A Bayesian Approach for Subset Selection in Contextual Bandits

Jialian Li[1], **Chao Du**[2], **Jun Zhu**[*1]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Lab, Bosch-Tsinghua Joint ML Center, Tsinghua University
[2]Alibaba Group
lijialia16@mails.tsinghua.edu.cn, duchao0726@gmail.com, dcszj@tsinghua.edu.cn

Abstract

Subset selection in Contextual Bandits (CB) is an important task in various applications such as advertisement recommendation. In CB, arms are attached with contexts and thus correlated in the context space. Proper exploration for subset selection in CB should carefully consider the contexts. Previous works mainly concentrate on the best one arm identification in linear bandit problems, where the expected rewards are linearly dependent on the contexts. However, these methods highly rely on linearity, and cannot be easily extended to more general cases. We propose a novel Bayesian approach for subset selection in general CB where the reward functions can be nonlinear. Our method provides a principled way to employ contextual information and efficiently explore the arms. For cases with relatively smooth posteriors, we give theoretical results that are comparable to previous works. For general cases, we provide a calculable approximate variant. Empirical results show the effectiveness of our method on both linear bandits and general CB.

Introduction

Subset selection (Soare, Lazaric, and Munos 2014) is an important task in decision-making. In this paper, we consider the subset selection task in a contextual bandit (CB) setting, where we have some featured arms, and the goal is to identify an optimal subset of arms by interacting with them and gaining observations. Practical problems like advertisement recommendation (Li et al. 2010) fall into this category. For example, a new advertiser might wish to identify a set of customers who are most probable to be interested in her advertisements. The goal is to conduct this subset selection with as few as possible interactions with the arms.

Formally, assume that we are facing a set of N arms (denoted by $[N] = \{1, 2, ..., N\}$) where each arm is attached with a context. Each time, we choose a subset of arms as an action set and obtain a corresponding reward observation. Our goal is to identify the optimal target set with a maximum expected reward based on the observations. We denote the sets of action sets and target sets as $\mathcal{I}$ and $\mathcal{Z}$ respectively. In this paper, we focus on the pure exploration setting for subset selection problems. That is, the training process focuses on gaining information for identifying the optimal target set.[1]

Recent works for pure exploration in CB mainly concentrate on single-arm identification for Linear Bandits (LB), where arm contexts and environment parameters are represented as vectors and the expected reward for each arm is the inner product of its context and the environment vector. Frequentist methods (Soare, Lazaric, and Munos 2014; Xu, Honda, and Sugiyama 2018) on LB construct confidence bounds to estimate the uncertainty of the expected rewards of arms. Comparing with Bayesian methods, frequentist methods are usually conservative and empirically less efficient (Eckman and Henderson 2018; Branke, Chick, and Schmidt 2007), since they require all confidence bounds to be small enough.

The Bayesian framework provides a way to handle uncertainty explicitly by maintaining a posterior distribution for parameters, which makes it possible to characterize the probability of being optimal directly. For example, BayesGap (Hoffman, Shahriari, and Freitas 2014) uses a posterior distribution to characterize the environment parameters in LB. When turning to choosing arms, it still constructs bounds for each arm and chooses the best one. However, it fails to exploit the contextual information for arm selection. For example, consider the three-arm LB with arms featured as $\boldsymbol{x}^1 = (1, 0)$, $\boldsymbol{x}^2 = (1, 0.1)$ and $\boldsymbol{x}^3 = (0, 1)$ and environment parameter $(1, 0.1)$ (Xu, Honda, and Sugiyama 2018). Then the key is to identify the optimal one between $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$. For this problem, pulling $\boldsymbol{x}^3$ can be more efficient in distinguishing $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ than pulling either of them. Therefore, in CB, arms with low rewards might be informative in exploration.

These best arm identification methods on LB cannot be extended to general CB subset selection for two reasons: (1) their constructions of confidence bounds highly rely on the linear property, and (2) their arm selection rules for best arm identification cannot be directly applied for general problems with $\mathcal{I}$ and $\mathcal{Z}$.

Considering the above issues, we propose Bayesian Resample Explore (BRE), a novel Bayesian method for sub-

*Corresponding author.
Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]There is another setting considering regret, where the goal is to improve the average performance over the whole training process. Thus the trade-off between exploitation and exploration is needed.

set selection in CB. Instead of using a posterior to estimate the reward upper bounds and find optimal arms as in BayesGap, we use a more flexible re-sampling method to find two near-optimal sets. Then we take actions to try to distinguish the two sets. Two new selection rules are designed to make use of the context correlations among arms. Our Bayesian method can be applied to both fixed-confidence and fixed-budget settings. For fixed-confidence problems, we use the posterior distributions to give a Monte Carlo stopping rule, which is empirically more efficient than bound-related stopping rules. We further provide a non-asymptotic sample complexity analysis for cases where posteriors are relatively smooth. An approximate variant is also presented for problems where posteriors are hard to calculate. We also reduce BRE to LB to compare with baseline methods. Empirical results validate the efficiency of BRE. Specifically, BRE outperforms previous methods like BayesGap and LinGapE (Xu, Honda, and Sugiyama 2018) on LB. 2

Related Work

Multi-Armed Bandit Problems

Multi-armed bandits (MAB) are the simplest bandit problems where arms have independent reward distributions. Kalyanakrishnan et al. (2012) provide LUCB, a frequentist solution for subset selection in MAB based on the classical UCB algorithm (Auer, Cesa-Bianchi, and Fischer 2002). Russo (2016) proposes a Bayesian method to solve the best arm identification in MAB. It modifies the famous regret minimization method Thompson Sampling (TS) (Thompson 1933) by assigning a probability to pull sub-optimal arms. Expected-Improvement (EI) (Henderson and Nelson 2006) is another Bayesian method for MAB, but it is less efficient than the TS-based method (Russo 2016). Some other works consider the fixed budget problem, such as the optimal-computing-budget allocation (OCBA) (Chen et al. 2000). Methods in MAB do not consider the correlation of arms and directly applying them to CB would lead to inefficiency.

Pure Exploration in LB

Many works concentrate on the pure exploration in Linear Bandits and attempt to exploit the arm correlation. Soare, Lazaric, and Munos (2014) construct the stopping condition for LB with the vectors of arms and propose some pre-defined arm selection strategies. Xu, Honda, and Sugiyama (2018) improve the pre-defined methods to a fully adaptive method via using arm correlation to choose actions. However, these frequentists' methods still suffer from a conservative stopping rule as they treat upper bounds of rewards independently and require uniform tightness for all bounds. For Bayesian methods, BayesGap (Hoffman, Shahriari, and Freitas 2014) characterizes the posterior distribution of the environments and gives a fixed-budget solution, but it ignores the exploration potential of sub-optimal arms.

Problem Formulation

Consider the task of subset selection in contextual bandits (CB) with N arms. Each arm $i \in [N]$ has a feature vector $\boldsymbol{x}^i \in \mathbb{R}^{m_x}$ with dimension $m_x \in \mathbb{N}^+$. Suppose that there exists an unknown environment parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{m_\theta}$ where $m_\theta \in \mathbb{N}^+$. At each time step, the algorithm can choose one or more arms and gains corresponding observations. Formally, define a set $\mathcal{I}$ that is composed of some subsets of $[N]$. Each time, a set $I \subseteq [N]$ from $\mathcal{I}$ is chosen and an immediate reward r_I will be returned. Here r_I might be either a scalar or a vector, and there might be noise added to r_I. The expectation of r_I is a function of $\boldsymbol{\theta}^*$ and all $\boldsymbol{x}^i$ for $i \in I$, denoted as $\mathbb{E}[r_I] = f(I, \boldsymbol{\theta}^*)$. The goal of the problem is to identify an optimal subset from some candidate subsets. We denote $\mathcal{Z}$, composed of some subsets of $[N]$, as the set of all candidate subsets. A target function $g(Z, \boldsymbol{\theta}^*)$ is given which maps each $Z \in \mathcal{Z}$ and $\boldsymbol{\theta}^*$ into a value. The goal is to choose a set $Z^* \in \mathcal{Z}$ which satisfies $Z^* = \operatorname{argmax}_{Z \in \mathcal{Z}} g(Z, \boldsymbol{\theta}^*)$. For convenience, we call sets in $\mathcal{I}$ as *action sets* and sets in $\mathcal{Z}$ as *target sets*. Denote $m_I = |\mathcal{I}|$ and $m_Z = |\mathcal{Z}|$.

Note that the observation function f and the target function g are not assumed to be the same, although in most existing works they are. In some practical problems, the testing environment might be a bit different from the training environment. The formulation here with functions f and g can easily distinguish the training and testing rewards.

There exist two settings for the subset selection problems: (1) the *fixed-budget* setting where the number of interactions is fixed; (2) the *fixed-confidence* setting where the algorithm doesn't stop until it guarantees a probability to be correct. Our work focuses on the fixed-confidence setting, following the *ranking and selection* problems (Kim and Nelson 2006). The good selection goal is used for recommendation, which aims to find a set Z such that, $g(Z, \boldsymbol{\theta}^*) > g(Z^*, \boldsymbol{\theta}^*) - \varepsilon$, with a probability no less than $1 - \delta$, for given ε and δ values.

The key for solving the subset selection problems is to choose proper actions so as to efficiently collect information about the unknown parameter $\boldsymbol{\theta}^*$ and to distinguish the optimal Z^* with other sets in $\mathcal{Z}$.

The contextual information in CB can be exploited in two ways. The first is to improve the estimation for $\boldsymbol{\theta}^*$ with observations from all time steps, since the environment parameters are shared. It is natural to use the Bayesian framework to estimate the uncertainty of $\boldsymbol{\theta}^*$. This is also the key idea for BayesGap in LB. The second way is to use contexts to help choose action sets, since arms are correlated by contexts. This is the point that BayesGap ignores.

Linear Bandits: If we pull one of the arm at each time step, we denote the action set as $\mathcal{I}_{arm} := \{\{i\}\}_{i=1}^N$. If $f(I, \boldsymbol{\theta}^*) = g(I, \boldsymbol{\theta}^*) = \sum_{i \in I} \boldsymbol{x}^{i\top} \boldsymbol{\theta}^*$ for any $I \in \mathcal{I}_{arm}$, this reduces to the Linear Bandit problem (LB).

Explore-K: If $\mathcal{Z}$ is composed of sets with K arms, this corresponds to Explore-K problems proposed in (Kalyanakrishnan and Stone 2010). Denote $\mathcal{Z}_K := \{Z \subseteq [N] : |Z| = K\}$ as the set of target sets for Explore-K.

Bayesian Framework

We now present a Bayesian framework for CB. We assume a prior distribution on the unknown environment parameters, which can be a non-informative prior if we do not have inductive bias.

Denote the chosen action set and the corresponding re-

ward at time step t as I_t and r_t respectively. Denote this pair as $d_t = (I_t, r_t)$ for conciseness. With the information $\mathcal{D}_t = \{d_s\}_{s=1}^t$, we attain a posterior distribution $p_t(\boldsymbol{\theta})$:

$$p_t(\boldsymbol{\theta}) := p(\boldsymbol{\theta}|\mathcal{D}_t) = \frac{p(\mathcal{D}_t|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{p(\mathcal{D}_t)} \propto \prod_{s=1}^{t} p(d_s|\boldsymbol{\theta})p_0(\boldsymbol{\theta}). \quad (1)$$

We consider a probability of good selection goal (PGS) (Eckman and Henderson 2018). Formally, for parameter $\boldsymbol{\theta}$, denote the event that set Z ($Z \in \mathcal{Z}$) is the ε-optimal set under $\boldsymbol{\theta}$ as $\mathcal{E}^{\varepsilon}(Z;\boldsymbol{\theta})$. That is, if $\mathcal{E}^{\varepsilon}(Z;\boldsymbol{\theta})$ is true, then $\forall Z' \in \mathcal{Z}$, $g(Z,\boldsymbol{\theta}) \geq g(Z',\boldsymbol{\theta}) - \varepsilon$. At time step t, the posterior p_t encodes the belief for the possible distribution over $\boldsymbol{\theta}^*$. Thus the probability that Z is the ε-optimal target set is defined as

$$\mathcal{P}_t^{\varepsilon}(Z) = \int_{\boldsymbol{\theta}} \mathbb{I}[\mathcal{E}^{\varepsilon}(Z;\boldsymbol{\theta})]p_t(\boldsymbol{\theta})d\boldsymbol{\theta}. \quad (2)$$

Here $\mathbb{I}[\cdot]$ denotes the indicator function.

The goal of the algorithm is to recommend a target set after some number of interactions with the environment. Following Kaufmann, Cappé, and Garivier (2016), a Bayesian subset selection algorithm can be divided into three rules:

Action set selection rule: At time step $t > 0$, with posterior $p_{t-1}(\boldsymbol{\theta})$, the algorithm selects one action set from $\mathcal{I}$.

Stopping rule: The algorithm stops at time step τ, if there exits a set $Z_\tau \in \mathcal{Z}$ such that $\mathcal{P}_\tau^{\varepsilon}(Z_\tau) > 1 - \delta$.

Recommendation rule: When the algorithm stops at time step τ, we recommend the set $\operatorname{argmax}_{Z\in\mathcal{Z}} \mathcal{P}_\tau^{\varepsilon}(Z)$.

There are three main advantages for the Bayesian framework. Firstly, the posterior distribution naturally characterizes the uncertainty of the environment parameters. For complex f or g functions, the construction of upper confidence bounds can be hard in frequentist's methods, while the posterior calculation simply follows the Bayes theorem. Further, many approximate inference methods can be applied for calculating posteriors with complex likelihood.

Secondly, frequentist's methods usually suffer from worse empirical performance because their stopping rule requires all bounds to be tight enough. Bayesian methods, in contrast, consider the joint distribution. This empirical gap can be shown from the comparison between UCB1 and TS in MAB (Chapelle and Li 2011).

The third appealing advantage comes from the posterior PGS guarantee. Current frequentist-based selection procedures require the fixed-confidence δ to be given and they need to guarantee high probability correctness for any configuration, making them conservative. Bayesian methods only use δ for the stopping rule and can be adaptive and sample efficient (Eckman and Henderson 2018).

However, two main challenges arise in implementing a Bayesian subset selection algorithm. The first is how to construct the subset selection rule with p_t so as to efficiently explore the environment. It is still not clear how we make use of p_t and context information to enlighten exploration. The other challenge is how to identify the stopping time since the exact probability $\mathcal{P}_t^{\varepsilon}$ can be hard to calculate or even intractable. Below we propose Bayesian Re-sample Exploration (BRE), a Bayesian approach to solve these challenges.

Method

We first propose the action set selection rule and the corresponding stopping rule. Then we introduce the whole BRE and theoretical results.

Action Set Selection Rule

The key for action set selection is to choose proper sets so as to separate Z^* with other target sets. Following the idea of Thompson Sampling (TS) (Russo et al. 2018), for time step t, a sampled $\boldsymbol{\theta}_t$ from the posterior distribution p_{t-1} is used to find its corresponding optimal target set Z_t. However, Z_t is not enough for exploration guidance. Russo (2016) in MAB uses a re-sample technique to assign a probability to choose sub-optimal arms. Inspired by this, we re-sample to find another potentially optimal target set. Since concentrating on the optimal one target set will allocate too many resources on the estimated optimal one, we turn to find two near-optimal candidate sets for exploration. The reason for choosing two sets is that the comparison between two candidates is easy to conduct. For more than two candidates, it is not straightforward to define an optimization target. Formally, we repeatedly sample $\boldsymbol{\theta}_t'$ until another potential optimal target set Z_t' is found such that $g(Z_t', \boldsymbol{\theta}_t') > g(Z_t, \boldsymbol{\theta}_t') + \varepsilon$.

Denote the indices Z_t and Z_t' as i_t and j_t. For any $i, j \in [m_Z]$, we define a new value

$$\kappa^{ij}(\boldsymbol{\theta}) = g(Z_i, \boldsymbol{\theta}) - g(Z_j, \boldsymbol{\theta}). \quad (3)$$

Denote $\hat{\kappa}_t^{ij}$ as the expectation of $\kappa^{ij}(\boldsymbol{\theta})$ over p_{t-1} and define $\kappa_t(\boldsymbol{\theta}) = \kappa^{i_t j_t}(\boldsymbol{\theta})$ if $\hat{\kappa}_t^{i_t j_t} \geq 0$ and $\kappa_t(\boldsymbol{\theta}) = \kappa^{j_t i_t}(\boldsymbol{\theta})$ otherwise. Denote $\hat{\kappa}_t$ as the posterior mean for $\kappa_t(\boldsymbol{\theta})$.

The goal for time step t turns to be increasing the probability of the event $\kappa_t(\boldsymbol{\theta}) > 0$. Below we propose two selection rules: (1) a general selection rule for general CB problems; (2) a ratio-based action selection rule for problems under specific assumptions.

General selection rule For a general CB problem, with Z_t and Z_t' chosen at time step t, we assume that action set I is chosen and reward r_I is returned. Then the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}_{t-1}, (I, r_I))$ can be calculated. Recall that we use the PGS setting where an error of ε is tolerable. Hence the probability of $\kappa_t(\boldsymbol{\theta}) > -\varepsilon$ becomes

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_{t-1},(I,r_I))}[\mathbb{I}[\kappa_t(\boldsymbol{\theta}) > -\varepsilon]]. \quad (4)$$

Notice that here the calculation of $p(\boldsymbol{\theta}|\mathcal{D}_{t-1}, (I, r_I))$ explicitly exploits the context information.

More specifically, the influence of each action set on the posteriors can be estimated using the context-related likelihood. Thus we can find the action set that is most informative for exploration.

Actually, r_I is not observable before actual actions, so we take expectation over the marginal distribution $p(r_I|\mathcal{D}_{t-1})$. Therefore, the estimation for κ_t after choosing I is

$$V_t(I) := \mathbb{E}_{p(r_I|\mathcal{D}_{t-1})}\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D}_{t-1},(I,r_I))}[\mathbb{I}[\kappa_t(\boldsymbol{\theta}) > -\varepsilon]]. \quad (5)$$

The *general subset selection rule* at time step t is

$$I_t = \operatorname*{argmax}_{I\in\mathcal{I}} V_t(I). \quad (6)$$

Further, when $V_t(I)$ is hard to calculate or even intractable, it might be possible to find a lower bound instead. Then the action set selection rule turns to maximize the lower bound term. One convenient construction of a lower bound is the posterior variance. Formally, consider a distribution $p(\boldsymbol{\theta})$ and we have

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{\theta})}[\mathbb{I}[\kappa_t(\boldsymbol{\theta}) > -\varepsilon]] \geq & \mathbb{E}_{p(\boldsymbol{\theta})}\left[\mathbb{I}[|\kappa_t(\boldsymbol{\theta}) - \hat{\kappa}_t| < \hat{\kappa}_t + \varepsilon]\right] \\ = & 1 - \mathbb{E}_{p(\boldsymbol{\theta})}\left[\mathbb{I}[|\kappa_t(\boldsymbol{\theta}) - \hat{\kappa}_t| \geq \hat{\kappa}_t + \varepsilon]\right] \\ \geq & 1 - \frac{\mathrm{Var}_p(\kappa_t(\boldsymbol{\theta}))}{(\hat{\kappa}_t + \varepsilon)^2} \geq 1 - \frac{\mathrm{Var}_p(\kappa_t(\boldsymbol{\theta}))}{\varepsilon^2}.\end{aligned}$$

The first inequality holds by adding $\hat{\kappa}_t$ to both sides; the second inequality holds with the Markov inequality. Now we propose another *general action set selection rule*

$$I_t = \operatorname*{argmin}_{I \in \mathcal{I}} \mathbb{E}_{p(r_I|\mathcal{D}_{t-1})} \mathrm{Var}_{p(\boldsymbol{\theta}|\mathcal{D}_{t-1},(I,r_I))}(\kappa_t(\boldsymbol{\theta})). \quad (7)$$

Ratio-based selection rule If the posteriors enjoy some nice properties, another ratio-based selection rule with nice theoretical guarantees can be given.

First we define some notations for assumptions. For $t > 0$, denote $n_t(I)$ to be the number of times that set $I \in \mathcal{I}$ has been chosen. Denote the probability simplex in $\mathbb{R}^m$ as Σ^m and then we define a vector $\boldsymbol{w}_t \in \Sigma^{m_I}$ where its kth element $w_{t,k} = n_t(I_k)/t$. For clarity, bold letters represent vectors.

Intuitively we expect the posterior for κ_t converges as t grows. Hence we define a *variance proxy function*:

Definition 1. *For $i, j \in [m_Z]$ and $i \neq j$, function $h^{ij}(t, \boldsymbol{w})$ is called a variance proxy function if it satisfies*

- *$h^{ij}(t_1, \boldsymbol{w}_{t_1}) \leq h^{ij}(t_2, \boldsymbol{w}_{t_2})$ if $t_1 \geq t_2$;*
- *there exits a $\tilde{\boldsymbol{w}}^{ij}$ and a decreasing function $\gamma^{ij}(t)$ such that for $t > 0$, $h^{ij}(t, \tilde{\boldsymbol{w}}^{ij}) \leq \gamma^{ij}(t)$.*

Based on h^{ij}, we give two assumptions below.

Assumption 1. *Up to time step t, with posterior p_t for $\boldsymbol{\theta}$, with a probability no less than $1 - \delta'$, for any $i, j \in [m_Z]$ and $i \neq j$, assume that there is a function $\zeta^{ij}(\delta')$ such that*

$$|\hat{\kappa}_t^{ij} - \kappa^{ij}(\boldsymbol{\theta}^*)| / \sqrt{h^{ij}(t, \boldsymbol{w}_t)} \leq \zeta^{ij}(\delta'). \quad (8)$$

Assumption 2. *At time step t, with posterior p_t for $\boldsymbol{\theta}$, for any $i, j \in [m_Z]$ and $i \neq j$, for some $\varepsilon' > 0$ and a non-increasing function $\ell^{ij}(\varepsilon')$, assume that*

$$p(|\kappa^{ij}(\boldsymbol{\theta}) - \hat{\kappa}_t^{ij}| / \sqrt{h^{ij}(t, \boldsymbol{w}_t)} > \varepsilon') \leq 2\exp(-\varepsilon'^2/2); \quad (9)$$

$$p((\kappa^{ij}(\boldsymbol{\theta}) - \hat{\kappa}_t^{ij}) / \sqrt{h^{ij}(t, \boldsymbol{w}_t)} > \varepsilon') \geq \ell^{ij}(\varepsilon'). \quad (10)$$

Assumption 1 indicates that the prior distribution is chosen such that the posteriors can converge to the true values. Note that we only assume the existence of $\tilde{w}$ and exact construction depends on concrete problems. For problems with nice properties, there are many simple choices for $\tilde{w}$, as we will show in a later section. Assumption 2 requires the posterior to converge neither too fast nor too slow. Simple problems like Gaussian stochastic bandits or linear bandits can easily satisfy the assumptions.

With above assumptions, we propose a *ratio-based action set selection rule*:

$$I_t = \operatorname*{argmin}_{k \in [m_I]: \tilde{w}_k^{ij} > 0} \frac{n_{t-1}(I_k)}{\tilde{w}_k^{ij}}. \quad (11)$$

The key idea is to make the action set ratios close to $\tilde{\boldsymbol{w}}^{ij}$, which is firstly used in the linear bandit algorithm LinGapE (Xu, Honda, and Sugiyama 2018). We extend it to CB with variance proxy function h^{ij} to characterize the converge property.

Stopping Rule

Usually, probability $\mathcal{P}_t^{\varepsilon}(Z)$ can be hard to calculate. Some work in MAB (Branke, Chick, and Schmidt 2007) uses the Slepian's inequality (Slepian 1962) to derive a lower bound of posterior PGS. This lower bound is constructed over arm pairs. However, since arms in contextual bandits share the same environment parameter, it would be better to directly consider $\mathcal{P}_t^{\varepsilon}(Z)$, rather than pair-wise bounds.

Hence based on the re-sample technique in above selection rules, we can use Monte Carlo estimations instead of accurate calculation. Samples from the posterior p_t can be used to estimate $\mathcal{P}_t^{\varepsilon}(Z)$ and check the stopping condition. This approximation is much easier than the exact calculation, at the cost of an extra sampling error.

Formally, assume that at time step τ, we have sampled $\boldsymbol{\theta}_\tau$ for c times, and Z_τ is ε-optimal for c' times. This process can be considered as sampling c data points from a Bernoulli distribution with a probability $\mathcal{P}_\tau^{\varepsilon}(Z_\tau)$ to get 1. Then with the Azuma's inequality, for $\varepsilon' > 0$, we have:

$$P\left(c'/c - \mathcal{P}_\tau^{\varepsilon}(Z_\tau) > \varepsilon'\right) \leq \exp(-2c\varepsilon'^2). \quad (12)$$

We choose $\exp(-2c\varepsilon'^2) = \delta_1$ and $c'/c - \varepsilon \geq 1 - \delta_2$. We can choose $\delta_1 + \delta_2 = \delta$. Therefore our stopping rule satisfies that when BRE stops, with a probability $1 - \delta_1$, $\mathcal{P}_\tau^{\varepsilon}(Z_\tau)$ is no less than $1 - \delta_2$. For proper c, we can find a corresponding c' as our stopping condition. For simplicity we choose $c' = c$ and we have

$$c = \ln(1/\delta_1)/(2\delta_2^2). \quad (13)$$

If we stop, we recommend Z_τ as the optimal set.

Bayesian Re-sample Exploration

Now we are ready to present Bayesian Re-sample Exploration (BRE) in detail. With a prior over $\boldsymbol{\theta}$, we maintain the posterior distribution $p_t(\boldsymbol{\theta})$ using observations up to time step t. Then at time step t, we follow the action set selection rule to choose I_t. If the stopping rule is satisfied at τ, we stop and recommend Z_τ as the optimal target set. Further, BRE can be simply applied in the fixed-budget setting by using the action set selection rules.

If Assumptions 1 and 2 are satisfied, we can use Eq. (11) to choose action sets and call this method BRE-R. If the Eq. (6) or (7) is used, we call this method as BRE-G. The algorithm framework for BRE-G is given in Algorithm 1.

Stopping Guarantee of BRE

Recall that our stopping rule is based on the Monte Carlo estimations. The error of BRE involves the Bayesian part

Algorithm 1 Bayes Re-sample Explore

Input: Prior $p_0(\boldsymbol{\theta})$, c
$t = 0, Terminal = False$
while $Terminal$ is $False$ **do**
  $t \leftarrow t+1, num = 1, OptSet = \{\}$
  while $|OptSet| < 2$ and $num < c$ **do**
    Sample $\boldsymbol{\theta}_t$ from $p_{t-1}(\boldsymbol{\theta})$
    $Z_t = \text{argmax}_{Z \in \mathcal{Z}} g(Z, \boldsymbol{\theta}_t)$
    $num+ = 1$
    $OptSet = OptSet \cup Z_t$
  end while
  if $|OptSet| == 2$ **then**
    Calculate κ_t and solve I_t with selection rules
    Choose action set I_t and get observation r_t
    Use (I_t, r_t) to update the posterior $p_t(\boldsymbol{\theta})$
  else
    $Teminal = True$
  end if
end while
Recommend Z_t

from the posterior distribution and the frequentist part from the Monte Carlo sampling process. Therefore we consider an expected posterior PGS to characterize both errors together and give a guarantee of stopping for BRE.

Theorem 1. *For $\boldsymbol{\theta}^*$ sampled from p_0, with c chosen as Eq.* (13)*, BRE stops at τ. Then with a probability $1-\delta_1$*

$$\mathbb{E}_{\boldsymbol{\theta}^*}\left[\mathbb{I}[\mathcal{E}(Z_\tau; \boldsymbol{\theta}^*)]\right] > 1 - \delta_2. \tag{14}$$

This theorem is exactly our designed stopping rule.

Sample Complexity for BRE-R

A non-asymptotic analysis of general problems is hard. Frequentist's upper-bound analysis is not suitable for sample-based methods. Analysis involving sample randomness is needed. With some techniques from Xu, Honda, and Sugiyama (2018) and Agrawal and Goyal (2013), below we propose our novel theoretical results for BRE-R.

Recall that Assumptions 1 and 2 are assumed to be satisfied. Denote that i^* to be the index of Z^* and $\ell = \text{argmin}_{i \in [m_Z], i \neq i^*} \ell^{i^* i}(\zeta^{i^* i}(\delta/3))$. Then we define

$$F_k^{ij} = \frac{1}{2} + \tilde{w}_k^{ij} \gamma^{ij\,-1} \left(\frac{\max(|\kappa^{ij}(\boldsymbol{\theta}^*)| + \varepsilon, \kappa^{i^* j}(\boldsymbol{\theta}^*), \kappa^{i^* i}(\boldsymbol{\theta}^*))^2}{\left(\zeta^{i^* i}(\delta/3) + \sqrt{2\ln(12 m_Z^2/\delta)}\right)^2} \right).$$

Theorem 2. *If the problem satisfies Assumptions 1 and 2, then with a probability no less than $1-\delta$, the stopping time τ for BRE-R is upper bounded by*

$$\max\left(\frac{4 \sum_{k \in [m_I]} \max_{i,j \in [m_Z]: i \neq j} F_k^{ij}}{\ell(1-\delta)}, \frac{16K \ln(2m_I/\delta)}{\ell^2(1-\delta)}, \frac{8\ln(2/\delta)}{(1-\delta)^2} \right). \tag{15}$$

Here the term $1/\ell/(1-\delta)$ characterize the extra samples caused by the Monte Carlo stopping process.

The key of BRE-R is to choose action sets such that their ratios are close to $\tilde{\boldsymbol{w}}^{ij}$. In Thm. 2, value F_k^{ij} characterizes the needed number of action set I_k to distinguish the (i,j) pair according to $\tilde{\boldsymbol{w}}^{ij}$. It can be seen that the sample complexity is in negative correlation with $\kappa^{i^* i}(\boldsymbol{\theta}^*)$ and $\kappa^{i^* j}(\boldsymbol{\theta}^*)$.

The proof divides time steps into two parts. The first part tries to distinguish all the sub-optimal target sets. The left part characterizes the extra steps caused by randomness, leading to the $1/(\ell(1-\delta))$ ratio. The complete proof is given in the Appendix . In LB, this result has a comparable order on LB with previous state-of-the-art work LinGapE, as discussed in Appendix .

Approximated BRE

One main difficulty of Bayesian methods is the complex calculation. To solve this, we use variational inference for posterior calculation and influence functions for action set selection. Hence we give an approximating approach, named BRE-A. Complete derivation is given in Appendix

Instead of exact calculation, we use variational inference to approximate $p(\boldsymbol{\theta}|\mathcal{D}_t)$. Let $q_\phi(\boldsymbol{\theta})$ be an distribution of $\boldsymbol{\theta}$ parameterized by ϕ. Recall $d_s = (I_s, r_s)$. We solve $\hat{\phi}_t$ with

$$\begin{aligned}\hat{\phi}_t &= \operatorname*{argmin}_\phi \mathrm{KL}[q_\phi(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathcal{D}_t)] \\ &= \operatorname*{argmin}_\phi (\mathrm{KL}[q_\phi(\boldsymbol{\theta}) \| p_0(\boldsymbol{\theta})] + \sum_{s=1}^{t} L(d_s, \phi)),\end{aligned}$$

where $L(d_s, \phi) := -\mathbb{E}_{q_\phi(\boldsymbol{\theta})}[\log p(r_s|\boldsymbol{\theta}, I_s)]$. If $p(\boldsymbol{\theta}|\mathcal{D}_t)$ lies in the function class of $q_\phi(\boldsymbol{\theta})$, $q_{\hat{\phi}_t}(\boldsymbol{\theta})$ equals to $p(\boldsymbol{\theta}|\mathcal{D}_t)$. Otherwise it is the best approximation of $p(\boldsymbol{\theta}|\mathcal{D}_t)$ from the distribution family. For convenience, we take $\hat{\phi}_t$ as the parameters of the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}_t)$.

Now assume that $d = (I, r_I)$ is the new observation for time step t. Then we can solve a ϕ with d weighted by ϵ

$$\hat{\phi}^\epsilon_{t,I,r_I} = \operatorname*{argmin}_\phi (\mathrm{KL}[q_\phi(\boldsymbol{\theta}) \| p_0(\boldsymbol{\theta})] + \sum_{s=1}^{t-1} L(d_s, \phi) + \epsilon L(d, \phi)).$$

With the result of Koh and Liang (2017), we define the "influence" of choosing action set I on the posterior of $\boldsymbol{\theta}$:

$$\begin{aligned}\mathcal{J}_{\boldsymbol{\theta}_t, I} &= \mathbb{E}_{p(r_I|\mathcal{D}_{t-1}, I)} \left[\frac{\mathrm{d}\hat{\phi}^\epsilon_{t,I,r_I}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \right] \\ &= \left(-H|_{\phi=\hat{\phi}_t}\right)^{-1} \mathbb{E}_{p(r_I|\mathcal{D}_{t-1}, I)} \left[\nabla_\phi L((I, r), \phi)\big|_{\phi=\hat{\phi}_t} \right],\end{aligned}$$

where $H = \nabla^2_\phi(\mathrm{KL}[q_\phi(\boldsymbol{\theta}) \| p_0(\boldsymbol{\theta})] + \sum_{s=1}^t L(d_s, \phi))$ is the Hessian. Then the influence for the variance of $\kappa_t(\boldsymbol{\theta})$ is defined as:

$$\mathcal{J}_{Var(\kappa_t), I} = \mathbb{E}_{p(r_I|\mathcal{D}_{t-1}, I)} \left[\frac{\mathrm{dVar}(\kappa_t(\boldsymbol{\theta}))}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \right] \tag{16}$$

$$= \hat{H} \mathbb{E}_{p(r_I|\mathcal{D}_{t-1}, I)} \left[\nabla_\phi L((I, r_I), \phi)\big|_{\phi=\hat{\phi}_t} \right], \tag{17}$$

where $\hat{H} = \left(\nabla_\phi \mathrm{Var}(\kappa_t(\boldsymbol{\theta}))\big|_{\phi=\hat{\phi}_t}\right)^\top \left(-H|_{\phi=\hat{\phi}_t}\right)^{-1}$.

Now we propose the selection rule for BRE-A:

$$I_t = \operatorname*{argmin}_{I \in \mathcal{I}} \mathcal{J}_{Var(\kappa_t), I}. \tag{18}$$

Linear Bandit Example

In this section, we apply BRE to Linear Bandits (LB). In LB, $\mathcal{I} = \mathcal{I}_{arm}$, i.e., we pull one arm at each time step. For convenience, we use I_t to represent selected arms instead of arm sets. In LB, denote $n = m_x = m_\theta$. For each arm i, we have $f(\{i\}, \boldsymbol{\theta}^*) = g(\{i\}, \boldsymbol{\theta}^*) = \boldsymbol{x}^{i\top}\boldsymbol{\theta}^*$. Assume that rewards are sampled from Gaussian distributions. That is, a rewards for choosing arm i is sampled from $\mathcal{N}(\boldsymbol{x}^{i\top}\boldsymbol{\theta}^*, v^2)$ for some fixed v.

We consider the Explore-K arm problem in LB. That is, our goal is to find K arms with maximum rewards. For two target sets Z_i and Z_j, we define their difference vector as

$$\boldsymbol{\xi}^{ij} = \sum_{z\in Z_i} \boldsymbol{x}^z - \sum_{z'\in Z_j} \boldsymbol{x}^{z'}. \tag{19}$$

Recall that the indices for Z_t and Z'_t are denoted as i_t and j_t. For convenience, we define $\boldsymbol{\xi}_t = \boldsymbol{\xi}^{i_t j_t}$ if $\hat{\kappa}_t^{i_t j_t} \geq 0$; $\boldsymbol{\xi}_t = \boldsymbol{\xi}^{j_t i_t}$ otherwise. Thus we have that $\kappa_t(\boldsymbol{\theta}) = \boldsymbol{\xi}_t^\top \boldsymbol{\theta}$.

Following Agrawal and Goyal (2013), we consider a prior $\mathcal{N}(0, \lambda A_0)$ for $\boldsymbol{\theta}$, where A_0 is an $n \times n$ identity matrix. With observation $\mathcal{D}_{t-1} = \{I_1, r_1, ..., I_{t-1}, r_{t-1}\}$ up to t, the posterior distribution $p_t(\boldsymbol{\theta})$ is $\mathcal{N}(\tilde{\boldsymbol{\theta}}_t, v^2 A_t^{-1})$, where $A_t = \lambda A_0 + \sum_{t'=1}^{t-1} x^{I_{t'}} x^{I_{t'}\top}$ and $\tilde{\boldsymbol{\theta}}_t = A_t^{-1} \sum_{t'=1}^{t-1} x^{I_{t'}} r_{t'}$. The marginal distribution over the direction of κ_t is thus $\mathcal{N}(\boldsymbol{\xi}_t^\top \tilde{\boldsymbol{\theta}}_t, v^2 \boldsymbol{\xi}_t^\top A_t^{-1} \boldsymbol{\xi}_t)$. Below we give three BRE action set selection rules for LB.

BRE-R: LB with Gaussian rewards enjoys quite nice properties. We choose the variance proxy function as

$$h^{ij}(t, w) = v^2 \boldsymbol{\xi}^{ij\top} \left(\lambda A_0 + t \sum_{k\in[N]} w_k * x^{I_i} x^{I_i\top}\right)^{-1} \boldsymbol{\xi}^{ij}. \tag{20}$$

To construct $\tilde{\boldsymbol{w}}^{ij}$ for each $\boldsymbol{\xi}^{ij}$. First we solve a^* with

$$a^* = \operatorname*{argmin}_{a\in\mathbb{R}^N} \sum_{k=1}^N |a_k|, \ s.t. \ \boldsymbol{\xi}^{ij} = \sum_{k=1}^N a_k \boldsymbol{x}^k. \tag{21}$$

With $\gamma^{ij}(t) = \frac{v^2}{t} \sum_{k=1}^N |a_k^*|$ and $\tilde{w}_k^{ij} = |a_k^*| / \sum_{k=1}^N |a_k^*|$, we can simply apply Eq. (11) to get BRE-R.

It can be proved that the above constructions satisfy Assumptions 1& 2. We defer the assumption check, detailed derivations, and complexity analysis to Appendix .

BRE-G: For LB, the variance of the posterior $p_t(\boldsymbol{\theta})$ is not influenced by the observed reward r_I. We denote $||\boldsymbol{y}||_A := \sqrt{\boldsymbol{y}^\top A \boldsymbol{y}}$. Then at time step t, BRE-G selection rule is

$$I_t = \operatorname*{argmin}_{i\in[N]} ||\boldsymbol{\xi}_t||_{(A_t + \boldsymbol{x}^i \boldsymbol{x}^{i\top})^{-1}}, \tag{22}$$

BRE-A: With the "influence" function calculation, the BRE-A arm selection rule for time step t turns to be:

$$I_t = \operatorname*{argmax}_{i\in[N]} (\boldsymbol{\xi}_t^\top A_t^{-1} \boldsymbol{x}^i)^2. \tag{23}$$

A detailed derivation for BRE-A is given in Appendix .

An interesting observation here is that Eq. (23) & (22) can be connected via the Sherman-Morrison formula:

$$\boldsymbol{\xi}^T (A + \boldsymbol{x}\boldsymbol{x}^\top)^{-1} \boldsymbol{\xi} = \boldsymbol{\xi}^T A^{-1} \boldsymbol{\xi} - (\boldsymbol{\xi}^\top A^{-1} \boldsymbol{x})^2 / (1 + \boldsymbol{x}^\top A^{-1} \boldsymbol{x}). \tag{24}$$

Thus we can consider BRE-A as an approximation method for BRE-G. Notice that we only need to calculate the inverse matrix A_t^{-1} for once using rule (23), while rule (22) needs to calculate the inverse matrix for each arm. This is extremely time efficient for problems with large N.

Discussion for LB

Three BRE methods can all be applied to LB directly, as shown above. Notice that the final selection rules for BRE-G in LB have similar forms to that in LinGapE, since LinGapE also employs variance to construct upper confidence bounds. Although BREs and LinGapE have similar selection rules, our sample-based optimal-arm choosing and stopping rules are shown to be more effective empirically.

For current BRE-R, the sample complexity bound has an order of $O(\sum_k \max_{i\neq j} \tilde{w}_k^{ij} / (|\kappa^{ij}(\boldsymbol{\theta}^*)| + \varepsilon)^2)$. This bound is relatively loose comparing to the bound of LinGapE. A small modification on posterior calculation could improve the bound to the same order of LinGapE. That is, a variant of BRE-R has comparable theoretical results with state-of-the-art frequentist method LinGapE on LB. However, the empirical performance of this variant is worse than current BRE-R, since it uses a larger variance. We conjecture that our current analysis for BRE-R in LB can be further improved. The complete discussion for BRE-R and its variant in LB is given in Appendix .

Experiments

We evaluate the performance of BRE by comparing with various baselines. Following (Xu, Honda, and Sugiyama 2018), we design experiments on synthetic data and a real-data simulated Explore-K problem. We test BRE on LB and some other general CB problems. Appendix provides detailed implementations for algorithms and further results.

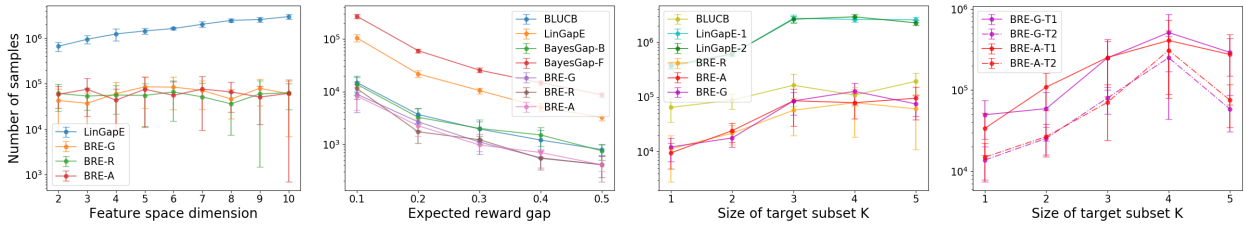
Baseline Methods

Here we introduce the comparison methods in detail.

BRE: We test BRE-G, BRE-R, and BRE-A.

Frequentist methods: For best one arm identification problems in LB, we use the state-of-the-art LinGapE as the frequentist baseline. We extend LinGapE to the Explore-K problem in two ways. For the Explore-K problem, we first use empirical mean to find K optimal arms subset Z_t. Then we aim to find two arms in Z_t and $[N]/Z_t$ respectively that they have maximum discrepancy and then choose one arm to distinguish them. Two standards are used to measure the discrepancy: one uses the ratio of bonus and the estimated reward, and the other uses the summation of them. We call them LinGapE-1 and LinGapE-2 respectively. Other methods such as $\mathcal{XY}$-static are not tested since they are sample inefficient, as shown in Xu, Honda, and Sugiyama (2018).

Other Bayesian methods: We extend LUCB (Kalyanakrishnan et al. 2012) to the Bayesian framework by sampling $\boldsymbol{\theta}_t$ from p_{t-1} and then pulling the Kth and $K+1$th optimal arms. We call it BLUCB and use the BRE stopping rule for it. We further extend BayesGap into the fixed-confidence setting, with two stopping rules: the BRE stopping rule



(a) Synthetic Data 1 (b) Synthetic Data 2 (c) Yahoo Linear Data (d) Yahoo General Data

Figure 1: The number of samples for best one arm identification problems

Arms	1	2	3	4	5	6
BRE-G	138	2.8E4	6	13	4	1
BRE-R	128	2.5E4	5	7	10	10
BRE-A	195	3.9E4	5	10	14	5
LinGapE	7.1E3	1.4E6	60	70	47	3
BLUCB	5.0E7	1	3	7	1	5.0E7
BayesGap-B	5.0E7	32	43	40	48	5.0E7
BayesGap-F	5.0E7	421	43	63	39	5.0E7

Table 1: Arms counts for $n = 5$

and the frequentist stopping rule of LinGapE. We call them BayesGap-B and BayesGap-F respectively.

We repeat each experiment for 10 times and plot means and standard deviations. We set a maximum budget: the interaction process will stop after pulling 10^8 times. For all experiments, we set $\delta = 0.05$, $\varepsilon = 0.0$ and $\lambda = 1$. The labels of the y axes are all the "number of samples".

Experiments on Best Arm Identification

We follow Xu, Honda, and Sugiyama (2018) to design 2 synthetic experiments. We set the reward distribution to be $\mathcal{N}({\boldsymbol{x}^i}^\top \boldsymbol{\theta}^*, 1)$ for each arm i.

Synthetic data 1: We choose $N = n + 1$ arms, with n varying from 2 to 10. The first n arms are the canonical base vectors of $\mathbb{R}^n$ and $\boldsymbol{x}^{n+1} = (\cos(0.01), \sin(0.01), 0, ..., 0)$. We set $\boldsymbol{\theta}^* = (2, 0, ..., 0)$. Thus, an algorithm should carefully distinguish $\boldsymbol{x}^1$ and $\boldsymbol{x}^{n+1}$ and should pull $\boldsymbol{x}^2$ frequently. Fig.1(a) shows the results. Note that we ignore BLUCB, BayesGap-B and BayesGap-F as they use samples more than 10^8. Instead, we give an example of arm counts for $n = 5$, as shown in Table 1. We can see that our BRE outperforms other methods. From the table, BLUCB and BayesGaps repeatedly pull arms 1 and 6, since they fail to use the arm correlations to choose arms. Fig.1(a) shows BRE methods are much faster than LinGapE. Further, LinGapE uses more samples as n gets larger since it requires all bounds to be tight, while BRE methods with an n-independent Monte Carlo stopping rule remain relatively stable.

Synthetic data 2: We choose canonical bases as the n arms and $\boldsymbol{\theta}^* = (\rho, 0, ..., 0)$. We choose $\rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. In this case, the correlation of arms is not very important and algorithms need to explore all arms independently. Fig. 1(b) shows the results. We can see that the problem becomes easier as ρ increases. All methods can efficiently find the optimal arm while BREs still perform the best. Comparison between BayesGap-F and BayesGap-B shows that the frequentist stopping rule is indeed empirically conservative.

Experiments on Subset Selection

We test the subset selection problems on *Yahoo! Webscope Dataset R6A* (Chu et al. 2009)[2]. We set $n = 4$, $N = 15$, $K \in \{1, 2, 3, 4, 5\}$. We construct data following LinGapE where the expected reward for arm i is ${\boldsymbol{x}^i}^\top \boldsymbol{\theta}^*$. The reward is 1 with a probability $({\boldsymbol{x}^i}^\top \boldsymbol{\theta}^* + 1)/2$ and -1 otherwise. We use Gaussian as the distribution class for variational inference and choose Gaussian prior for $\boldsymbol{\theta}$. We train BRE-G, BRE-R, BRE-A, LinGapE-1, LinGapE-2 and BLUCB.

We further vary the target function with: (1) $g(Z, \boldsymbol{\theta}) = \sum_{i \in Z} \exp({\boldsymbol{x}^i}^\top \boldsymbol{\theta})$ and, (2) $g(Z, \boldsymbol{\theta}) = \sum_{i \in Z} {\tilde{\boldsymbol{x}^i}}^\top \boldsymbol{\theta}^*$, where $\tilde{\boldsymbol{x}^i}$ is a noised vector near $\boldsymbol{x}^i$. We test them with BRE-G and BRE-A and name the two targets with T1 and T2.

The results are plotted in Fig. 1(c) and 1(d). BRE methods still can perform the best. For varied target functions, BRE-G and BRE-A can still efficiently explore the target sets. Empirically, we observe that the sample number is mostly influenced by the reward gap between the optimal and the near-optimal target sets, rather than the size of K. This is why the $K = 4$ problem seems to be harder than $K = 5$. Further details and discussion are given in Appendix .

Conclusion

We propose a Bayesian approach, BRE, for subset selection in CB problems. BRE uses samples from posterior distributions to find two near-optimal sets and chooses action sets to distinguish the two sets. Further, we use a Monte Carlo stopping rule to efficiently find the posterior stopping condition. A theoretical analysis for specific cases is given and we also propose an approximation method for calculation convenience. We test our method with previous baselines on synthetic and real data based simulations. Our method outperforms frequentist methods and other Bayesian methods.

Acknowledgements

This work was supported by NSFC Projects (Nos. 62061136001, 61620106010, U19B2034, U1811461), Beijing NSF Project (No. L172037), Beijing Academy of Artificial Intelligence (BAAI), and the Xplorer Prize.

[2]https://webscope.sandbox.yahoo.com/

References

Agrawal, S.; and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 127–135.

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

Branke, J.; Chick, S. E.; and Schmidt, C. 2007. Selecting a selection procedure. *Management Science* 53(12): 1916–1932.

C., H.; Abramowitz, M.; and Stegun, I. A. 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. *Journal of the American Statal Association* 59(308): 1324.

Chapelle, O.; and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, 2249–2257.

Chen, C.-H.; Lin, J.; Yücesan, E.; and Chick, S. E. 2000. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10(3): 251–270.

Chu, W.; Park, S.-T.; Beaupre, T.; Motgi, N.; Phadke, A.; Chakraborty, S.; and Zachariah, J. 2009. A case study of behavior-driven conjoint analysis on Yahoo! Front Page Today module. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1097–1104.

Eckman, D. J.; and Henderson, S. G. 2018. Guarantees on the probability of good selection. In *2018 Winter Simulation Conference (WSC)*, 351–365. IEEE.

Henderson, S. G.; and Nelson, B. L. 2006. *Handbooks in operations research and management science: simulation*, volume 13. Elsevier.

Hoffman, M.; Shahriari, B.; and Freitas, N. 2014. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, 365–374.

Kalyanakrishnan, S.; and Stone, P. 2010. Efficient Selection of Multiple Bandit Arms: Theory and Practice. In *International Conference on Machine Learning*, volume 10, 511–518.

Kalyanakrishnan, S.; Tewari, A.; Auer, P.; and Stone, P. 2012. PAC Subset Selection in Stochastic Multi-armed Bandits. In *International Conference on Machine Learning*, volume 12, 655–662.

Kaufmann, E.; Cappé, O.; and Garivier, A. 2016. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research* 17(1): 1–42.

Kim, S.-H.; and Nelson, B. L. 2006. Selecting the best system. *Handbooks in operations research and management science* 13: 501–534.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*, 1885–1894.

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

Russo, D. 2016. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, 1417–1418.

Russo, D. J.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, Z.; et al. 2018. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning* 11(1): 1–96.

Slepian, D. 1962. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal* 41(2): 463–501.

Soare, M.; Lazaric, A.; and Munos, R. 2014. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, 828–836.

Thompson, W. R. 1933. On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25(3-4): 285–294.

Xu, L.; Honda, J.; and Sugiyama, M. 2018. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, 843–851.