Operations Research Letters 49 (2021) 365–371

Contents lists available at ScienceDirect

Operations Research Letters

www.elsevier.com/locate/orl

Best arm identification in generalized linear bandits

Abbas Kazerouni [a], Lawrence M. Wein [b,*]

[a] Facebook, Menlo Park, CA 94025, United States of America
[b] Graduate School of Business, Stanford University, Stanford, CA 94305, United States of America

ARTICLE INFO

Article history:
Received 4 January 2021
Received in revised form 25 March 2021
Accepted 31 March 2021
Available online 6 April 2021

Keywords:
Best arm identification
Generalized linear bandits
Sequential clinical trial

ABSTRACT

We consider the best-arm identification problem in generalized linear bandits: each arm has a vector of covariates, there is an unknown vector of parameters that is common across the arms, and a generalized linear model captures the dependence of rewards on the covariate and parameter vectors. The goal is to identify a near-optimal arm with high probability while minimizing the number of arm pulls (i.e., the sampling budget). We propose the first algorithm for this problem and provide theoretical guarantees on its accuracy and sampling efficiency.

1. Introduction

The multi-armed bandit problem is a prototypical model for optimizing the tradeoff between exploration and exploitation. We consider a pure-exploration version of the bandit problem known as the best-arm identification problem, where the goal is to minimize the number of arm pulls required to select an arm that is - with sufficiently high probability – sufficiently close to the best arm. We assume that each arm has an observable vector of covariates or features, and there is an unknown vector of parameters (of the same dimension as the vector of features) that is common across arms. Hence, with every pull of an arm, the decision maker refines the estimate of the unknown parameter vector and learns simultaneously about all arms. Whereas in a linear bandit the mean reward of an arm is the linear predictor (i.e., the inner product of the parameter vector and the feature vector), in our generalized linear model the mean reward is related to the linear predictor via a link function, which allows for mean rewards that are nonlinear in the linear predictor, as well as binary or integer rewards (via, e.g., logistic or Poisson regression). Our motivation for studying this version of the bandit problem comes from drug design, where arms correspond to drugs, the covariate vectors describe biological properties of the drugs, and the inverse link function in the generalized linear model captures the nonlinear and perhaps binary relationship between the covariates and the experimental outcomes.

There is a vast literature on best-arm identification in the multi-arm bandit setting with independent arms, where pulling one arm does not reveal any information about the reward of other arms ([2] and references therein). Rather than assume independent arms, our formulation considers parametric arms, where each arm has a covariate vector and there is an unknown parameter vector that is common across arms. There has been considerable work on the linear parametric bandit (i.e., the mean reward of an arm is the inner product of its covariate vector and the parameter vector) under the minimum-regret objective ([8] and references therein) as well as alternative probabilistic models of arm dependence (e.g., [9] and references therein). Relevant for our purposes, the generalized linear parametric bandit, which uses an inverse link function to relate the linear predictor and the mean reward, has been studied under regret minimization [3,6].

However, relatively little work exists on best-arm identification in parametric bandits. Major progress has been made in [10] by proposing the first adaptive algorithm for best-arm identification in linear bandits. These authors design a gap-based exploration algorithm by employing the standard confidence sets constructed in the literature for linear bandits under regret minimization. We adapt the gap-based exploration algorithm of [10] from the linear setting to the generalized linear case, which requires us to derive confidence sets for reward gaps between different pairs of arms. In the regret-minimization setting, the typical approach to the linear bandit is to develop a confidence set for the unknown parameter vector that governs the rewards of all arms, whereas in the best-arm setting, a confidence set on the reward gaps is needed. In the best-arm identification for the linear bandit [10], the authors were able to convert the confidence set for the parameter vector into efficient confidence sets for the reward gaps. However, this ap-

* Corresponding author at: 655 Knight Way, Stanford, CA 94305-5015, United States of America.
E-mail addresses: kazerouni@fb.com (A. Kazerouni), lwein@stanford.edu (L.M. Wein).

https://doi.org/10.1016/j.orl.2021.03.011
0167-6377/

proach breaks down in the generalized linear bandit; i.e., naively converting the confidence set for the parameter vector in [3] into confidence sets for reward gaps between arms leads to extremely loose confidence sets, which strongly degrades the performance of the gap-based exploration algorithm. Rather than use this indirect method, we build gap confidence sets directly from the data.

The remainder of this paper is organized as follows. In Section 2, we formulate the best-arm identification problem for generalized linear bandits. We describe our algorithm in Section 3 and establish theoretical guarantees in Section 4. We provide concluding remarks in Section 5.

2. Problem formulation

Consider a decision maker who is seeking to find the best (i.e., highest mean reward) among a set of K available arms. We let $[K] = \{1, 2, \cdots, K\}$ denote the set of possible arms. There is a feature vector $x^a \in \mathbb{R}^d$ associated with arm a, for $a \in [K]$. These feature vectors are known to the decision maker and each summarizes the available information about the corresponding arm.

The decision maker chooses an arm a_t to play in each round t. We employ a generalized linear model [7] and assume that a stochastic reward r_t is observed whose distribution lies within the exponential family; i.e., having a probability density function of the form

$$p(r_t) = \exp(r_t \theta^\intercal x^{a_t} - b(\theta^\intercal x^{a_t}) + c(r_t)), \tag{1}$$

where $\theta \in \mathbb{R}^d$ is an unknown parameter that governs the reward of all arms, $b(\cdot)$ and $c(\cdot)$ are real functions such that $\mu(z) = \frac{d}{dz} b(z)$ for all $z \in \mathbb{R}$, where $\mu : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function known as the inverse link function. In our model, the rewards of every arm are governed by the same inverse link function. Hence, according to the properties of the exponential family, we have

$$\mathbb{E}[r_t | a_t] = \mu_{a_t} = \mu(\theta^\intercal x^{a_t}). \tag{2}$$

Given a_t and θ, the r_t's are independent random variables. Different choices for the function $\mu(z)$ in (2) result in modeling different reward structures inside the exponential family. For example, choosing $\mu(z) = e^z$ and $\mu(z) = 1/(1+e^{-z})$ correspond to a Poisson regression model and a logistic regression model, respectively.

Let $a^* = \arg\max_{a \in [K]} \mu_a$ be the optimal arm; i.e., the arm with the highest expected reward. Throughout this study, we break ties randomly whenever the arg max operator returns more than one value. By exploring different arms, the decision maker is trying to find the optimal arm as soon as possible based on the noisy observations. Let τ be a stopping time that dictates whether enough evidence has been gathered to declare the optimal arm. The declared optimal arm is denoted by $\hat{a}_\tau$. An exploration strategy can be represented by $\mathcal{S} = (\mathcal{A}, \tau)$, where, at any time t, $\mathcal{A}$ is a function mapping from the previous observations $\{(a_i, r_i)\}_{i=1}^{t-1}$ to the arm a_t to be played next, and τ determines whether enough information has been gathered to declare the optimal arm, $\hat{a}_\tau$. Because finding the exact optimal arm may require a prohibitively large amount of exploration, the performance of an exploration strategy is evaluated via the following relaxed criterion.

Definition 1. Given $\epsilon > 0$ and $\delta \in (0, 1)$, an exploration strategy $\mathcal{S} = (\mathcal{A}, \tau)$ is said to be $(\epsilon, \delta)-$optimal if

$$\mathbb{P}\left[\mu(\theta^\intercal x^{a^*}) - \mu(\theta^\intercal x^{\hat{a}_\tau}) \geq \epsilon\right] \leq \delta. \tag{3}$$

In this definition, ϵ denotes an acceptable region around the optimal arm and $1 - \delta$ represents the confidence in identifying an arm within this region. This criterion relaxes the notion of optimality by allowing the exploration strategy to return a sufficiently good – but not necessarily optimal – arm. With this definition in place, the decision maker's goal is to design an $(\epsilon, \delta)-$optimal exploration strategy with the smallest possible stopping time.

Before proceeding to the algorithm, we introduce additional notation and state a set of regularity assumptions. We let $\mathcal{X} = \{x^a\}_{a \in [K]}$ denote the set of feature vectors and assume that feature vectors, the unknown reward parameter and the rewards are bounded; i.e., there exist $S, L, R > 0$ such that $\|\theta\|_2 \leq S$, $\|x\|_2 \leq L$ $\forall x \in \mathcal{X}$, and $r_t \leq R$ almost surely for all t. We also assume that μ is continuously differentiable, Lipschitz continuous with constant k_μ and satisfies $c_\mu = \inf_{\theta:\|\theta\| \leq S,\ x \in \mathcal{X}} \dot{\mu}(\theta^\intercal x) > 0$. For example, in the case of logistic regression, $R = 1, k_\mu = 1/4$ and c_μ depends on $\sup_{\theta:\|\theta\| \leq S,\ x \in \mathcal{X}} |\theta^\intercal x|$. While our algorithm and analysis will hold with any upper bound on k_μ and any lower bound on c_μ, we assume that k_μ and c_μ are known and given. We define the gap between any two arms $i, j \in [K]$ to be $\Delta(i, j) = \mu_i - \mu_j = \mu(\theta^\top x^i) - \mu(\theta^\top x^j)$ and define the optimal gap associated to an arm $i \in [K]$ as

$$\Delta_i = \begin{cases} \mu(\theta^\top x^{a^*}) - \mu(\theta^\top x^i) & \text{if } i \neq a^* \\ \mu(\theta^\top x^i) - \max_{j \neq i} \mu(\theta^\top x^j) & \text{if } i = a^*. \end{cases} \tag{4}$$

Finally, for any positive semi-definite matrix A, we let $\|x\|_A = \sqrt{x^\top A x}$.

3. The proposed algorithm

In this section, we propose an exploration strategy for the problem formulated in Section 2. Following [10], our algorithm consists of the following steps:

1. Build confidence sets for the pairwise gaps between arms,
2. Identify the potential best arm and an alternative arm that has the most ambiguous gap with the best arm,
3. Play an arm to reduce this ambiguity.

These steps are repeated sequentially until the ambiguity in step 2 drops below a certain threshold.

The confidence sets are derived in Subsection 3.1 and the algorithm is presented in Subsection 3.2.

3.1. Confidence sets for gaps

To build the confidence sets for reward gaps, we follow ideas in [3] but develop confidence sets directly for gaps instead of arm rewards.

Let $x_l = x^{a_l}$ be shorthand notation for the feature vector associated with the arm played in period l and let $H_{t-1} = \{(x_1, r_1), (x_2, r_2), \cdots, (x_{t-1}, r_{t-1})\}$ be the history of actions played and random rewards observed prior to period t. For any $t > 0$, let $M_t = \sum_{l=1}^{t-1} x_l x_l^\top$ and assume it is nonsingular for any $t > E$ for some fixed value $E > d$. We let $\lambda_0 > 0$ be the minimum eigenvalue of M_{E+1} and define $\kappa = \sqrt{3 + 2\log(1 + 2L^2/\lambda_0)}$. Given the observations by the start of period t, by (1) the Maximum Likelihood (ML) estimate of the reward parameter, θ_t, solves the equation

$$\sum_{l=1}^{t-1} (r_l - \mu(\theta_t^\intercal x_l)) x_l = 0. \tag{5}$$

Based on the estimated reward parameter, we can take

$$\Delta_t(i, j) = \mu(\theta_t^\top x^i) - \mu(\theta_t^\top x^j) \tag{6}$$

as an estimate for the gap between arms $i, j \in [K]$, which is a function of the observations made prior to period t.

Given $\delta \in (0, 1)$, let

$$C_t = \alpha \sqrt{2d \log t \log\Big(\frac{\pi^2 d t^2}{6\delta}\Big)}, \tag{7}$$

where α is a tunable parameter and C_t is a time-varying quantity that scales the width of the confidence sets for all pairs of arms. We set

$$\alpha = \frac{2\kappa R}{c_\mu} \tag{8}$$

in the theoretical analysis in Section 4, and set α so as to achieve robust performance across a variety of scenarios in the computational study in [5]. We consider the following confidence set for the gap between any two arms $i, j \in [K]$ based on the observations made prior to period t:

$$\mathcal{C}_t(i,j) = \left\{ \Delta \in \mathbb{R} : |\Delta - \Delta_t(i,j)| \le C_t \max_{c,c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}} \right\}. \tag{9}$$

The confidence set defined in (9) is centered around the estimated gap between the two arms in (6) and, as we will show in Section 4, contains the true gap with high probability. To further simplify the representation, we let

$$\beta_t(i,j) = C_t \max_{c,c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}} \tag{10}$$

represent the width of the confidence set in (9). In (10), the matrix M_t^{-1} is the inverse of M_t. This matrix norm reflects the fact that the shape of the confidence sets depends heavily on the previously observed data; i.e., we have higher confidence in directions where more data have been observed. The right side of (10) is derived by linearizing the inverse link function, and the maximization takes a conservative approach by using worst-case slopes c and c'.

Although we follow the basic approach in [3] in deriving these confidence sets, it is worth mentioning that they cannot be deduced from the results presented in [3]. More precisely, an immediate use of the results in that paper gives rise to a confidence set that is similar to (9) except that $\max_{c,c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}}$ is replaced by the factor $\left(\|x^i\|_{M_t^{-1}} + \|x^j\|_{M_t^{-1}}\right)$, which would be much looser because the feature vectors x^i and x^j could be close to each other. On the other hand, an even tighter bound would depend on $\|x_i - x_j\|_{M_t^{-1}}$, but – as with earlier work – we have been unable to derive such a bound.

While linearizing the inverse link function introduces some looseness into the analysis, empirical methods such as the likelihood method, the lack-of-fit method or the bootstrap method do not provide a theoretical guarantee and require the sequential data that are generated to be independent and identically distributed. Moreover, the conservative approach of taking worst-case slopes c and c' in (10) is required to obtain a theoretical bound, but introduces additional looseness into the analysis. We alleviate this looseness by downscaling the theoretical length of the confidence set in the numerical results in [5].

3.2. The algorithm

With the confidence sets established, we are now ready to describe our proposed algorithm. Details and successive steps of the proposed algorithm are presented in Algorithms 1-3. Following the gap-based exploration scheme, the proposed algorithm consists of two major components that are described below.

Algorithm 1: GLGapE.

Input: $\mathcal{X}, \epsilon, \delta, E, \alpha, S, L, R, c_\mu, k_\mu$
Output: approximately best arm $\hat{a}_\tau$
Initial Exploratory Phase: Play E random arms and gather observations in H_E, and initialize $T_a(t)$'s for the played arms
for $t > E$ **do**
  `//select a gap to examine`
  $i_t, j_t, B(t), y_t \leftarrow$ select-gap$(\mathcal{X}, H_{t-1}, c_\mu, k_\mu)$
  if $B(t) \le \epsilon$ **then**
    return i_t as the best arm
  end if
  `//select an action`
  $a_t \leftarrow$ select-arm$(\mathcal{X}, H_{t-1}, T(t), y_t)$
  Play a_t
  Observe r_t
  $H_t = H_{t-1} \cup \{(x_t, r_t)\}$
  $T_a(t) = T_a(t-1) + 1$
end for

Algorithm 2: Select-gap.

Input: $\mathcal{X}, H, c_\mu, k_\mu$
Output: a gap to be examined
Find the ML estimate θ_t
$i_t \leftarrow \arg\max_{i \in [K]} \mu(\theta_t^\top x^i)$
$j_t \leftarrow \arg\max_{j \in [K], j \neq i_t} \Delta_t(j, i_t) + \beta_t(i_t, j)$
Find $c_1, c_2 = \arg\max_{c,c' \in [c_\mu, k_\mu]} \|cx^{i_t} - c'x^{j_t}\|_{M_t^{-1}}$
Let $y_t = c_1 x^{i_t} - c_2 x^{j_t}$
$B(t) \leftarrow \Delta_t(j_t, i_t) + \beta_t(i_t, j_t)$
return $i_t, j_t, B(t), y_t$

Algorithm 3: Select-arm.

Input: $\mathcal{X}, H, T, y$
Output: arm to be played
Find $w_1^*, \cdots, w_K^*$ as the solution of $\arg\min_{\{w_a\}} \sum_{a=1}^K |w_a|$ s.t. $\sum_{a=1}^K w_a x^a = y$
Determine $p_a = \frac{|w_a^*|}{\sum_{k=1}^K |w_k^*|}$ for $a = 1, \cdots, K$
Find $\hat{a}_t = \arg\min_{a \in [K]: p_a > 0} \frac{T_a(t)}{p_a}$
return $\hat{a}_t$

Selecting a gap to explore: The algorithm starts by playing $E > d$ random arms such that M_{E+1} is nonsingular. While we have not optimized this parameter for the algorithm, taking $E = \min\{K, 3d\}$ has worked well in all of our numerical studies [5]. At any subsequent period $t > E$, the algorithm first finds the empirically best arm $i_t = \arg\max_{i \in [K]} \mu(\theta_t^\top x^i)$. Then, to check whether this arm is within ϵ distance of the true optimal arm, the algorithm takes a pessimistic approach. In particular, it finds another arm that is the most advantageous over arm i_t within the gap confidence sets; i.e., $j_t = \arg\max_{j \in [K], j \neq i_t} \Delta_t(j, i_t) + \beta_t(j, i_t)$. Note that this pessimistic gap consists of two components: the estimated gap and the uncertainty in the gap.

If this pessimistic gap is less than ϵ, the algorithm stops and declares the empirically best arm as the optimal arm. Otherwise, it selects an arm to reduce the uncertainty component in the identified gap. According to (10), this uncertainty is governed by $\|y_t\|_{M_t^{-1}}$ where $y_t = c_1 x^{i_t} - c_2 x^{j_t}$ and $c_1, c_2 = \arg\max_{c,c' \in [c_\mu, k_\mu]} \|cx^{i_t} - c'x^{j_t}\|_{M_t^{-1}}$.

Selecting an arm: While there are different ways to reduce $\|y_t\|_{M_t^{-1}}$, we follow the approach in [10]. With a slight abuse of notation, let us define $M_{\mathbf{z}_n} = \sum_{i=1}^n z_i z_i^\top$ for any sequence of feature vectors $\mathbf{z}_n = (z_1, z_2, \ldots, z_n) \in \mathcal{X}^n$, where n represents a generic time period. Let $\mathbf{z}_n^*(y) = \arg\min_{\mathbf{z}_n \in \mathcal{X}^n} \|y\|_{M_{\mathbf{z}_n}^{-1}}$ be the sequence of feature vectors that would have minimized the uncertainty in the

direction of y. For each arm $a \in [K]$, let $p_a(y)$ denote the relative frequency of x^a appearing in the sequence $\mathbf{z}_n$ when $n \to \infty$. As has been shown in Section 5.1 of [10],

$$p_a(y) = \frac{|w_a^*(y)|}{\sum_{k=1}^K |w_k^*(y)|}, \tag{11}$$

where $w^*(y) \in \mathbb{R}^K$ is the solution of the linear program

$$\min_{w \in \mathbb{R}^K} \|w\|_1 \quad \text{s.t.} \quad \sum_{k=1}^K w_k x^k = y. \tag{12}$$

To minimize the uncertainty in the direction of y_t, the algorithm plays the arm

$$a_{t+1} = \operatorname*{arg\,min}_{a \in [K]: p_a(y_t) > 0} \frac{T_a(t)}{p_a(y_t)}, \tag{13}$$

where $T_a(t)$ is the number of times arm a has been played prior to period t.

4. Theoretical analysis

In this section, we provide theoretical guarantees for the performance of the proposed algorithm. We prove that the algorithm indeed finds an $(\epsilon, \delta)-$optimal arm in Subsection 4.1 and provide an upper bound on the stopping time of the algorithm in Subsection 4.2.

4.1. (ϵ, δ)-optimality of the proposed algorithm

We start by proving that the confidence sets constructed in Section 3.1 hold with high probability at all times.

Proposition 1. *Fix δ and t such that $0 < \delta < \min(1, d/e)$ and $t > \max(2, d)$, where d is the feature dimension. Then the following holds with probability at least $1 - \delta$:*

$$\forall x, x' \in \mathcal{X}: \ \left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \le D_t(\delta) \max_{c, c' \in [c_\mu, k_\mu]} \|cx - c'x'\|_{M_t^{-1}}, \tag{14}$$

where

$$D_t(\delta) = \frac{2\kappa R}{c_\mu}\sqrt{2d \log t \log(d/\delta)}. \tag{15}$$

The proof is a slight modification of the proof of Proposition 1 in [3].

Proof. Let $g_t(\beta) = \sum_{l=1}^{t-1} \mu(\beta^\top x_l) x_l$. According to (5), the ML estimate θ_t satisfies $g_t(\theta_t) = \sum_{l=1}^{t-1} r_l x_l$. By the mean value theorem, there exist points z, z' with $c = \dot\mu(z), c' = \dot\mu(z')$ such that

$$\begin{aligned}
&\left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \\
&= \left|\left[\mu(\theta_t^\intercal x) - \mu(\theta^\intercal x)\right] - \left[\mu(\theta_t^\intercal x') - \mu(\theta^\intercal x')\right]\right|, \\
&= \left|\left[\dot\mu(z)(\theta_t - \theta)^\top x - \dot\mu(z')(\theta_t - \theta)^\top x'\right]\right|, \\
&= \left|(\theta_t - \theta)^\top (cx - c'x')\right|.
\end{aligned} \tag{16}$$

On the other hand, by the fundamental theorem of calculus, we have

$$g_t(\theta_t) - g_t(\theta) = G_t(\theta_t - \theta), \tag{17}$$

where

$$G_t = \int_0^1 \nabla g_t(s\theta + (1-s)\theta_t)\, ds.$$

The definition of $g_t(\beta)$ implies that for any β,

$$\nabla g_t(\beta) = \sum_{l=1}^{t-1} x_l x_l^\top \dot\mu(\beta^\top x_l).$$

It follows that $G_t \succeq c_\mu M_t \succeq c_\mu M_d \succeq \lambda_0 I \succ 0$. Hence, G_t and G_t^{-1} are positive definite and nonsingular. Therefore, from (16) and (17), it follows that

$$\begin{aligned}
&\left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \\
&= \left|(cx - c'x')^\top G_t^{-1}\left[g_t(\theta_t) - g_t(\theta)\right]\right|, \\
&\le \|cx - c'x'\|_{G_t^{-1}} \|g_t(\theta_t) - g_t(\theta)\|_{G_t^{-1}}.
\end{aligned} \tag{18}$$

The inequality $G_t \succeq c_\mu M_t$ implies that $G_t^{-1} \preceq c_\mu^{-1} M_t^{-1}$, and hence $\|y\|_{G_t^{-1}} \le \frac{1}{\sqrt{c_\mu}} \|y\|_{M_t^{-1}}$ for arbitrary $y \in \mathbb{R}^d$. Thus, by (18) we get

$$\begin{aligned}
&\left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \\
&\le \frac{2}{c_\mu} \|cx - c'x'\|_{M_t^{-1}} \|g_t(\theta_t) - g_t(\theta)\|_{M_t^{-1}}.
\end{aligned} \tag{19}$$

As has been shown in the proof of Proposition 1 in [3],

$$\|g_t(\theta_t) - g_t(\theta)\|_{M_t^{-1}} \le \kappa R \sqrt{2d \log t \log(d/\delta)} \tag{20}$$

holds with probability at least $1 - \delta$. Combining (19) and (20) shows that

$$\begin{aligned}
&\left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \\
&\le D_t(\delta) \|cx - c'x'\|_{M_t^{-1}}
\end{aligned} \tag{21}$$

holds with probability at least $1 - \delta$. Finally, because $c, c' \in [c_\mu, k_\mu]$, taking the maximum over c, c' on the right side of (21) completes the proof. □

The following theorem, which is a direct consequence of Proposition 1, shows that the confidence sets in (9) contain the true gaps at all times with high probability. We define event $\mathcal{E}$ to be

$$\mathcal{E} = \{\forall\, t > \max(2, d), \forall i, j \in [K] : \Delta(i, j) \in \mathcal{C}_t(i, j)\}, \tag{22}$$

and throughout this section we set α according to (8).

Theorem 1. *Let δ be such that $0 < \delta < \min(1, d/e)$. Then event $\mathcal{E}$ occurs with probability at least $1 - \delta$.*

Proof. For any $t > \max(2, d)$, let $\delta_t = \frac{6\delta}{\pi^2 t^2}$ and define the event $\mathcal{E}_t$ as

$$\mathcal{E}_t = \left\{\forall x, x' \in \mathcal{X}: \ \left|\left[\mu(\theta_t^\intercal x) - \mu(\theta_t^\intercal x')\right] - \left[\mu(\theta^\intercal x) - \mu(\theta^\intercal x')\right]\right| \le D_t(\delta_t) \max_{c, c' \in [c_\mu, k_\mu]} \|cx - c'x'\|_{M_t^{-1}}\right\}.$$

Applying Proposition 1 for δ_t implies that $\mathbb{P}[\mathcal{E}_t] \ge 1 - \delta_t$. The union bound then gives

$$\mathbb{P}[\cap_{t=1}^\infty \mathcal{E}_t] \ge 1 - \sum_{t=1}^\infty \delta_t = 1 - \frac{6\delta}{\pi^2} \sum_{t=1}^\infty \frac{1}{t^2} = 1 - \delta.$$

The proof is completed by noting that $\mathcal{E} = \cap_{t=1}^\infty \mathcal{E}_t$. □

With the confidence sets established, the next theorem proves $(\epsilon, \delta)-$optimality of the proposed algorithm.

Theorem 2. *Let $\epsilon > 0$ and $0 < \delta < \min(1, d/e)$ be arbitrary. Then the proposed algorithm is $(\epsilon, \delta)-$optimal; i.e., at its stopping time, the algorithm returns an arm $\hat{a}_\tau$ such that*

$$\mathbb{P}[\Delta(a^*, \hat{a}_\tau) > \epsilon] < \delta.$$

Proof. Let τ be the stopping time of the algorithm and let $\hat{a}_\tau$ be the returned arm. Suppose that $\Delta(a^*, \hat{a}_\tau) > \epsilon$. Then, according to line 5 of Algorithm 1, we have

$$\Delta(a^*, \hat{a}_\tau) > \epsilon \geq B(\tau) \geq \Delta_\tau(a^*, \hat{a}_\tau) + \beta_\tau(a^*, \hat{a}_\tau).$$

From this, we get that

$$\Delta(a^*, \hat{a}_\tau) - \Delta_\tau(a^*, \hat{a}_\tau) > \beta_\tau(a^*, \hat{a}_\tau),$$

which means that event $\mathcal{E}$ does not occur. According to Theorem 1, this can happen with probability at most δ. Thus, $\Delta(a^*, \hat{a}_\tau) > \epsilon$ happens with probability at most δ. □

4.2. An upper bound on the stopping time

In this subsection, we study the sample complexity of the proposed algorithm. The following theorem provides an upper bound on the number of experiments the proposed algorithm needs to carry out before identifying the optimal arm. Before stating the result, let us take $\Delta_{\min} = \min_{i \in [K]} \Delta_i$ to be the smallest gap and for any $\epsilon > 0$, define

$$H_\epsilon = \frac{18Kk_\mu}{\max(3\epsilon, \epsilon + \Delta_{\min})^2}, \tag{23}$$

which represents the complexity of the exploration problem in terms of the problem parameters.

Theorem 3. *Let τ be the stopping time of the proposed algorithm. Then*

$$\tau \leq \frac{64d\kappa^2R^2}{c_\mu^2} H_\epsilon \left(\log\left[\frac{64d\kappa^2R^2}{c_\mu^2} H_\epsilon \left(\pi\sqrt{\frac{d}{6\delta}} + 1 \right) \right] + \frac{c_\mu}{4\kappa R}\sqrt{\frac{K+1}{dH_\epsilon}} \right)^2 \tag{24}$$

is satisfied with probability at least $1 - \delta$. In asymptotic notation, (24) can be expressed as

$$\tau = O\left(\frac{d\kappa^2R^2}{c_\mu^2} H_\epsilon \left[\log\left(\frac{d^{3/4}\kappa R H_\epsilon^{1/2}}{c_\mu \delta^{1/4}} \right) \right]^2 \right). \tag{25}$$

Theorem 3 provides an upper bound on the stopping time of the proposed algorithm in terms of the parameters of the exploration problem; a lower bound is left for future work. As expected, the number of experiments required by the proposed algorithm before declaring a near-optimal arm decreases in the reward tolerance (ϵ) and the error probability (δ), and increases in the number of features (d) and arms (K). While the dependence on K (both here and in [10]) is disappointing, it is a by-product of our algorithm requiring a stopping mechanism to conclude that sufficient evidence has been gathered to declare an arm ϵ-optimal. In terms of the dependence on the dimension (d), the reward bound (R) and the complexity parameter (H_ϵ), the sample complexity in (25) is similar to that derived in [10] for linear bandits. Recall that k_μ and c_μ are characteristics of the function μ on its domain and are independent of how the arms' rewards are distributed in this domain. The main difference between (25) and the bound in [10] is the appearance of the factor k_μ/c_μ^2 in the complexity bound (25), which encodes the difficulty of learning the inverse link function μ.

Also, as the inverse link function μ becomes flatter on the boundaries of the input domain (i.e., has smaller c_μ), more samples are required to distinguish between different pairs of arms. Indeed, for the logistic bandit, where $\mu(z) = \frac{1}{1+e^{-z}}$ and $\dot{\mu}(z) = \frac{e^{-z}}{(1+e^{-z})^2}$, the minimum derivative c_μ can be exponentially small, and the extreme flatness in the logistic link can introduce large problem-dependent constants, which is a problem that also plagues earlier work (e.g., [3]). Consequently, in our numerical results in [5], we circumvent this issue by downsizing the theoretical length of the confidence sets.

The remainder of this subsection is devoted to proving Theorem 3, which requires a set of preliminaries. We start by introducing some additional notation. For any $i, j \in [K]$ and any t, let

$$(c_1^{ij}(t), c_2^{ij}(t)) = \underset{c, c' \in [c_\mu, k_\mu]}{\arg\min} \; \|cx^i - c'x^j\|_{M_t^{-1}},$$

and define $y_t^{ij} = c_1^{ij}(t)x^i - c_2^{ij}(t)x^j$. Note that with this notation, y_t defined in Algorithm 2 can be represented as $y_t = c_1^{i_t j_t}(t)x^{i_t} - c_2^{i_t j_t}(t)x^{j_t}$. For any $y \in \mathbb{R}^d$, let $\rho(y)$ be the optimal value of the linear program in (12); i.e.,

$$\rho(y) = \sum_{a=1}^{K} |w_a^*(y)|. \tag{26}$$

For any two real numbers a, b, we use the shorthand notation $a \vee b = \max(a, b)$.

Our proof for Theorem 3 relies on a number of results from the literature, which we state here for completeness. The following two lemmas are proved in [10].

Lemma 1 *(Lemma 1 of [10]). For any $y \in \mathbb{R}^d$, we have*

$$\|y\|_{M_t^{-1}} \leq \sqrt{\frac{\rho(y)}{Q_y(t)}},$$

where

$$Q_y(t) = \min_{k \in [K]:\, p_k(y) > 0} \frac{T_k(t)}{p_k(y)}.$$

Lemma 2 *(Lemma 4 of [10]). When event $\mathcal{E}$ holds, $B(t)$ satisfies the following bounds:*

1. *If either i_t or j_t is the best arm:*

$$B(t) \leq \min(0, -(\Delta_{i_t} \vee \Delta_{j_t}) + \beta_t(i_t, j_t)) + \beta_t(i_t, j_t),$$

2. *If neither i_t nor j_t is the best arm:*

$$B(t) \leq \min(0, -(\Delta_{i_t} \vee \Delta_{j_t}) + 2\beta_t(i_t, j_t)) + \beta_t(i_t, j_t).$$

The following lemma is proved in [1].

Lemma 3 *(Proposition 6 of [1]). For any $t > 0$, let $q(t) = a\sqrt{t} + b$ and $l(t) = \log(t)$, for some $a > 0$. Define $t^* = \frac{4}{a^2}\left[\log\left(\frac{4}{a^2}\right) - b\right]^2$. Then, for any positive t such that $t \geq t^*$, we have $q(t) > l(t)$.*

The following lemma establishes an upper bound on the solution of the linear program in (12).

Lemma 4. *Let $y_t = c_1^{i_t j_t}(t)x^{i_t} - c_2^{i_t j_t}(t)x^{j_t}$ and let $w^*(y_t)$ be the solution of (12). Then we have*

$$\|w^*(y_t)\|_\infty \le 2k_\mu.$$

Proof. Let $w' \in \mathbb{R}^K$ be such that $w'_{i_t} = c_1^{i_t j_t}(t)$, $w'_{j_t} = -c_2^{i_t j_t}(t)$ and all other elements of w' are zero. Clearly, w' satisfies the constraint in (12). Therefore, we have

$$\begin{aligned}\|w^*(y_t)\|_\infty \le \|w^*(y_t)\|_1 \le \|w'\|_1 &= |c_1^{i_t j_t}(t)| + |c_2^{i_t j_t}(t)| \\ &\le k_\mu + k_\mu = 2k_\mu. \quad \square\end{aligned}$$

With the above lemmas in place, we are ready to prove the following theorem.

Theorem 4. *If event $\mathcal{E}$ occurs, then the stopping time of the proposed algorithm satisfies*

$$\tau \le H_\epsilon C_\tau^2 + K + 1.$$

Proof. Suppose that event $\mathcal{E}$ occurs. Let $k \in [K]$ be an arbitrary arm, and let $t_k < \tau$ be the last round in which arm k was pulled. Because $B(t_k) > \epsilon$, Lemma 2 implies that

$$\min\left(0, -(\Delta_{i_{t_k}} \vee \Delta_{j_{t_k}}) + 2\beta_{t_k}(i_{t_k}, j_{t_k})\right) + \beta_{t_k}(i_{t_k}, j_{t_k}) \ge B(t_k) \ge \epsilon,$$

which in turn gives rise to the following three inequalities:

$$\epsilon \le \beta_{t_k}(i_{t_k}, j_{t_k}),\ \ \epsilon + \Delta_{i_{t_k}} \le 3\beta_{t_k}(i_{t_k}, j_{t_k}),\ \ \epsilon + \Delta_{j_{t_k}} \le 3\beta_{t_k}(i_{t_k}, j_{t_k}). \tag{27}$$

Rearranging (27) yields

$$\beta_{t_k}(i_{t_k}, j_{t_k}) \ge \max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right). \tag{28}$$

From Lemma 1, we have

$$Q_{y_{t_k}}(t_k) \le \frac{\rho(y_{t_k})}{\|y_{t_k}\|^2_{M_{t_k}^{-1}}},$$

which by substituting $\beta_{t_k}(i_{t_k}, j_{t_k}) = C_{t_k}\|y_{t_k}\|_{M_{t_k}^{-1}}$ from (10) gives

$$Q_{y_{t_k}}(t_k) \le \frac{\rho(y_{t_k})C_{t_k}^2}{\beta_{t_k}(i_{t_k}, j_{t_k})^2}. \tag{29}$$

Combining (28) and (29) gives

$$Q_{y_{t_k}}(t_k) \le \frac{\rho(y_{t_k})C_{t_k}^2}{\max\left(\epsilon, \frac{\epsilon+\Delta_{i_{t_k}}}{3}, \frac{\epsilon+\Delta_{j_{t_k}}}{3}\right)^2}. \tag{30}$$

Because arm k was selected by Algorithm 3 in round t_k, it follows that

$$T_k(t_k) = p_k(y_{t_k})Q_{y_{t_k}}(t_k). \tag{31}$$

Then, from (30), we get

$$\begin{aligned}
T_k(\tau) &= T_k(t_k) + 1 && \text{by construction,}\\
&= p_k(y_{t_k})Q_{y_{t_k}}(t_k) + 1 && \text{by (31),}\\
&\le \frac{p_k(y_{t_k})\rho(y_{t_k})}{\max\left(\epsilon, \frac{\epsilon+\Delta_{i_{t_k}}}{3}, \frac{\epsilon+\Delta_{j_{t_k}}}{3}\right)^2}C_{t_k}^2 + 1 && \text{by (30),}\\
&= \frac{|w_k^*(y_{t_k})|}{\max\left(\epsilon, \frac{\epsilon+\Delta_{i_{t_k}}}{3}, \frac{\epsilon+\Delta_{j_{t_k}}}{3}\right)^2}C_{t_k}^2 + 1 && \text{by (11) and (26),}\\
&\le \frac{\|w^*(y_{t_k})\|_\infty}{\max\left(\epsilon, \frac{\epsilon+\Delta_{i_{t_k}}}{3}, \frac{\epsilon+\Delta_{j_{t_k}}}{3}\right)^2}C_{t_k}^2 + 1,\\
&\le \frac{2k_\mu}{\max\left(\epsilon, \frac{\epsilon+\Delta_{i_{t_k}}}{3}, \frac{\epsilon+\Delta_{j_{t_k}}}{3}\right)^2}C_{t_k}^2 + 1 && \text{by Lemma 4,}\\
&\le \frac{2k_\mu}{\max\left(\epsilon, \frac{\epsilon+\min_{i\in[K]}\Delta_i}{3}\right)^2}C_{t_k}^2 + 1\\
&\le \frac{18k_\mu}{\max(3\epsilon, \epsilon+\Delta_{\min})^2}C_{t_k}^2 + 1\\
&\le \frac{18k_\mu}{\max(3\epsilon, \epsilon+\Delta_{\min})^2}C_{\tau}^2 + 1. && (32)
\end{aligned}$$

Note that $\tau = 1 + \sum_{k=1}^K T_k(\tau)$. Hence, it follows from (23) and (32) that

$$\tau \le H_\epsilon C_\tau^2 + K + 1,$$

which completes the proof. $\square$

We are now in a position to prove Theorem 3.

Proof of Theorem 3. Suppose that event $\mathcal{E}$ holds. Then

$$\begin{aligned}
\tau &\le H_\epsilon C_\tau^2 + K + 1 \ \text{ by Theorem 4,}\\
&= \frac{16d\kappa^2R^2}{c_\mu^2}H_\epsilon \log(\tau)\log\left(\pi\sqrt{\frac{d}{6\delta}}\tau\right) + K + 1 \ \text{ by (7),}\\
&\le \frac{16d\kappa^2R^2}{c_\mu^2}H_\epsilon \left(\log\left[\left(\pi\sqrt{\frac{d}{6\delta}} + 1\right)\tau\right]\right)^2 + K + 1, && (33)
\end{aligned}$$

where (33) follows by Jensen's inequality and the logarithmic arithmetic-geometric mean inequality (i.e., $\sqrt{\log x \log y} \le \frac{1}{2}(\log x + \log y) \le \log\left(\frac{x+y}{2}\right)$). Applying the inequality $x + y \le (\sqrt{x} + \sqrt{y})^2$ for any $x, y > 0$ to (33) yields

$$\sqrt{\tau} \le \frac{4\kappa R\sqrt{d}}{c_\mu}\sqrt{H_\epsilon}\log\left[\left(\pi\sqrt{\frac{d}{6\delta}} + 1\right)\tau\right] + \sqrt{K+1}. \tag{34}$$

Let $c_1 = \frac{4\kappa R\sqrt{d}}{c_\mu}\sqrt{H_\epsilon}$, $c_2 = \pi\sqrt{\frac{d}{6\delta}} + 1$ and define $a = \frac{1}{c_1\sqrt{c_2}}$, $b = -\frac{\sqrt{K+1}}{c_1}$. With a change of variable $z = c_2\tau$, (34) can be written as

$$a\sqrt{z} + b \le \log(z).$$

According to Lemma 3, this is possible only if

$$z \le z^* = \frac{4}{a^2}\left[\log\left(\frac{4}{a^2}\right) - b\right]^2.$$

Substituting for a, b and z shows that (34) holds only if

$$\tau \leq \frac{64d\kappa^2R^2}{c_\mu^2}H_\epsilon\left(\log\left[\frac{64d\kappa^2R^2}{c_\mu^2}H_\epsilon\left(\pi\sqrt{\frac{d}{6\delta}}+1\right)\right] + \frac{c_\mu}{4\kappa R}\sqrt{\frac{K+1}{dH_\epsilon}}\right)^2.$$

On the other hand, according to Theorem 1, event $\mathcal{E}$ holds with probability at least $1-\delta$. This completes the proof. □

5. Concluding remarks

Numerical results in an earlier version of this paper [5] show that – at least with $\delta = 0.05$ and $\epsilon = 0.1$ (i.e., being 95% confident of selecting an arm with a cure rate within 10% of optimal) – an optimal arm can be identified very quickly with our algorithm: when there are a large number of arms, the number of experiments performed can be much less than the total number of arms. This is achieved by learning about all arms whenever an arm is played. Indeed, our algorithm easily outperforms the GapE algorithm [4], which treats arms as independent with binary rewards. The numerical results in [5] also reveal that the amount of exploration undertaken by the proposed algorithm depends more strongly on the feature dimension d than on the number of arms K. This is due to the fact that the proposed algorithm builds confidence sets for the reward parameter, which is of dimension d, rather than for the reward of each arm separately. In addition, the stopping time in the computational results decreases with c_μ in a manner consistent with Theorem 3. Our algorithm, coupled with an approach to dimensionality reduction of the feature dimension such as PCA or machine learning tools, has the potential to streamline some nonlinear problems in drug design and other experimental settings.

References

[1] A. Antos, V. Grover, C. Szepesvári, Active learning in heteroscedastic noise, Theor. Comput. Sci. 411 (29–30) (2010) 2712–2728.
[2] S. Chen, T. Lin, I. King, M.R. Lyu, W. Chen, Combinatorial pure exploration of multi-armed bandits, Adv. Neural Inf. Process. Syst. (2014) 379–387.
[3] S. Filippi, O. Cappe, A. Garivier, C. Szepesvári, Parametric bandits: the generalized linear case, Adv. Neural Inf. Process. Syst. (2010) 586–594.
[4] V. Gabillon, M. Ghavamzadeh, A. Lazaric, Best arm identification: a unified approach to fixed budget and fixed confidence, Adv. Neural Inf. Process. Syst. (2012) 3212–3220.
[5] A. Kazerouni, L.M. Wein, Best arm identification in generalized linear bandits, arXiv:1905.08224, 2019.
[6] L. Li, Y. Lu, D. Zhou, Provably optimal algorithms for generalized linear contextual bandits, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 2071–2080.
[7] P. McCullagh, J.A. Nelder, Generalized Linear Models, second edition, Chapman & Hall/CRC, Boca Raton, FL, 1989.
[8] P. Rusmevichientong, J.N. Tsitsiklis, Linearly parameterized bandits, Math. Oper. Res. 35 (2) (2010) 395–411.
[9] D. Russo, B. Van Roy, Learning to optimize via posterior sampling, Math. Oper. Res. 39 (4) (2014) 1221–1243.
[10] L. Xu, J. Honda, M. Sugiyama, A fully adaptive algorithm for pure exploration in linear bandits, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2018, pp. 843–851.