

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Best Arm Identification in Generalized Linear Bandits

Abbas Kazerouni

Department of Electrical Engineering, Stanford University, Stanford, CA 94305, abbask@stanford.edu

Lawrence M. Wein

Graduate School of Business, Stanford University, Stanford, CA 94305, lwein@stanford.edu

Motivated by drug design, we consider the best-arm identification problem in generalized linear bandits. More specifically, we assume 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 problem is to minimize the number of arm pulls required to identify an arm that is sufficiently close to optimal with a sufficiently high probability. Building on recent progress in best-arm identification for linear bandits (Xu et al. 2018), we propose the first algorithm for best-arm identification for generalized linear bandits, provide theoretical guarantees on its accuracy and sampling efficiency, and evaluate its performance in various scenarios via simulation.

Key words: best arm identification, generalized linear bandits, sequential clinical trial

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. 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). Hence, with every pull of an arm, the decision maker refines the estimate of the unknown parameter vector and learns simultaneously about all arms.

Our motivation for studying this version of the bandit problem comes from drug design. The field of drug design is immense (Martin 2010). There are different types of drugs (e.g., organic small molecules or biopolymer-based, i.e. biopharmaceutical, drugs), several phases of development, and a variety of experimental and computational methods used to select compounds with favorable properties. Three characteristics of our bandit problem make it ideally suited to certain subproblems within the drug design process. The first characteristic is the desire to run as few experiments as possible with the goal of selecting a particular compound (i.e., an arm in our model) for the next stage of analysis or testing. Hence, the actual performance of the selected arms during testing is of secondary importance (these pre-clinical tests do not involve human subjects), giving rise to a pure exploration problem. Second, drugs can often be described by a vector of features, which may include calculated properties of molecules, or the three-dimensional structure of the molecule and how it relates to the three-dimensional structure of the biological target. Finally, in many of these experimental settings, a generalized linear model provides a better fit than the linear model. This is true when the outcome is binary; e.g., in animal experiments, there is either survival or death, or either the presence or absence of disease, and some *in vitro* assays are qualitative (i.e., binary output). It is also true when the outcome is counting data (e.g., the number of cells or animals that survived or died), or when the relationship between the outcome and the best linear predictor is nonlinear.

There may also be other potential applications of this model, e.g., where the aim is to choose the best (non-personalized) ad for an advertising campaign, and the observed output is binary (e.g., the ad led to a purchase or a click-through).

Literature Review. A recent review (Bubeck and Cesa-Bianchi 2012) categorizes bandit problems into three types: Markovian (maximizing discounted rewards, often in a Bayesian setting, and exemplified by the celebrated result in Gittins (1979)), stochastic (typically minimizing regret in a frequentist setting, with independent and identically distributed rewards (Lai and Robbins 1985)), and adversarial (a worst-case setting, where an adversary chooses the rewards (Auer et al. 2002)). We study the stochastic problem, but consider best-arm identification rather than regret minimization. In passing, we note that the ranking-and-selection problem (Kim and Nelson 2006) in the simulation literature has a similar goal to best-arm identification, although evaluating an alternative requires running simulation experiments.

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 (Even-Dar et al. 2006, Bubeck et al. 2009, Audibert et al. 2010, Gabillon et al. 2012, Kalyanakrishnan et al. 2012, Karnin et al. 2013, Chen et al. 2014). Different algorithms have been developed for variants of this setting, most of which use gap-based exploration where arms are played to reduce the uncertainty about the gaps between the rewards of pairs of arms. Because playing an arm reveals information only about that arm, each arm needs to be played several times to reduce the uncertainty about its reward. As such, these algorithms can be practically implemented only when the number of available arms is relatively small.

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 bandits (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 (e.g., Auer (2002), Rusmevichientong and Tsitsiklis (2010) and references therein) as well as alternative probabilistic models of arm dependence (e.g., Russo and Van Roy (2014) and references therein). In addition, contextual bandit models allow additional side information in each round, which can model, e.g., patient information in clinical trials or consumer information in online

advertising (e.g., Wang et al. (2005), Seldin et al. (2011), Goldenshluger and Zeevi (2013)). 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 (Filippi et al. 2010, Li et al. 2017).

However, relatively little work exists on best-arm identification in parametric bandits. The first analysis of best-arm identification in linear bandits (Soare et al. 2014) proposed a static exploration algorithm that was inspired by the transductive experiment design principle (Yu et al. 2006). This algorithm determines the sequence of to-be-played actions before making any observations and fails to adapt to the observed rewards. Recently (Xu et al. 2018), major progress has been made via 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.

Our Contribution. We adapt the gap-based exploration algorithm of Xu et al. (2018) 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 (Xu et al. 2018), the authors were able to convert the confidence set for the parameter vector into efficient confidence sets for the reward gaps. However, this approach breaks down in the generalized linear bandit; i.e., naively converting the confidence set for the parameter vector in Filippi et al. (2010) 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 simulation results in Section 5 and offer concluding remarks in Section 6.

2. Problem Formulation

Consider a decision maker who is seeking to find the best among a set of K available arms. We let $[K] = \{1, 2, \dots, K\}$ 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. We employ a generalized linear model (McCullagh and Nelder 1989) and assume that the reward of each arm a has a particular distribution in an exponential family with mean

$$\mu_a = \mu(\theta^\top x^a), \quad (1)$$

where $\theta \in \mathbb{R}^d$ is an unknown parameter that governs the reward of all arms and $\mu : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing function known as the inverse link function. Different choices for the function $\mu(z)$ in (1) 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.

The decision maker chooses an arm to play in each round t . If arm a_t is played in round t , a stochastic reward r_t is observed, which satisfies

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

Let $a^* = \arg \max_{a \in [K]} \mu_a$ be the optimal arm; i.e., the arm with the highest expected reward. 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}_τ . 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}_τ . 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 (ϵ, δ) -optimal if

$$\mathbb{P} \left[\mu(\theta^\top x^{a^*}) - \mu(\theta^\top x^{\hat{a}^\tau}) \geq \epsilon \right] \leq \delta. \quad (2)$$

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 (ϵ, δ) -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 $c_\mu = \inf_{\theta: \|\theta\| \leq S, x \in \mathcal{X}} \dot{\mu}(\theta^\top x) > 0$. For example, in the case of logistic regression, $R = 1, k_\mu = 1/4$ and c_μ depends on $\sup_{\theta: \|\theta\| \leq S, x \in \mathcal{X}} |\theta^\top x|$. 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

$$\Delta_i = \begin{cases} \mu(\theta^\top x^{a^*}) - \mu(\theta^\top x^i) & \text{if } i \neq a^* \\ \mu(\theta^\top x^i) - \max_{j \neq i} \mu(\theta^\top x^j) & \text{if } i = a^*. \end{cases} \quad (3)$$

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 Xu et al. (2018), 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 Filippi et al. (2010) but develop confidence sets directly for gaps instead of arm rewards.

Let $(a_1, r_1), (a_2, r_2), \dots, (a_{t-1}, r_{t-1})$ be the history of actions played and random rewards observed prior to period t , and let $x_l = x^{a_l}$ be shorthand notation for the feature vector associated with the arm played in period l . For any $t > 0$, let $M_t = \sum_{l=1}^{t-1} x_l x_l^\top$ be the empirical covariance matrix 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 , the Maximum Likelihood (ML) estimate of the reward parameter, θ_t , solves the equation

$$\sum_{l=1}^{t-1} (r_l - \mu(\theta_t^\top x_l)) x_l = 0. \quad (4)$$

Based on the estimated reward parameter, we can take

$$\Delta_t(i, j) = \mu(\theta_t^\top x^i) - \mu(\theta_t^\top x^j) \quad (5)$$

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

$$C_t = \alpha \sqrt{2d \log t \log \left(\frac{\pi^2 dt^2}{6\delta} \right)}, \quad (6)$$

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

$$\alpha = \frac{2\kappa R}{c_\mu} \quad (7)$$

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 Section 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 :

$$\mathcal{C}_t(i, j) = \left\{ \Delta \in \mathbb{R} : |\Delta - \Delta_t(i, j)| \leq C_t \max_{c, c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}} \right\}. \quad (8)$$

The confidence set defined in (8) is centered around the estimated gap between the two arms in (5) and, as we will show in Section 4, contains the true gap with high probability. To further simplify the representation, we let

$$\beta_t(i, j) = C_t \max_{c, c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}} \quad (9)$$

represent the width of the confidence set in (8).

Although we follow the basic approach in Filippi et al. (2010) in deriving these confidence sets, it is worth noting that they cannot be deduced from the results presented in Filippi et al. (2010). More precisely, an immediate use of the results in that paper gives rise to a very loose confidence set that is similar to (8) except that $\max_{c, c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_t^{-1}}$ is replaced by the looser factor $(\|x^i\|_{M_t^{-1}} + \|x^j\|_{M_t^{-1}})$.

Algorithm 1: GLGapE

Input: $\mathcal{X}, \epsilon, \delta, E, \alpha, S, L, R$

Output: approximately best arm \hat{a}_τ

- 1: **Initial Exploratory Phase:** Play E random arms and gather observations in D_E , and initialize $T_a(t)$'s for the played arms
 - 2: **for** $t > E$ **do**
 - 3: //select a gap to examine
 - 4: $i_t, j_t, B(t), y_t \leftarrow \text{select-gap}(\mathcal{X}, D_{t-1})$
 - 5: **if** $B(t) \leq \epsilon$ **then**
 - 6: **return** i_t as the best arm
 - 7: **end if**
 - 8: //select an action
 - 9: $a_t \leftarrow \text{select-arm}(\mathcal{X}, D_{t-1}, T(t), y_t)$
 - 10: Play a_t
 - 11: Observe r_t
 - 12: $D_t = D_{t-1} \cup \{(x_t, r_t)\}$
 - 13: $T_a(t) = T_a(t-1) + 1$
 - 14: **end for**
-

Algorithm 2: select-gap

Input: \mathcal{X}, D

Output: a gap to be examined

- 1: Find the ML estimate θ_t
 - 2: $i_t \leftarrow \arg \max_{i \in [K]} \mu(\theta_t^\top x^i)$
 - 3: $j_t \leftarrow \arg \max_{j \in [K], j \neq i_t} \Delta_t(j, i_t) + \beta_t(i_t, j)$
 - 4: Find $c_1, c_2 = \arg \max_{c, c' \in [c_\mu, k_\mu]} \|cx^{i_t} - c'x^{j_t}\|_{M_t^{-1}}$
 - 5: Let $y_t = c_1 x^{i_t} - c_2 x^{j_t}$
 - 6: $B(t) \leftarrow \Delta_t(j_t, i_t) + \beta_t(i_t, j_t)$
 - 7: **return** $i_t, j_t, B(t), y_t$
-

Algorithm 3: select-arm

Input: \mathcal{X}, D, T, y

Output: arm to be played

- 1: Find w_1^*, \dots, w_K^* as the solution of $\arg \min_{\{w_a\}} \sum_{a=1}^K |w_a|$ s.t. $\sum_{a=1}^K w_a x^a = y$
 - 2: Determine $p_a = \frac{|w_a^*|}{\sum_{k=1}^K |w_k^*|}$ for $a = 1, \dots, K$
 - 3: Find $\hat{a}_t = \arg \min_{a \in [K]: p_a > 0} \frac{T_a(t)}{p_a}$
 - 4: **return** \hat{a}_t
-

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.

Selecting a Gap to Explore: The algorithm starts by playing $E > d$ random arms such that the empirical covariance matrix M_{E+1} is nonsingular. 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 (9), this uncertainty is governed by $\|y_t\|_{M_t^{-1}}$ where $y_t = c_1 x^{i_t} - c_2 x^{j_t}$ and $c_1, c_2 = \arg \max_{c, c' \in [c_\mu, k_\mu]} \|c x^{i_t} - c' x^{j_t}\|_{M_t^{-1}}$.

Selecting an Arm: While there are different ways to reduce $\|y_t\|_{M_t^{-1}}$, we follow the approach in Xu et al. (2018). 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 $\mathbf{z}_n = (z_1, z_2, \dots, z_n) \in \mathcal{X}^n$, 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 $n \rightarrow \infty$. As has been shown in Section 5.1 of Xu et al. (2018),

$$p_a(y) = \frac{|w_a^*(y)|}{\sum_{k=1}^K |w_k^*(y)|}, \quad (10)$$

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

$$\min_{w \in \mathbb{R}^K} \|w\|_1 \quad \text{s.t.} \quad \sum_{k=1}^K w_k x^k = y. \quad (11)$$

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

$$a_{t+1} = \arg \min_{a \in [K]: p_a(y_t) > 0} \frac{T_a(t)}{p_a(y_t)}, \quad (12)$$

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 (ϵ, δ) -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 $1 - \delta$:*

$$\forall x, x' \in \mathcal{X}: \quad |[\mu(\theta_t^\top x) - \mu(\theta_t^\top x')] - [\mu(\theta^\top x) - \mu(\theta^\top x')]| \leq D_t(\delta) \max_{c, c' \in [c_\mu, k_\mu]} \|cx - c'x'\|_{M_t^{-1}}, \quad (13)$$

where

$$D_t(\delta) = \frac{2\kappa R}{c_\mu} \sqrt{2d \log t \log(d/\delta)}. \quad (14)$$

The proof is a slight modification of the proof of Proposition 1 in Filippi et al. (2010).

Proof. Let $g_t(\beta) = \sum_{l=1}^{t-1} \mu(\beta^\top x_l) x_l$. According to (4), the ML estimate θ_t satisfies $g_t(\theta_t) = 0$.

By the mean value theorem, there exist points z, z' with $c = \dot{\mu}(z), c' = \dot{\mu}(z')$ such that

$$\begin{aligned} |[\mu(\theta_t^\top x) - \mu(\theta_t^\top x')] - [\mu(\theta^\top x) - \mu(\theta^\top x')]| &= |[\mu(\theta_t^\top x) - \mu(\theta^\top x)] - [\mu(\theta_t^\top x') - \mu(\theta^\top x')]|, \\ &= |[\dot{\mu}(z)(\theta_t - \theta)^\top x - \dot{\mu}(z')(\theta_t - \theta)^\top x']|, \\ &= |(\theta_t - \theta)^\top (cx - c'x')|. \end{aligned} \quad (15)$$

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

$$g_t(\theta_t) - g_t(\theta) = G_t(\theta_t - \theta), \quad (16)$$

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 (15) and (16), it follows that

$$\begin{aligned} |[\mu(\theta_t^\top x) - \mu(\theta_t^\top x')] - [\mu(\theta^\top x) - \mu(\theta^\top x')]| &= |(cx - c'x')^\top G_t^{-1} [g_t(\theta_t) - g_t(\theta)]|, \\ &\leq \|cx - c'x'\|_{G_t^{-1}} \|g_t(\theta_t) - g_t(\theta)\|_{G_t^{-1}}. \end{aligned} \quad (17)$$

The inequality $G_t \succeq c_\mu$ implies that $G_t^{-1} \preceq c_\mu^{-1} M_t^{-1}$, and hence $\|y\|_{G_t^{-1}} \leq \frac{1}{\sqrt{c_\mu}} \|y\|_{M_t^{-1}}$ for arbitrary $y \in \mathbb{R}^d$. Thus, by (17) we get

$$|[\mu(\theta_t^\top x) - \mu(\theta_t^\top x')] - [\mu(\theta^\top x) - \mu(\theta^\top x')]| \leq \frac{2}{c_\mu} \|cx - c'x'\|_{M_t^{-1}} \|g_t(\theta_t) - g_t(\theta)\|_{M_t^{-1}}. \quad (18)$$

As has been shown in the proof of Proposition 1 in Filippi et al. (2010),

$$\|g_t(\theta_t) - g_t(\theta)\|_{M_t^{-1}} \leq \kappa R \sqrt{2d \log t \log(d/\delta)} \quad (19)$$

holds with probability at least $1 - \delta$. Combining (18) and (19) shows that

$$|[\mu(\theta_t^\top x) - \mu(\theta_t^\top x')] - [\mu(\theta^\top x) - \mu(\theta^\top x')]| \leq D_t(\delta) \|cx - c'x'\|_{M_t^{-1}} \quad (20)$$

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 (20) completes the proof. \square

The following theorem, which is a direct consequence of Proposition 1, shows that the confidence sets in (8) contain the true gaps at all times with high probability. We define event \mathcal{E} to be

$$\mathcal{E} = \{\forall t > \max(2, d), \forall i, j \in [K] : \Delta(i, j) \in \mathcal{C}_t(i, j)\}, \quad (21)$$

and throughout this section we set α according to (7).

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

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

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

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

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

With the confidence sets established, the next theorem proves (ϵ, δ) -optimality of the proposed algorithm.

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

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

$$H_\epsilon = \frac{18Kk_\mu}{\max(3\epsilon, \epsilon + \Delta_{\min})^2}, \quad (22)$$

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*

$$\tau \leq \frac{64d\kappa^2 R^2}{c_\mu^2} H_\epsilon \left(\log \left[\frac{64d\kappa^2 R^2}{c_\mu^2} H_\epsilon \left(\pi \sqrt{\frac{d}{6\delta}} + 1 \right) \right] + \frac{c_\mu}{4\kappa R} \sqrt{\frac{K+1}{dH_\epsilon}} \right)^2 \quad (23)$$

is satisfied with probability at least $1 - \delta$. In asymptotic notation, (23) can be expressed as

$$\tau = O \left(\frac{d\kappa^2 R^2}{c_\mu^2} H_\epsilon \left[\log \left(\frac{d^{3/4} \kappa R H_\epsilon^{1/2}}{c_\mu \delta^{1/4}} \right) \right]^2 \right). \quad (24)$$

Theorem 3 provides an upper bound on the stopping time of the proposed algorithm in terms of the parameters of the exploration problem. 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). In terms of the dependence on dimension (d), reward bound (R) and the complexity parameter (H_ϵ), the sample complexity in (24) is similar to that derived in Xu et al. (2018) for linear bandits. The main difference is the appearance of the factor κ/c_μ in the complexity bound (24), which encodes the difficulty of learning the inverse link function μ . 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.

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

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

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 (11); i.e.,

$$\rho(y) = \sum_{a=1}^K |w_a^*(y)|. \quad (25)$$

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 Xu et al. (2018).

LEMMA 1 (Lemma 1 of Xu et al. (2018)). *For any $y \in \mathbb{R}^d$, we have*

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

where

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

LEMMA 2 (Lemma 4 of Xu et al. (2018)). *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 Antos et al. (2010).

LEMMA 3 (Proposition 6 of Antos et al. (2010)). *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 (11).

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 (11). Then we have

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

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 (11). Therefore, we have

$$\|w^*(y_t)\|_\infty \leq \|w^*(y_t)\|_1 \leq \|w'\|_1 = |c_1^{i_t j_t}(t)| + |c_2^{i_t j_t}(t)| \leq k_\mu + k_\mu = 2k_\mu.$$

□

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

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

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

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

which in turn gives rise to the following three inequalities:

$$\epsilon \leq \beta_{t_k}(i_{t_k}, j_{t_k}), \quad \epsilon + \Delta_{i_{t_k}} \leq 3\beta_{t_k}(i_{t_k}, j_{t_k}), \quad \epsilon + \Delta_{j_{t_k}} \leq 3\beta_{t_k}(i_{t_k}, j_{t_k}). \quad (26)$$

Rearranging (26) yields

$$\beta_{t_k}(i_{t_k}, j_{t_k}) \geq \max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right). \quad (27)$$

From Lemma 1, we have

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

which by substituting $\beta_{t_k}(i_{t_k}, j_{t_k}) = C_{t_k} \|y_{t_k}\|_{M_{t_k}^{-1}}$ from (9) gives

$$T_{y_{t_k}}(t_k) \leq \frac{\rho(y_{t_k})C_{t_k}^2}{\beta_{t_k}(i_{t_k}, j_{t_k})^2}. \quad (28)$$

Combining (27) and (28) gives

$$T_{y_{t_k}}(t_k) \leq \frac{\rho(y_{t_k})C_{t_k}^2}{\max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right)^2}. \quad (29)$$

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

$$T_k(t_k) = p_k(y_{t_k})T_{y_{t_k}}(t_k). \quad (30)$$

Then, from (29), we get

$$\begin{aligned} T_k(\tau) &= T_k(t_k) + 1 && \text{by construction,} \\ &= p_k(y_{t_k})T_{y_{t_k}}(t_k) + 1 && \text{by (30),} \\ &\leq \frac{p_k(y_{t_k})\rho(y_{t_k})}{\max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right)^2} C_{t_k}^2 + 1 && \text{by (29),} \\ &= \frac{|w_k^*(y_{t_k})|}{\max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right)^2} C_{t_k}^2 + 1 && \text{by (10) and (25),} \\ &\leq \frac{\|w^*(y_{t_k})\|_\infty}{\max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right)^2} C_{t_k}^2 + 1, \\ &\leq \frac{2k_\mu}{\max\left(\epsilon, \frac{\epsilon + \Delta_{i_{t_k}}}{3}, \frac{\epsilon + \Delta_{j_{t_k}}}{3}\right)^2} C_{t_k}^2 + 1 && \text{by Lemma 4,} \\ &\leq \frac{2k_\mu}{\max\left(\epsilon, \frac{\epsilon + \min_{i \in [K]} \Delta_i}{3}\right)^2} C_{t_k}^2 + 1 \\ &\leq \frac{18k_\mu}{\max(3\epsilon, \epsilon + \Delta_{\min})^2} C_{t_k}^2 + 1 \\ &\leq \frac{18k_\mu}{\max(3\epsilon, \epsilon + \Delta_{\min})^2} C_\tau^2 + 1. \end{aligned} \quad (31)$$

Note that $\tau = 1 + \sum_{k=1}^K T_k(\tau)$. Hence, it follows from (22) and (31) that

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

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

$$\begin{aligned}
 \tau &\leq H_\epsilon C_\tau^2 + K + 1 \text{ by Theorem 4,} \\
 &= \frac{16d\kappa^2 R^2}{c_\mu^2} H_\epsilon \log(\tau) \log\left(\pi\sqrt{\frac{d}{6\delta}}\tau\right) + K + 1 \text{ by (6),} \\
 &\leq \frac{16d\kappa^2 R^2}{c_\mu^2} H_\epsilon \left(\log\left[\left(\pi\sqrt{\frac{d}{6\delta}} + 1\right)\tau\right]\right)^2 + K + 1,
 \end{aligned} \tag{32}$$

where (32) follows by Jensen's inequality and the logarithmic arithmetic-geometric mean inequality (i.e., $\sqrt{\log x \log y} \leq \frac{1}{2}(\log x + \log y) \leq \log\left(\frac{x+y}{2}\right)$). Applying the inequality $x + y \leq (\sqrt{x} + \sqrt{y})^2$ for any $x, y > 0$ to (32) yields

$$\sqrt{\tau} \leq \frac{4\kappa R\sqrt{d}}{c_\mu} \sqrt{H_\epsilon} \log\left[\left(\pi\sqrt{\frac{d}{6\delta}} + 1\right)\tau\right] + \sqrt{K+1}. \tag{33}$$

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 $z = c_2\tau$, (33) can be written as

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

According to Lemma 3, this is possible only if

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

Substituting for a, b and z shows that (33) holds only if

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

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

5. Numerical Results

In this section, we test the proposed algorithm in various scenarios using synthetic data in Subsection 5.2 and drug design data in Subsection 5.3. The experimental details are described in Subsection 5.1.

5.1. Experimental Details

All scenarios considered here involve a logistic inverse link function with binary observations. Recall that the proposed algorithm starts with an exploratory phase in which E random arms are played to ensure that M_{E+1} is nonsingular. Although we have not optimized this parameter in our numerical studies, we use $E = \min(K, 3d)$ throughout all simulations here.

The confidence sets defined in (6) and (8) rely on the free parameter α . Our theoretical analysis in Section 4 suggests the use of $\alpha = \frac{2\kappa R}{c_\mu}$ from (7). However, the theoretical confidence sets derived for generalized linear settings in Filippi et al. (2010) are loose and the performance can be improved by scaling these sets. Although Section 4.2 of Filippi et al. (2010) uses an asymptotic analysis to shrink these sets in the regret-minimization setting (a finite-sample result in Theorem 1 of Li et al. (2017) provides a rigorous justification of this analysis), we simply note that all the rewards and reward gaps are in $[0, 1]$ in the case of a logistic reward function while the width of the confidence set defined in (8) may be larger than one. Consequently, we set α so as to constrain the confidence set widths in (8) to the interval $[0, 1]$, which can be achieved by setting $\frac{2\kappa R}{c_\mu}$ times $\max_{i,j \in [K]} \beta_{E(i,j)}$ equal to 1, yielding via (6) and (9)

$$\alpha = \frac{1}{\frac{2\kappa R}{c_\mu} \sqrt{2d \log E \log \left(\frac{\pi^2 d E^2}{6\delta} \right) \max_{i,j \in [K]} \max_{c,c' \in [c_\mu, k_\mu]} \|cx^i - c'x^j\|_{M_E^{-1}}}}, \quad (34)$$

which is determined at the end of the exploratory phase of length E .

Throughout our simulations, we fix the confidence parameter $\delta = 0.05$, which theoretically guarantees the accuracy of the proposed algorithm with 95% confidence. The results in Figs. 1-5 are generated by averaging over tens of random realizations of the problem. In all cases, the proposed algorithm achieved the desired accuracy (as dictated by ϵ) with probability more than 0.95.

5.2. Synthetic Data Simulations

To create a synthetic data set, we generate the reward parameter θ by sampling from a d -dimensional standard multivariate Gaussian distribution. The feature vector associated with

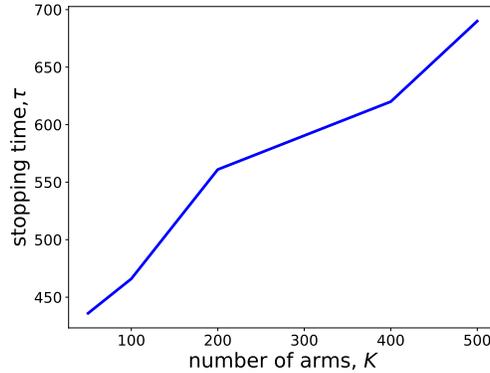


Figure 1 The amount of exploration required by the proposed algorithm versus the number of available arms, with tolerance $\epsilon = 0.1$, confidence parameter $\delta = 0.05$ and $d = 10$ features.

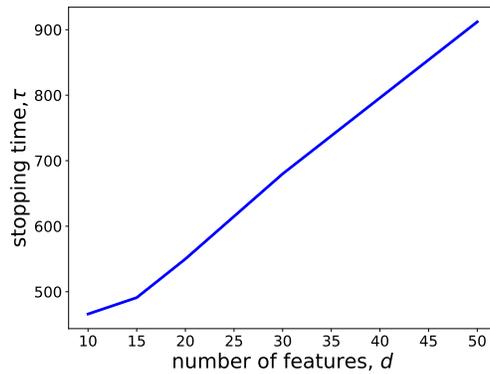


Figure 2 The amount of exploration required by the proposed algorithm versus the number of features, with tolerance $\epsilon = 0.1$, confidence parameter $\delta = 0.05$ and $K = 10$ arms.

each of the K arms is randomly generated such that each of its components is uniformly distributed in $[-1, 1]$. Acting as if we do not know the reward parameter θ , we run the proposed algorithm to identify the best among the available K arms under different scenarios. Recalling that the parameters L , R and S are defined via $\|x\|_2 \leq L$, $r_t \leq R$ and $\|\theta\|_2 \leq S$, we have $L \leq d$ and $R = 1$. Note that the parameter S is used only in determining c_μ , which we treat as given in our simulations; hence, there is no need to specify S .

We first fix the tolerance parameter $\epsilon = 0.1$ and the feature dimension $d = 10$, and plot the stopping time τ as a function of the number of arms K (Fig. 1). We then fix $\epsilon = 0.1$ and the number of arms $K = 100$, and compute τ as a function of the feature dimension d (Fig. 2). These plots reveal that the amount of exploration undertaken by the proposed algorithm depends more

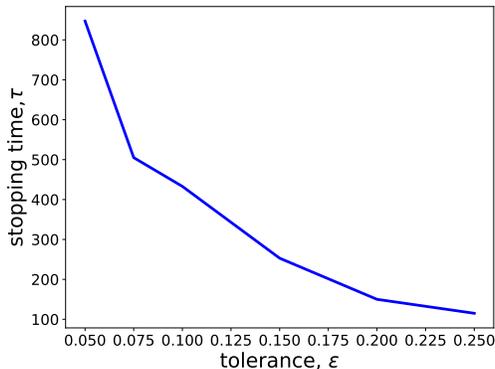


Figure 3 The amount of exploration required by the proposed algorithm versus the tolerance level ϵ , with confidence parameter $\delta = 0.05$ $K = 100$ arms, and $d = 10$ features.

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.

Fixing the feature dimension $d = 10$ and the number of arms $K = 100$, we plot the stopping time τ versus the tolerance parameter ϵ (Fig. 3). As expected, a smaller tolerance requires a larger amount of exploration in order to guarantee that the identified arm is within the acceptable distance of the optimal arm.

Recall that exploration algorithms for independent arms require playing each arm many times. To assess the improvement over this approach, we implement the GapE algorithm (Gabillon et al. 2012), which treats arms as independent with binary rewards, to an instance of this synthetic scenario with $K = 50$ arms, $d = 10$ features and $\epsilon = 0.1$. On average, the GapE algorithm requires 60k experiments before identifying a near-optimal arm, while our proposed algorithm identifies a near-optimal arm after only 436 experiments.

We know of no other algorithms that are natural comparisons to our proposed algorithm: the algorithms in Xu et al. (2018) and Soare et al. (2014) use a linear model that does not accommodate binary outcomes, and regret-minimization algorithms are not typically tested against best-arm identification algorithms because the former do not have any stopping criteria, which precludes us from comparing our algorithm to the one in Filippi et al. (2010).

5.3. Drug Design Simulations

In this subsection, we consider the problem of identifying the most effective among a large set of designed drugs for a certain disease. We use the ChEMBL data repository (Gaulton et al. 2011) and download the data set describing 982 small molecules designed for targeting the Cyclin-Dependent Kinase 2/Cyclin A protein complex. The covariates for each small molecule are associated with its three-dimensional structure, which can be represented as a 1024-dimensional binary vector. To construct such a representation, we compute the Morgan Fingerprint (Riniker and Landrum 2013) of each molecule with radius 4. Then, by applying a Principal Components Analysis (PCA) to these sparse high-dimensional binary vectors, we generate a feature vector with the desired dimension.

The generalized linear model allows for the random experimental output to have values that are continuous, integer or binary, depending on the link function that is used. If we want to use a model with binary data (i.e., the logit inverse link function), we would need binomial or Bernoulli data for each arm. However, we have been unable to find data like these. Rather, the output data in the data set from Gaulton et al. (2011) are the affinity of each small molecule, as measured by the IC_{50} (the concentration of an inhibitor where the biological function, typically the binding to a protein, is reduced by half). The IC_{50} is a deterministic number (there are no data in Gaulton et al. (2011) on the possible measurement errors) that is an indicator of the effectiveness of each small molecule.

To adapt these outcome data into our modeling framework, we convert these deterministic IC_{50} values into random binary outcomes as follows. First, we assume the drug produces a cure if the IC_{50} value is less than 10^6 nM (nanomolar), and does not produce a cure if the IC_{50} is greater than or equal to 10^6 nM. Next, given the feature vectors and these artificial cure outcomes, we fit a logistic regression model and estimate the cure rate of each drug. Finally, by assuming that these estimated cure rates are the unknown expected rewards of the drugs, we apply the proposed algorithm to find the best drug; i.e., in our simulations, each time a drug is tested, we observe a Bernoulli random variable with its mean equal to the drug's expected reward.

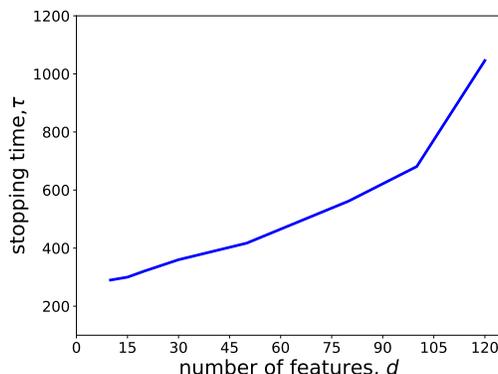


Figure 4 The amount of exploration required by the proposed algorithm versus the number of features in the drug design simulations, with tolerance $\epsilon = 0.1$, confidence parameter $\delta = 0.05$ and $K = 400$ arms.

We first fix the tolerance parameter $\epsilon = 0.1$ and the number of arms $K = 400$ by randomly choosing 400 of the 982 molecules in Gaulton et al. (2011). The stopping time τ is plotted against the feature dimension d , which is varied using PCA (Fig. 4). As in the synthetic data simulations, this relationship is increasing and – when the number of features is smaller than 50 – the proposed algorithm finishes exploration before all the arms are explored.

To roughly assess the effectiveness of the proposed algorithm in its choice of best arm, we note that the 982 IC_{50} values in the data set range from 0.3nM to 1.5×10^6 nM, and that the best (i.e., lowest-valued) among the 400 molecules in the experiment has an $IC_{50} = 0.4$ nM. The algorithm’s declared optimal arm has an $IC_{50} = 11.99$ nM. However, because considerable information is lost when we binarize the data (e.g., two drugs with IC_{50} values of 0.5nM and 10^5 nM would both be assigned a binary value of one for a cure), we should not expect our algorithm to identify the drug with the lowest IC_{50} .

We next fix $\epsilon = 0.1$ and the feature dimension $d = 20$, and plot the stopping time τ versus the number of arms K (Fig. 5), where for each value of K we randomly choose K arms out of the 982 small molecules in Gaulton et al. (2011). When there are more than 200 arms in Fig. 5, the number of experiments carried out by the proposed algorithm is smaller than the total number of arms.

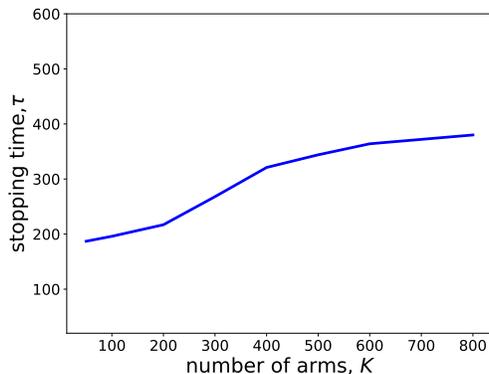


Figure 5 The amount of exploration required by the proposed algorithm versus the number of arms in the drug design simulations, with tolerance $\epsilon = 0.1$, confidence parameter $\delta = 0.05$ and $d = 20$ features.

6. Conclusion

Motivated by drug design, we consider the best-arm identification problem for a generalized linear bandit, 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. We use ideas from Filippi et al. (2010) on building confidence sets, but apply these directly to confidence sets for gaps between pairs of arms rather than confidence sets for rewards of arms, which allows us to generalize the results in Xu et al. (2018) from the linear bandit model to the generalized linear bandit model. After performing our analysis, we became aware of recent work (Li et al. 2017) that improves the regret rate in Filippi et al. (2010) by a factor of \sqrt{d} . We leave it to future work to investigate whether the results in Li et al. (2017) can be used to improve our theoretical bounds.

Numerical results 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) – can be identified very quickly: when there are a large number of arms, the number of experiments performed can be much less than the total number of arms (Fig. 5). This is achieved by learning about all arms whenever an arm is played. 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.

Acknowledgments

This research was supported by the Graduate School of Business, Stanford University (L.M.W.) and a Stanford Graduate Fellowship (A.K.).

References

- Antos, A., Grover, V., and Szepesvri, C. (2010). Active learning in heteroscedastic noise. *Theoretical Computer Science* 411(29-30), 2712-2728.
- Audibert, J. Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. *Proceedings of the Twenty-third Annual Conference on Learning Theory*, 41-53.
- Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(3), 397-422.
- Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. (2002). The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing* 32(1):48-77.
- Bubeck, S., and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1), 1-122.
- Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. *International Conference on Algorithmic Learning Theory* (Springer, Berlin), 23-37.
- Chen, S., Lin, T., King, I., Lyu, M.R., and Chen, W. (2014). Combinatorial pure exploration of multi-armed bandits. *Advances in Neural Information Processing Systems*, 379-387.
- Even-Dar, E., Mannor, S. and Mansour, Y., (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research* 7, 1079-1105.
- Filippi, S., Cappe, O., Garivier, A., and Szepesvri, C. (2010). Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 586-594.
- Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems*, 3212-3220.

- Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J.P. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* 40(D1), D1100-D1107.
- Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)* 41(2), 148-164.
- Goldenshluger, A., and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems* 3(1), 230-261.
- Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC Subset Selection in Stochastic Multi-armed Bandits. *Proceedings of the 29th International Conference on Machine Learning*, 655-662.
- Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. *Proceedings of the 30th International Conference on Machine Learning*, 1238-1246.
- Kazerouni, A., Ghavamzadeh, M., Abbasi, Y., and Van Roy, B. (2017). Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 3910-3919.
- Kim, S.H., and Nelson, B.L. (2006). Selecting the best system. In *Simulation, Handbooks in Operations Research and Management Science*, Editors, Henderson, S.G. and Nelson, B.L., 13, 501-534, Elsevier (Amsterdam).
- Lai, T.L., and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1), 4-22.
- Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. *Proceedings of the 34th International Conference on Machine Learning*, 2071-2080.
- Martin, Y. C. (2010). *Quantitative Drug Design: A Critical Introduction*, Second Edition. CRC Press (Boca Raton, FL).
- McCullagh, P. and Nelder, J.A., (1989). *Generalized Linear Models*, Second Edition. Chapman & Hall/CRC (Boca Raton, FL).
- Riniker, S., and Landrum, G.A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* 5(1), 26.
- Rusmevichientong, P., and Tsitsiklis, J.N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research* 35(2), 395-411.

- Russo, D., and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4), 1221-1243.
- Seldin, Y., Auer, P., Shawe-Taylor, J.S., Ortner, R., and Laviolette, F. (2011). PAC-Bayesian analysis of contextual bandits. *Advances in Neural Information Processing Systems*, 1683-1691.
- Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 828-836.
- Wang, C.C., Kulkarni, S.R., and Poor, H.V. (2005). Arbitrary side observations in bandit problems. *Advances in Applied Mathematics* 34(4), 903-938.
- Xu, L., Honda, J. and Sugiyama, M. (2018). A fully adaptive algorithm for pure exploration in linear bandits. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 843-851.
- Yu, K., Bi, J., and Tresp, V. (2006). Active learning via transductive experimental design. *Proceedings of the 23rd International Conference on Machine Learning*, 1081-1088.