)^{-1}\left(\frac{\mu_{0*}}{\sigma_{0*}^2} + \frac{a_*(l)\overline{W}_{l,*}}{\sigma_*^2 b_*(l)}\right), \left(\frac{1}{\sigma_{0*}^2} + \frac{a_*(l)^2}{\sigma_*^2 b_*(l)}\right)^{-1}\right).\end{aligned}$$

Model-based RL needs the ability to simulate the state transition, or roughly speaking, the capability to simulate $\overline{W}_{l+1,*}|\overline{W}_{l,*}$. Note that $f(\overline{W}_{l+1,*}|\overline{W}_{l,*}) = \int f(\mu_*|\overline{W}_{l,*}) f(\overline{W}_{l+1,*}|\mu_*, \overline{W}_{l,*}) d\mu_*$. Therefore, we only need to first sample μ_* from $\mu_*|\overline{W}_{l,*}$, and then sample $\overline{W}_{l+1,*}$ from $\overline{W}_{l+1,*}|\mu_*, \overline{W}_{l,*}$. We have

$$\overline{W}_{l+1,*}|\mu_*, \overline{W}_{l,*} \sim N\left(\frac{\overline{W}_{l,*} * N_{1:l,*} + \mu_T * N_{1:l+1,*}}{N_{1:l+1,*}}, \frac{1}{N_{1:l+1,*}}\sigma_*^2\right).$$

A.2 Role of information Gain

A.2.1 Proof the fixed-horizon Bayesian decision rule. Consider a fixed-horizon Bayesian optimal decision scenario where a final decision is made at T, if is no opportunity cost for running the experiment (i.e., $c = 0$) and ignore the running experiment impact (third term in Eq 2), then the objective becomes:

$$\begin{aligned}&\mathbb{E}_\rho^\pi\Big[\sum_{t=1}^{T} R_t|\mathcal{H}_1 = h\Big] = \mathbb{E}_\rho^\pi\Big[R_T|\mathcal{H}_1 = h\Big] \\ =\ &\mathbb{E}_\rho\Big[(\mathbb{E}_\rho(\delta \mid \mathcal{H}_T) \times H \times N - c_h) \times \mathbb{I}\big[\pi_T(\mathcal{H}_T) = 1\big] \mid \mathcal{H}_1 = h\Big]\end{aligned}$$

where the first equality is due to that only the final decision is allowed, and the second equality is from our reward definitions above. The optimal policy is hence $\pi_T(\mathcal{H}_T) = 2 - \mathbb{I}\big[\mathbb{E}_\rho(\delta \mid \mathcal{H}_T) \times H \times N > c_h\big]$

A.2.2 Proof of Lemma 1.

$$\begin{aligned}&\mathbb{E}_\rho^{\pi_2}\Big[\sum_{t=t'}^{T} R_t \mid \mathcal{H}_{t'} = h\Big] - \mathbb{E}_\rho^{\pi_1}\Big[\sum_{t=t'}^{T} R_t \mid \mathcal{H}_{t'} = h\Big] \\ =\ &\mathbb{E}_\rho^{\pi_2}\Big[R'_t + R_{t'+1} \mid \mathcal{H}_{t'} = h\Big] - \mathbb{E}_\rho^{\pi_1}\Big[R'_t \mid \mathcal{H}_{t'} = h\Big] \\ =\ &\mathbb{E}_\rho^{\pi_2}\Big[R_{t'+1} \mid \mathcal{H}_{t'} = h\Big] - \mathbb{E}_\rho^{\pi_1}\Big[R'_t \mid \mathcal{H}_{t'} = h\Big] \\ -\ &c + \mathbb{E}_\rho^{\pi_1}\Big[R'_t \mid \mathcal{H}_{t'} = h\Big] * N_{1:t'+1,Tr} \\ &\text{(Plug in Equation (3))} \\ =\ &\mathbb{E}_\rho\Big[\big(\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times (H + T - t' - 1) \times N\big) \\ &\times\mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\ &-\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times (H + T - t') \times N \times \mathbb{I}[\pi_{1,t}(h) = 1] \\ &-c + \mathbb{E}_\rho^{\pi_1}\Big[R'_t \mid \mathcal{H}_{t'} = h\Big] * N_{1:t'+1,Tr} \\ =\ &\Big\{\mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\ &-\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'} = h\big] \times \mathbb{I}[\pi_{2,t}(h) = 1]\Big\} \times N \times (H + T - t') \\ &-\mathbb{E}_\rho\Big[\mathbb{E}_\rho\big[\delta \mid \mathcal{H}_{t'+1}\big] \times \mathbb{I}[\pi_{2,t}(\mathcal{H}_{t'+1}) = 1] \mid \mathcal{H}_{t'} = h\Big] \\ &-c + \mathbb{E}_\rho^{\pi_1}\Big[R'_t \mid \mathcal{H}_{t'} = h\Big] * N_{1:t'+1,Tr}.\end{aligned}$$

B ADDITIONAL NUMERICAL DETAILS AND RESULTS

B.1 Real data analysis with heuristic ground truths

In this section, we directly run different methods on historical trajectories in and show the results in Table 2. However, Since we do not know the ground truth of the treatment effect (and hence the correct decision and the impacts), it is impossible to evaluate various methods. We adopt a heuristic approach that uses the posterior mean after 4 weeks as the ground truth. Accordingly, *all decision accuracy-related metrics (type-I error, power, FDR) should be understood as the differences with the fixed-horizon Bayesian decision rule, and all utility-related metrics are measured using the posterior mean.* We emphasize that, this is a heuristic approach to provide users a sense of what may happen compared with running the full horizon of experiments, and we do not recommend over-interpreting this set of results.

We apply an affine transformation on the utility-related components and use another unit E, due to confidentiality consideration. This does not affect the conclusions. We do not compare with fixed-horizon Bayesian procedure, as by design they are directly based on the ground truths. From the results, we can see that RL yields significantly higher (empirical) utility, consistent with our design. In particular, as mentioned in the introduction, the signal in our applications is typically weak and existing methods are too conservative.

Table 2: Meta-analysis results on real experiments with heuristic-based ground truths. The number after each method name indicates the tuning parameter being used. We omitted results with some tuning parameters that have very poor performance.

Method	% Early Terminated Experiments	(Empirical) Type I	(Empirical) Power	(Empirical) FDR	Average Weeks	(Empirical) Average Opportunity Cost (E)	(Empirical) Average Launch Impact (E)	(Empirical) Average Experiment Impact (E)	(Empirical) Average Cumulative Reward (E)
FFHT	0.0%	0.0%	0.05	0.0%	4.0	0.31	0.58	-0.02	0.25(0.12)
alpha-spending	1.94%	0.0%	0.05	0.0%	3.97	0.31	0.58	-0.02	0.25(0.12)
BF 3	83.68%	0.9%	0.05	15.0%	1.71	0.0	0.15	0.01	0.16(0.03)
BF 10	57.73%	0.15%	0.01	10.0%	2.51	0.02	0.07	0.02	0.07(0.02)
BF 30	32.32%	0.08%	0.0	14.29%	3.23	0.06	0.02	0.03	-0.02(0.01)
POS 3	84.62%	0.83%	0.05	14.86%	1.68	0.0	0.15	0.01	0.16(0.03)
JZS 3	0.43%	0.0%	0.0	0.0%	3.99	0.31	0.01	-0.02	-0.32(0.02)
AVP	0.31%	0.15%	0.01	22.22%	3.99	0.31	0.03	-0.02	-0.3(0.03)
RL	98.96%	27.48%	0.61	32.62%	1.81	0.03	1.38	-0.01	**1.34**(0.32)

Table 3: Simulation results.

Method	% Early Terminated Experiments	Type I	Power	FDR	Average Weeks	Average Opportunity Cost (E)	Average Launch Impact (E)	Average Experiment Impact (E)	Average Cumulative Reward (E)
FFHT 4	0.0%	0.0%	0.0	0.01%	4.0	0.15	0.24	-0.0	0.1(0.01)
alpha-spending	15.23%	0.01%	0.0	0.02%	3.82	0.14	0.24	-0.0	0.11(0.01)
BFHT	0.0%	0.1%	0.01	0.13%	4.0	0.15	0.39	-0.0	0.24(0.01)
BF 3	90.83%	0.49%	0.01	0.38%	1.64	0.03	0.26	-0.0	0.24(0.02)
BF 10	61.27%	0.24%	0.01	0.25%	2.6	0.08	0.35	-0.0	0.28(0.02)
BF 30	38.3%	0.1%	0.01	0.15%	3.22	0.11	0.34	-0.0	0.24(0.02)
POS 3	90.4%	0.48%	0.01	0.37%	1.66	0.03	0.26	-0.0	0.24(0.02)
JZS 3	7.2%	0.0%	0.0	0.01%	3.89	0.14	0.13	-0.0	-0.01(0.01)
AVP	33.37%	0.14%	0.01	0.18%	3.24	0.11	0.35	-0.01	0.25(0.02)
RL	98.90%	25.36%	0.72	27.37%	2.33	0.06	0.38	-0.0	0.32(0.02)

B.2 Simulated data analysis

We simulate 3000 experiment trajectories and compare the performance of proposed RL model with baseline methods in Table 3. The data generation process is as follows.

- Sample size. We assume the number of customers shown in each week follows the beta-geometric model. ([36] without the censoring in the likelihood function.) The parameters α and β of the beta distribution were randomly simulated from a uniform distribution, with $\alpha \sim Uniform(0.1, 1)$ and $\beta \sim Uniform(4, 60)$. Assume the total number of customers is 10K.
- Sample weekly responses for 4 weeks. We followed the normal-normal model in (4), with parameters as $\mu_{0C} = 0.1$, $\sigma_{0C} = 2$, $\mu_{0Tr} = 0.1$, $\sigma_{0Tr} = 2.83$ and $\sigma_C = \sigma_T = 100$.
- The total weekly opportunity cost is 1.5×10^8, decomposed to each experiment according to their sample sizes. Huddle costs are set as 0.