Minimax Concave Penalized Multi-Armed Bandit Model with High-Dimensional Convariates

Xue Wang [*1] **Mike Mingcheng Wei** [*2] **Tao Yao** [*1]

Abstract

In this paper, we propose a Minimax Concave Penalized Multi-Armed Bandit (MCP-Bandit) algorithm for a decision-maker facing high-dimensional data with latent sparse structure in an online learning and decision-making process. We demonstrate that the MCP-Bandit algorithm asymptotically achieves the optimal cumulative regret in the sample size T, $O(\log T)$, and further attains a tighter bound in both the covariates dimension d and the number of significant covariates s, $O(s^2(s + \log d))$. In addition, we develop a linear approximation method, the 2-step Weighted Lasso procedure, to identify the MCP estimator for the MCP-Bandit algorithm under non-i.i.d. samples. Using this procedure, the MCP estimator matches the oracle estimator with high probability. Finally, we present two experiments to benchmark our proposed the MCP-Bandit algorithm to other bandit algorithms. Both experiments demonstrate that the MCP-Bandit algorithm performs favorably over other benchmark algorithms, especially when there is a high level of data sparsity or when the sample size is not too small.

1. Introduction

Individual-level data have become increasingly accessible in the Internet era, and decision-makers have accelerated data accumulation with extraordinary speed in a variety of industries, such as health-care, retail, and advertising. The growing availability of user-specific data, such as medical records, demographics, geographic, browsing/shopping history, etc., provides decision-makers with unprecedented opportunities to tailor decisions to individual users. For example, doctors (i.e., decision-makers) can personalize treatments for patients (i.e., users) based on their medical history, clinical tests, and biomarkers (i.e., user-specific data). These data are often collected sequentially over time, during which decision-makers adaptively learn to predict users' responses to each decision as a function of users' covariates (i.e., the exploration phase) and optimally adjust decisions to maximize their rewards (i.e., the exploitation phase) – an *online* learning and decision-making process. We will adopt the multi-armed bandit model (Robbins 1952) to study this process.

Individual-level data are typically presented in a high-dimensional fashion, which poses significant computational and statistical challenges. In particular, traditional statistic methods, such as Ordinary Least Squares (OLS), require a substantial number of samples (e.g., the sample size must be larger than the covariates dimension) in order to be deemed computationally feasible. Yet, under the high-dimensional data structure, learning the accurate predictive models requires even more data samples, which are obtained through costly trials or experiments. Learning algorithms, such as Lasso (Tibshirani 1996) and Minimax Concave Penalized (MCP) (Zhang et al. 2010), have been developed to recover the latent sparse data structure for high-dimensional data. Therefore, compared to traditional statistic methods, Lasso and MCP use significantly fewer data samples and deliver better performance in high-dimensional settings.

In this paper, we propose a new algorithm, the MCP-Bandit algorithm, for online learning and decision-making processes in high-dimensional settings. Our algorithm follows the ideas of the bandit model to balance the exploration-and-exploitation trade-off and adopts the MCP estimator to expedite the convergence of our parameter estimations to their true values and to improve their statistical performances. Since we focus on the multi-arm bandit model that mixes the exploitation and exploration phases, samples generated under the exploitation phase are typically not i.i.d., which significantly challenges the existing MCP literature. Therefore, we adopt a matrix perturbation technique to derive new oracle inequalities for MCP under non-i.i.d samples. To our best knowledge, our work is the first one which applies the MCP techniques to handle non-i.i.d samples. In

[*] Equal contribution [1] Pennsylvania State University, University Park, PA, USA [2] University at Buffalo, Buffalo, NY, USA. Correspondence to: Xue Wang <xzw118@psu.edu>, Mike Mingcheng Wei <mcwei@buffalo.edu>, Tao Yao <taoyao@psu.edu>.

Proceedings of the 35^{th} *International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

addition, although it is statistically favorable to adopt the MCP estimator, solving the MCP estimator (a NP-complete problem) could be computationally challenging. We propose a linear approximation method, 2-step weighted Lasso procedure (2sWL), under the bandit setting as an efficient solution approach to tackle this challenge. It also guarantees that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability.

We theoretically demonstrate that the MCP-Bandit algorithm can notably improve the cumulative regret bound comparing to existing high-dimensional bandit algorithms and attain the optimal regret bound on the sample size dimension. In particular, we benchmark our MCP-Bandit algorithm to an oracle counterpart where all parameters are common knowledge and adopt the expected cumulative regret (i.e., the difference in rewards achieved by the oracle case and our MCP-Bandit algorithm) as the performance measure. We show that the maximal cumulative regret of the MCP-Bandit algorithm over T users (i.e., a sample size of T) is at most $O(\log T)$, which is the optimal/lowest theoretical bound for all possible algorithms (Goldenshluger et al. 2013) and improves the $O((\log T)^2)$ bound of the Lasso-Bandit algorithm developed in Bastani & Bayati (2015). It is worth noting that the sparse structure of the high-dimensional data typically implies that the dimension of significant covariates (i.e., the covariates with non-zero coefficients) is much smaller than that of all covariates (i.e. $s \leq O(\log d)$). We show that the cumulative regret of the MCP-Bandit algorithm in the covariates dimension, d, and the number of significant covariates, s, is bounded by $O(s^2(s + \log d))$, which is a tighter bound than the Lasso-Bandit algorithm ($O(s^2 \log^2 d)$) in Bastani & Bayati (2015).

At last, through one synthetic-data-based experiment and one real-data-based experiment (i.e., Warfarin Dosing experiment), we evaluate the MCP-Bandit algorithm's performance compared to other state-of-the-art bandit algorithms designed both in low-dimensional settings and in high-dimensional settings. We find that the MCP-Bandit algorithm performs favorably in both experiments, especially when the data sparsity level is high. Furthermore, when the sample size is not extremely small, the MCP-Bandit algorithm appears to be the most beneficial. These observations suggest that the MCP-Bandit algorithm delivers great performance for high-dimensional data and provides a smooth transaction from data-poor regime to data-rich regime for decision-makers.

2. Literature

This research is closely related to the exploration-exploitation trade off in the multi-armed bandit literature. Generally, there are two approaches to model users' reward functions. The decision-maker could make no parametric assumption on the reward functions (Yang et al. 2002; Rigollet & Zeevi 2010), but these algorithms' performances degenerate exponentially as the covariates' dimension grows. Therefore, we follow the second approach, a parametric approach, and focus on the case where the arm rewards follow a linear function of users' covariates (Auer 2002; Rusmevichientong & Tsitsiklis 2010; Chu et al. 2011; Agrawal & Goyal 2013). Under this approach, Dani et al. (2008), Abbasi-Yadkori et al. (2011), and Abbasi-Yadkori et al. (2012) show that the expected cumulative regret is bounded by $O(\sqrt{T})$ in both low-dimensional and high-dimensional settings. This bound is further improved to $O(\log T)$ by Goldenshluger et al. (2013) under a OLS-Bandit algorithm in a low-dimensional setting and to $O((\log T)^2)$ by Bastani & Bayati (2015) under the Lasso-Bandit algorithm in a high-dimensional setting. This is a significant improvement from $O(\sqrt{T})$, especially as the sample size becomes larger. Our research closely follows Goldenshluger et al. (2013) and Bastani & Bayati (2015) and shows that the expected cumulative regret for our proposed the MCP-Bandit algorithm is bounded by $O(\log T)$ in both low-dimensional and high-dimensional settings. This regret bound is essentially the lowest theoretical bound for all possible algorithms (Goldenshluger et al. 2013). Besides the improved dependence on sample size dimension T, the MCP-Bandit algorithm will provide a better bound in the covariates dimension d. In the literature, the dimensionality's dependence is common to be polynomial in d. (Auer 2002; Rusmevichientong & Tsitsiklis 2010; Chu et al. 2011; Agrawal & Goyal 2013; Goldenshluger et al. 2013). Such polynomial dependence in d can be quite costly and prohibit the practical adoption of these algorithms in high-dimensional settings. Recently the Lasso-Bandit algorithm proposed by (Bastani & Bayati, 2015) reduces the dimensionality's dependence to be log-polynomial in d, i.e., $O(\log^2 d)$. However, the price to pay is that the Lasso-Bandit algorithm could only attain a suboptimal dependence in T. The proposed MCP bandit algorithm achieves a tighter log-polynomial dependence in d, (i.e.,$O(\log d)$) and the optimal dependence in T, (i.e., $O(\log T)$) simultaneously.

Our research is also connected to the literature of statistical learning algorithms that have been developed to recover the latent sparse structure and, therefore, provide a good performance guarantee even under limited samples in high-dimensional settings. In particular, Lasso, proposed by Tibshirani (1996), is able to identify a sparse subset of user covariates and produce good estimations using limited samples. However, the Lasso estimator can be biased (Fan & Li 2001). To address this issue, MCP has been proposed by Zhang et al. (2010) and is shown to be unbiased and can reach near optimal statistical performance, both theoretically and numerically. Although the MCP estimator has statistical performance that is more desirable, solving the MCP estimator is an NP-Complete problem due to the non-convexity penalty function (Liu et al.). The literature has since proposed various algorithms, such as MIPGO (Liu et al. 2016) and LLA (Zou 2006; Fan et al. 2014; 2015), to overcome this computational hurdle. We also contribute to this line of research by establishing the MCP estimator' convergence rate and regret bounds in the multi-armed bandit setting with non-i.i.d. samples and by developing an efficient 2sWL procedure for the MCP estimator in high-dimensional settings.

3. Problem Formulation

Following the settings of Auer (2002) and others, we present a standard bandit problem. Consider a sequential arrival process $t \in \{1, 2, ..., T\}$. At each time step t, a single user, prescribed by a vector of high-dimensional covariates, $\boldsymbol{x}_t \in \mathbb{R}^{1\times d}$, arrives. All covariates vectors $\{\boldsymbol{x}_t\}_{t\geq 0}$ are observable to a decision-maker and are i.i.d. distributed according to an *unknown* distribution $\mathcal{P}_x$. The decision-maker has access to a decision/arm set $\mathcal{K} = \{1, 2, ..., K\}$, and the reward for decision $i \in \mathcal{K}$ on a user with a covariates vector $\boldsymbol{x}$ is defined as:

$$R_i(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}_i + \epsilon, \quad (1)$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^{1\times d}$ is the unknown coefficient vector for decision $i \in \mathcal{K}$ and ϵ follows the sub-gaussian distribution. Following a standard assumption in the bandit literature (Rusmevichientong & Tsitsiklis 2010) to avoid trivial decisions, we assume that both covariates vector $\boldsymbol{x}$ and coefficient vector $\boldsymbol{\beta}_i$ are upper bounded so that the maximum regret at every time step will also be upper bounded.

In addition, the parameter vector $\boldsymbol{\beta}_i$ for $i \in \mathcal{K}$ is high-dimensional with latent sparse structures (i.e., the true value for the parameter vector $\boldsymbol{\beta}_i^{true}$ is sparse). We denote $S_i = \{j : \beta_{i,j}^{true} \neq 0\}$ as the index set for significant covariates (i.e., the covariates with non-zero coefficient parameters). This index set is also *unknown* to the decision-maker, and we define the maximum number of significant covariates for all arms as s (i.e., $s = \max_{i\in\mathcal{K}} |S_i|$), which is typically much smaller than the dimension of the covariates vector.

Note that the decision parameter vector $\boldsymbol{\beta}_i$ is unknown, but through a sequential online learning opportunities, the decision-maker could partially resolve the uncertainty and maximize its expected reward. We denote the decision-maker's policy as $\pi = \{\pi_t\}_{t\geq 0}$, where $\pi_t \in \mathcal{K}$ is the decision prescribed by policy π at time t. To benchmark the performance of policy π, we introduce an oracle policy $\pi^* = \{\pi_t^*\}_{t\geq 0}$ under which the decision-maker knows the true values of the covariates vector $\boldsymbol{\beta}_i^{true}$ for all $i \in \mathcal{K}$ and chooses the best decision to maximize its expected reward $\pi_t^* \doteq \arg\max_i \mathbb{E}_\epsilon[(R_i(\boldsymbol{x}_t))]$. Obviously, the decision-maker's reward is upper-bounded by the oracle policy. Accordingly, we define the decision-maker's expected regret at time t for the observed user covariates $\boldsymbol{x}_t$ under policy π as $r_t \doteq \mathbb{E}_\epsilon\left[\max_i R_i(\boldsymbol{x}_t) - R_{\pi_t}(\boldsymbol{x}_t)\right]$, which is the expected reward difference between the optimal oracle policy π^* and the decision-maker's policy π at time t. Our goal is to explore the policy π that minimizes the cumulative regret up to time T, $R_T \doteq \sum_{i=1}^T r_t$.

To analyze the regret, we present two technical assumptions.

Assumption 1 There exists a $C_0 > 0$ such that for $i \neq j \in \mathcal{K}$, $\mathbb{P}_{\boldsymbol{x}}\{|\boldsymbol{x}^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)| \in (0, \kappa]\} \leq C_0\kappa$ for $\kappa > 0$.

The first assumption is often referred to as the Margin Condition and is first introduced in the classification literature by Tsybakov et al. (2004). Goldenshluger et al. (2013) and Bastani & Bayati (2015) adopt this assumption to the linear bandit model. The Margin Condition ensures only a fraction of covariates can be drawn near the boundary hyperplane $\boldsymbol{x}^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) = 0$ in which rewards for both decisions are nearly equal. Clearly, if a large proportion of covariates are drawn from the vicinity of the boundary hyperplane, then for any bandit algorithm, a small estimation error in the decision parameter vectors ($\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$) will lead decision-makers to choose the wrong decision and perform poorly.

Assumption 2 There exists a partition $\mathcal{K}_o$ and $\mathcal{K}_s$ for $\mathcal{K}$. For $i \in \mathcal{K}_s$, we will have $\boldsymbol{x}^T\boldsymbol{\beta}_i + h < \max_{j\neq i} \boldsymbol{x}^T\boldsymbol{\beta}_j$ for some $h > 0$. For $i \in \mathcal{K}_o$, we define $U_i \doteq \left\{\boldsymbol{x} \,\middle|\, \boldsymbol{x}^T\boldsymbol{\beta}_i > \max_{j\neq i} \boldsymbol{x}^T\boldsymbol{\beta}_j + h\right\}$. There exist p^* such that $\min_{i\in\mathcal{K}_o} \mathbb{P}\{\boldsymbol{x} \in U_i\} \geq p^*$ for $p^* > 0$.

The second assumption is the Arm Optimality Condition (Goldenshluger et al. 2013; Bastani & Bayati 2015) and ensures that the decision parameter vectors for optimal decisions can be eventually learned, as the sample size increases. In particular, this Arm Optimality Condition separates decisions to an optimal subset (denoted by $\mathcal{K}_o$) and a suboptimal subset ($\mathcal{K}_s$): Decision i in $\mathcal{K}_o$ must be strictly optimal for some users' covariates vectors (denoted by set U_i); otherwise, decision j in $\mathcal{K}_s$ must be strictly suboptimal for all users' covariates vectors. Therefore, even if there is a small estimation error for decision i in $\mathcal{K}_o$, decision-makers are more likely to choose decision i for a user with a covariates vector draw from the set U_i. Accordingly, as sample size T increases, decision-makers could improve their estimations for decision parameter vectors for optimal arms.

These two assumptions are directly adopted from the multi-armed bandit literature and have been shown to be satisfied for all discrete distributions with finite support and a very large class of continuous distributions (see Bastani & Bayati 2015 for detailed examples and discussions).

4. MCP-Bandit Algorithm

In the big data era, one of the major challenges for online learning is the high dimensionality coupled with a limited sample size. The Lasso estimator is proposed to tackle this hurdle. Yet, the Lasso estimator could perform suboptimally due to the bias introduced by its penalty function, especially when the magnitude of true parameters is not small. One way to address this issue is to construct a new penalty function that renders an unbiased estimator and improves the sparse structure discovery. To this end, we will adopt the novel MCP approach to achieve this goal.

4.1. MCP Estimation

Consider that the true parameter vector $\boldsymbol{\beta}^{true}$ is sparse with a significant covariates index set $S = \{j : \beta_j^{true} \neq 0\}$, then the oracle estimator, under which the decision-maker has perfect knowledge of the index set S, can be presented by setting $\beta_j = 0$ for $j \in S^c$ and solving

$$\hat{\boldsymbol{\beta}}_O(\boldsymbol{X}, \boldsymbol{y}) \doteq \arg\min_{\beta \in S} \{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2\}. \tag{2}$$

It is worth noting that under the oracle estimator, the decision-maker can directly ignore the irrelevant covariates (by forcing their corresponding coefficients to be zero) and essentially reduce the high-dimensional problem to a low-dimensional estimation problem. Therefore, the oracle estimator is the best estimator we can attain from the data, and we will have the following result:

Lemma 1 Let s be the cardinality of S and $\boldsymbol{X}_S$ be the significant covariates matrix. If $n \geq s$, the estimator from (2) will satisfies the following inequality with probability $1 - \exp(-O(n))$:

$$\|\hat{\boldsymbol{\beta}}_O(\boldsymbol{X}, \boldsymbol{y}) - \boldsymbol{\beta}^{true}\|_2 \leq \sqrt{2\sigma^2 \cdot \frac{\text{eig}_{\max}(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S)}{\text{eig}_{\min}(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S)^2}}\sqrt{\frac{s}{n}},$$

where $\text{eig}_{\max}(\cdot)$ and $\text{eig}_{\min}(\cdot)$ denote the maximum and minimum eigenvalues.

[All proofs are in the supplement file] In practice, however, the significant covariates index set S is typically unknown. Therefore, we introduce the MCP approach to learn and recover this latent sparse structure. Specifically, we define the MCP penalty function as $P_{\lambda,a}(x) \doteq \int_0^{|x|} \max(0, \lambda - \frac{1}{a}|t|)dt$, where a and λ are positive parameters defined by the decision-maker. Using this MCP penalty function, we can present the MCP estimator as follows:

$$\hat{\boldsymbol{\beta}}_M(\boldsymbol{X}, \boldsymbol{y}, \lambda) \doteq \arg\min_{\beta}\{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \sum_{j=1}^{d} P_{\lambda,a}(\beta_j)\} \tag{3}$$

Denote the index set for non-zero coefficients solutions in Equation (3) as $\mathcal{J} \doteq \{j : \hat{\beta}_j \neq 0\}$. If the absolute value of every non-zero element in the MCP estimator is greater than $a\lambda$, then $P_{\lambda,a}(\cdot)$ become constant parameters for all $j \in \mathcal{J}$. Therefore, we will have $P_{\lambda,a}(\beta_j) = \frac{1}{2}a\lambda^2$ for $j \in \mathcal{J}$; and $P_{\lambda,a}(\beta_j) = 0$, otherwise. In other words, solving the MCP estimator is equivalent to the following problem: $\arg\min_\beta \left\{\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2\right\}$, where $\beta_j = 0$ for $j \in \mathcal{J}^c$.

If $\mathcal{J} = S$, then we conclude that the MCP estimator converges to the oracle estimator. Solving the MCP problem could be challenging. Liu et al. (2016; 2017) have shown that it is an NP-complete problem to find the MCP estimator by globally solving Equation (3). In the next subsection, we propose a local linear approximation method (i.e., the 2sWL procedure) to tackle this computational challenge and demonstrate that the estimator solved by this procedure will match the oracle estimator with high probability.

4.2. 2-Step Weighted Lasso Procedure

Let $\boldsymbol{w} = \{w_j\}$ be positive weights, and we define a weighted Lasso estimator as follows:

$$\hat{\boldsymbol{\beta}}_W(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) \doteq \arg\min_{\beta}\{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \sum_{j=1}^{d} w_j|\beta_j|\}.$$

Then, the 2sWL procedure consists of the following two steps. First, we solve a standard Lasso problem where all positive weights are set to a given parameter λ_0. Second, we use the Lasso estimator obtained in the first step to update the weights vector $\boldsymbol{w}$ by taking the first-order derivatives of the MCP penalty function, and then applying this updated weight vector, we solve the weighted Lasso problem to obtain the MCP estimator. The procedures of 2sWL at time t can be described as follows:

2-Step Weighted Lasso (2sWL) Procedure:

```
    Require: input parameters $a$ and $\lambda_0$
    Step 1: solve a standard Lasso problem
        $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_W(\boldsymbol{X}, \boldsymbol{y}, \lambda_0)$;
    Step 2: update $w_j = \begin{cases} P'_{\lambda,a}(|\beta_{1,j}|) & \text{, for } \beta_{1,j} \neq 0 \\ \lambda_0 & \text{, for } \beta_{1,j} = 0 \end{cases}$
        and solve a weighted Lasso Problem
            $\hat{\boldsymbol{\beta}}_{2sWL}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}) = \hat{\boldsymbol{\beta}}_W(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w})$.
```

Note that it is equivalent to solve the Lasso problem twice in the 2sWL procedure; therefore, the worst-case computation complexity for 2sWL is in the same order as for the standard Lasso problem. In practice, we can initialize the second step procedure with a warm start from the first step of the Lasso procedure, which further reduces the computation time.

Next, we will show that the MCP estimator identified by the 2sWL procedure can recover the oracle estimator with high probability. To this end, we will need the standard Compatibility Condition for Lasso estimator (Bühlmann & Van De Geer 2011), where we denote ϕ as the compatibility constant, to handle high dimensional data with sparse structure. The Compatibility Condition for Lasso estimators is analogous to the standard positive-definite assumption for the OLS estimator but less restrictive (e.g., the Compatibility Condition allows collinearity in the covariates matrix).

Proposition 1 Let $\min\{|\beta_j^{true}| : \beta_j^{true} \neq 0, j = 1, 2, .., d\} > (4\frac{s}{\phi^2} + a)\lambda$, then the following MCP inequality holds with probability $\exp(-O(n) + O(\log d))$:

$$\|\hat{\boldsymbol{\beta}}_{2sWL}(\boldsymbol{X}, \boldsymbol{y}, \lambda) - \boldsymbol{\beta}^{true}\|_2 > \sqrt{\frac{2\sigma^2 \mathrm{eig}_{\max}(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S)}{\mathrm{eig}_{\min}(\frac{1}{n}\boldsymbol{X}_S^T\boldsymbol{X}_S)^2}}\sqrt{\frac{s}{n}}.$$

We can further show that as the sample size increases, the MCP estimator converges to the true parameter at the optimal convergence rate (Wang et al. 2014).

Proposition 2 Set $\lambda = O(\sqrt{\log d/n})$. If the sample size exceeds a certain threshold (i.e., $n \gtrsim O(s^2 \cdot \log d)$), the convergence rate for the MCP estimator under the 2swL procedure satisfies $\|\hat{\boldsymbol{\beta}}_{2sWL}(\boldsymbol{X}, \boldsymbol{y}, \lambda) - \boldsymbol{\beta}^{true}\|_1 = O(s\sqrt{1/n})$, which matches the optimal convergence rate and is faster than that of the Lasso estimator (e.g.,$O(s\sqrt{\log d/n})$).

Together with the fact that the MCP estimator is an unbiased estimator, we believe that adopting the MCP estimator in multi-armed bandit model in high-dimensional settings could improve the decision-maker's reward and curb its expected cumulative regret.

4.3. MCP-Bandit Algorithm

After establishing the MCP estimator's statistical property, we are ready to present our proposed the MCP-Bandit algorithm. The proposed algorithm combines forced samples from all decisions at a pre-determined time sequence during which the decision-maker myopically selects a prescribed decision. In particular, we follow the forced sampling sequence developed by Goldenshluger et al. (2013) for the two-arms setting and by Bastani & Bayati (2015) for the multi-arms setting:

Forced sampling sequences: For a given positive integer $q \in \mathbb{Z}^+$, designed by the decision-maker, we define a sequence of forced samples for decision i as follows: $\mathcal{T}_i \doteq \{(2^n - 1) \cdot Kq + j | n \in \{0, 1, 2, ...\}$ and $j \in \{q(i-1)+1, q(i-1)+2, ..., iq\}\}$. At each prescribed time $t \in \mathcal{T}_i$, the decision-maker will myopically select decision i. Up to time t, we define the set of forced sampling sequences for decision i as $\mathcal{T}_{i,t}$, where the cardinality of this forced sample set is at least $\lfloor Kq \log t \rfloor$. We further denote the MCP estimator based on the forced sampling sequence $\mathcal{T}_i$ as $\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t-1}, \lambda_1)$.

All-sample sequences: In addition to myopically select decision i according to the prescribed forced sampling sequences, the decision-maker could choose decision i by comparing rewards among all decisions and pick decision i to maximize its reward. We denote $\mathcal{S}_{i,t}$ as the set of times decision i is selected, $\mathcal{S}_{i,t} = \{t' | \pi_{t'} = i \text{ for } 1 \leq t' \leq t\}$, and use $\hat{\boldsymbol{\beta}}_M(\mathcal{S}_{i,t-1}, \lambda_{2,t})$ to represent the MCP estimator based on this all sample set. Clearly, the forced sampling set $\mathcal{T}_{i,t}$ is a subset of $\mathcal{S}_{i,t}$.

MCP-Bandit Algorithm

```
Require: input parameters $q, h, \lambda_1, \lambda_{2,0}$
    Initialize $\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,0}, \lambda_1)$ and $\hat{\boldsymbol{\beta}}_M(\mathcal{S}_{i,0}, \lambda_{2,0})$ for $i \in \mathcal{K}$
  for $t = 1, 2....$ do
    Observe $\boldsymbol{x}_t$
    If $t \in \mathcal{T}_i$ for $i = 1, 2, ..., K$
        Set $\pi_t$ to $i$
    Else
        Update $\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t-1}, \lambda_1)$ for $i \in \mathcal{K}$ with 2sWL
        $\hat{K} = \{i | \boldsymbol{x}_t^T \hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t-1}, \lambda_1) \geq$
                $\max_{j \in K}\{\boldsymbol{x}_t^T \hat{\boldsymbol{\beta}}_M(\mathcal{T}_{j,t-1}, \lambda_1)\} - h/2\}$
        Update $\hat{\boldsymbol{\beta}}_M(\mathcal{S}_{i,t-1}, \lambda_{2,t-1})$ for $i \in \hat{K}$ with 2sWL
        $\pi_t = \arg\max_{i \in \hat{K}} \left\{\boldsymbol{x}_t^T \hat{\boldsymbol{\beta}}_M(\mathcal{S}_{i,t-1}, \lambda_{2,t-1})\right\}$
    End If
    Set $\mathcal{S}_{\pi_t,t}$ to $\mathcal{S}_{\pi_t,t-1} \cup t$ and $\lambda_{2,t}$ to $\lambda_{2,0}\sqrt{\frac{\log t + \log d}{t}}$
    Play arm $\pi_t$ and observes $y_t$
  end for
```

The MCP-Bandit algorithm can be described and executed as follows. If the current time t is prescribed in the forced sample sequence $\mathcal{T}_i$, then the decision-maker will select decision i. Otherwise, the decision-maker will first estimate the MCP estimator based on the forced sampling sequence $\mathcal{T}_i$ before time t, $\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t-1}, \lambda_1)$, via the aforementioned 2sWL procedure, and then construct a decision subset $\hat{K}$, in which all decisions are within $h/2$ of the maximum possible reward. Note that any decision that is not in this subset $\hat{K}$ will be a suboptimal decision for the current user. Finally, the decision-maker uses all samples to re-estimate the MCP estimator $\hat{\boldsymbol{\beta}}_M(\mathcal{S}_{i,t-1}, \lambda_{2,t})$, based on which the decision-maker will compare the reward performance for all decisions in the subset $\hat{K}$ and select the decision that generates the highest expected reward.

The following Theorem is the main result of this paper and establishes the expected cumulative regret upper bound for the MCP-Bandit algorithm.

Theorem 1 When $q \geq q_0$, $K \geq 2$, $d \geq 1$, $T \geq t_0$, and we take $\lambda_1 = (\phi^2 p^* h)/(64 s x_{\max})$, $\lambda_{2,0} = \phi^2 x_{\max}\sqrt{\sigma^2(\log t + \log d)/((p^*)^3 t)}/2s$, and $\lambda_{\min} > 0$. The expected cumulative regret of the MCP-Bandit algorithm is upper-bounded at time T by

$$\begin{aligned} R_T &\leq 2(Kq)^2 b x_{\max} + 2C_1 qKb x_{\max}\log T \\ &+ 2Kb x_{\max}\log T \\ &+ (4Kb x_{\max}(C_2+1) + \frac{C_0 s^2\sigma^2\lambda_{\max}}{2\lambda_{\min}^2}p^*)\log T \\ &= O(s^2(s+\log d)\log T), \end{aligned} \qquad (4)$$

where $x_{\max}$ and b are upper bounds for covariate $\boldsymbol{X}$ and parameters $\boldsymbol{\beta}$, C_1, and C_2 are positive constants independent on T and d, $q_0 \gtrsim O(s^2 \log d)$, and $t_0 \gtrsim O((Kq)^2)$.

Under a low-dimensional multi-armed bandit model setting, Goldenshluger et al. (2013) show that the lower-bound on the expected cumulative regret is $O(\log T)$, which is also applicable (i.e., is the lower-bound) to our high-dimensional setting. Theorem 1 demonstrates that the maximal expected cumulative regret of the MCP-Bandit algorithm over T users is upper-bounded by $O(\log T)$. Therefore, the MCP-Bandit algorithm achieves the optimal expected cumulative regret in sample size T. This result comes from the fact that we can ensure $O(\log T)$ forced samples at time T, and therefore the MCP estimator will match the oracle estimator with high probability, which leads to the $\log T$ dependence. In addition, when compared to the Lasso-Bandit algorithm proposed by Bastani & Bayati (2015) in high-dimensional settings, the MCP-Bandit algorithm reduces the regret upper-bound from $O(\log^2 T)$ to $O(\log T)$.

Theorem 1 also shows that the expected cumulative regret of the MCP-Bandit algorithm in the covariates dimension d is upper-bounded by $O(\log d)$, which is also a tighter bound than that of the Lasso-Bandit algorithm $O(\log^2 d)$ and other classic bandit algorithms (e.g., OLS bandit in Goldenshluger et al. 2013 or OFUL in Abbasi-Yadkori et al. 2011), which typically yield polynomial dependence in d.

5. Key Steps of Regret Analysis for the MCP-Bandit Algorithm

In this session, we will brief key steps in establishing the expected cumulative regret upper-bound for the MCP-Bandit algorithm in Theorem 1. We first highlight the influence of the non-i.i.d. data, inherited from the multi-armed Bandit model, on statistical convergence properties of the MCP estimator and further prove a general oracle MCP inequality for these non-i.i.d. data. Then, we will apply this result to establish the convergence properties for both forced-sample and all-sample estimators and to provide the corresponding cumulative regret. Finally, the total expected cumulative regret can be established by adding up the regret in these estimators. The main structure and sequence of our proving steps described above are first introduced by Bastani & Bayati (2015), who present their analysis for the expected regret for the LASSO-Bandit algorithm in this sequence. We will follow their presentation structure, but with different proving techniques and convergence properties, to illustrate the key steps in analyzing the MCP-Bandit algorithm.

5.1. Oracle Inequality for Non-i.i.d. Data

We first show a general result for the MCP estimator under non-i.i.d. data. Consider a linear model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}^{n\times d}$ is the design matrix, $\boldsymbol{y}^{n\times 1}$ is the response vector and $\boldsymbol{\epsilon}^{n\times 1}$ is the i.i.d $\sigma-$subgaussians. Denote $\mathcal{A}$ as the index set for a sub-sample in $\boldsymbol{X}$ and $\boldsymbol{y}$. The MCP estimator for this linear model is $\hat{\boldsymbol{\beta}}_M(\mathcal{A},\lambda) \doteq \arg\min\left\{\frac{1}{|\mathcal{A}|}\|\boldsymbol{X}^{\mathcal{A}}\boldsymbol{\beta} - \boldsymbol{y}^{\mathcal{A}}\|_2^2 + P_\lambda(\boldsymbol{\beta})\right\}$, where $\boldsymbol{X}^{\mathcal{A}}$ is the covariate matrix with sample indexed by $\mathcal{A}$. Note that if samples in $\mathcal{A}$ are not i.i.d, then standard MCP convergence results (Fan et al., 2014; 2015) can not be directly applied. Yet, as there are samples generated via the forced sample sequences (e.g., $\mathcal{T}_i$ for $i = 1,...,K$), there must exists a subset $\mathcal{A}^{'} \subseteq \mathcal{A}$ such that all samples in this subset are i.i.d from distribution $\mathcal{P}_{\boldsymbol{X}}$, that is, $\{X_t | t \in \mathcal{A}^{'}\} \sim \mathcal{P}_{\boldsymbol{X}} \times ...\mathcal{P}_{\boldsymbol{X}}$. The next step is to show that when the cardinality of $\mathcal{A}^{'}$ (i.e., $|\mathcal{A}^{'}|$) is large enough, $\hat{\boldsymbol{\beta}}_M(\mathcal{A},\lambda)$ will not be far away from an all i.i.d. sample estimator and converge to the true parameter. We formally summarize this result in Proposition 3:

Proposition 3 If $|\mathcal{A}^{'}|/|\mathcal{A}| \geq c_0/2 > 0$, $\|\boldsymbol{\beta}^{true}_{\min}\|_{\min} > (4s/\phi^2 + a)\lambda$, $|\mathcal{A}| \geq \frac{1024 s z^3_{\max}\log d}{c_0\phi^2}$, $\mathrm{eig}_{\min}\mathbb{E}[\boldsymbol{x}_S^T\boldsymbol{x}_S] = \lambda_{\min} > 0$ and $\mathrm{eig}_{\max}(\frac{1}{|\mathcal{A}|}(\boldsymbol{X}_S^{\mathcal{A}})^T\boldsymbol{X}_S^{\mathcal{A}}) \leq \lambda_{\max}$, then the oracle inequality $\|\hat{\boldsymbol{\beta}}_M(\mathcal{A},\lambda) - \boldsymbol{\beta}^{true}\|_2 > \sqrt{\frac{\lambda_{\max}\sigma^2}{2\lambda^2_{\min}}}\sqrt{\frac{s}{|\mathcal{A}|}}$ holds for $t > 0$ with probability $\exp(-O(|\mathcal{A}|) + O(\log d))$, where $\boldsymbol{X}_S^{\mathcal{A}}$ is the significant covariates matrix with sample indexed by $\mathcal{A}$.

5.2. Oracle Inequality for Forced-sample Estimator

In the forced sample set $\mathcal{T}_{i,t}$, each sample is drawn i.i.d from the whole population. Denote $\mathcal{T}^{'}_{i,t} = \mathcal{T}_{i,t} \cup U_i$, where U_i is the set that decision i is the optimal choice. First, we need to show that up to time t, $|\mathcal{T}^{'}_{i,t}|$ and $|\mathcal{T}_{i,t}|$ are not too small with high probability. By the design of the forced sampling sequence, we will have $|\mathcal{T}_{i,t}| \geq q_0 \log t$, $q \geq \lceil 4q_0 \rceil$, and $t \geq (Kq)^2$. If we define an indicator $z_{i,t}$ to indicate whether $\boldsymbol{x}_{i,t} \in \mathcal{T}^{'}_{i,t}$, then $z_{i,t}$ will be i.i.d Bernoulli random variable and $E[z_{i,t}] \geq p^*$. Thus $|\mathcal{T}^{'}|$ follows Binomial$(|\mathcal{T}_{i,t}|, E[z_{i,t}])$, from the Chernoff bound for Binomial random variable: $\mathbb{P}\left\{\frac{|\mathcal{T}^{'}_{i,t}|}{|\mathcal{T}_{i,t}|} \geq \frac{p_i^*}{2}\right\} \geq 1 - \frac{1}{t}$. To apply the results in Proposition 3, we need to show that $|\mathcal{T}^{'}_{i,t}| \geq \frac{2^{11} s x^3_{\max}\log d}{c_0\phi^2} = \frac{2^{12} s x^3_{\max}\log d}{p_i^*\phi^2}$. As $|\mathcal{T}^{'}_{i,t}| \geq q_0 \log t$ holds, if $q_0 \geq \frac{2^{12} s x^3_{\max}\log d}{p_i^*\phi^2}$ and $\log t \geq 1$, then the sample size requirement will be satisfied. If we set $q_0 \geq \frac{8\sigma^2\lambda_{\max}s^2x^2_{\max}}{\lambda^2_{\min}h^2}$, for $\log t \geq 1$ the following inequality will hold with probability $\exp(-O(\log d \log t))$:

$$\|\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t},\lambda) - \boldsymbol{\beta}^{true}\|_1 > \frac{h}{4x_{\max}}. \tag{5}$$

5.3. Oracle inequality for All-sample estimator

Next, to show the oracle inequality for all-sample MCP estimator, we will start with the following Proposition (i.e. Proposition 3 in Bastani & Bayati (2015)) to establish the oracle inequality for the Lasso estimator:

Proposition 4 (Proposition 3 of Bastani & Bayati (2015)) The all-sample estimator satisfies the oracle inequality, $\|\hat{\boldsymbol{\beta}}_L(\mathcal{S}_{i,t}) - \boldsymbol{\beta}^{true}\|_1 > 16x_{\max}\sqrt{\frac{\log t + \log d}{(p_i^*)^3 c_1 t}}$, with probability $\frac{1}{t} + 2\exp\left(-\frac{(p_i^*)^2(c_2 \wedge \frac{1}{2})}{16}t\right)$, where c_1 and c_2 are positive constants.

Let $\lambda = \frac{\phi^2 x_{\max}}{2s_0}\sqrt{\frac{\log t + \log d}{(p^*)^3 c_1 t}}$. Then, for a given minimum signal strength $\|\boldsymbol{\beta}^{true}\|_{\min} > (\frac{32s}{\phi^2} + a)\lambda$, Proposition 4 directly suggests that $\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^{true}\|_\infty \leq \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^{true}\|_1 \leq \frac{32s}{\phi^2}\lambda$.

Therefore, the first-order derivatives for the MCP penalty function with respect to the parameters of significant covariates will be zero. On the other hand, for non-significant covariates, the first-order derivatives for the MCP penalty function will be large numbers:

$$\begin{array}{ll} \partial P_\lambda(\hat{\boldsymbol{\beta}}_S) = 0 & \text{, if } |\hat{\boldsymbol{\beta}}_S|_{\min} > a\lambda \\ \partial P_\lambda(\hat{\boldsymbol{\beta}}_{S^c}) \geq \partial P_\lambda(\frac{32s}{\phi^2}\lambda) & \text{, if } |\hat{\boldsymbol{\beta}}_{S^c}|_{\max} \leq \frac{32s}{\phi^2}\lambda \end{array}$$

Accordingly, we use these first-order derivatives as the weights to the weighted Lasso problem, and its oracle inequality can be stated as in the following proposition.

Proposition 5 When $t \geq (Kq)^2$ and $q \gtrsim O(s^2 \log d)$, the all sample estimator $\hat{\boldsymbol{\beta}}_{2sWL}(\mathcal{S}_{i,t}, \lambda)$, where $\lambda = \frac{\phi^2 x_{\max}}{2s_0}\sqrt{\frac{\log t + \log d}{(p^*)^3 c_1 t}}$, is an oracle solution and satisfies the oracle inequality: $\mathbb{P}\left\{\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}^{true}\|_2 \geq \sqrt{\frac{\sigma^2 \lambda_{\max} s}{2\lambda_{\min}^2 p^* t}}\right\} \leq \zeta$, where $\zeta = \exp(-O(t) + O(\log d)) + O(1/t)$.

5.4. Bounding the Cumulative Expected Regret

We now bound the cumulative regret for the MCP-Bandit algorithm by dividing our time periods $[T]$ into three groups and providing a upper bound for each group.

The first group contains all samples with $t \leq (Kq)^2$ and all forced samples. When $t \leq (Kq)^2$, we do not have sufficient samples to accurately estimate covariates parameter vectors, the decision performance under the MCP-Bandit algorithm will be sub-optimal comparing to that of the oracle case. Combined with the fact that forced sample size up to time T is on the order of $O(\log T)$, we can bound the cumulative regret by their worst case performance: $2(Kq)^2 b x_{\max} + 2C_1 \log T K q b x_{\max}$. Next, we segment the $t > (Kq)^2$ without forced samples case into two groups, depending on whether we can accurately estimate covariates parameter vectors by using only forced samples.

The second group includes scenarios where $t > (Kq)^2$ and forced sample based estimators are not accurate enough. In particular, we define $A_t \doteq \left\{\|\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t}, \lambda) - \boldsymbol{\beta}^{true}\|_1 \leq \frac{h}{4x_{\max}}\right\}$. When A_t doesn't hold, the forced sample based estimator vector $\hat{\boldsymbol{\beta}}_M(\mathcal{T}_{i,t}, \lambda)$ is not near the true parameter vector $\boldsymbol{\beta}^{true}$. Under those scenarios, our decisions will be sub-optimal with high probability. Note that the size of forced samples increases in t, so the probability of event A_t not occurring decreases in time t. Through Equation (5), we can bound the cumulative regret for the second group by $2Kbx_{\max} \log T$.

The third group includes cases where $t > (Kq)^2$ and forced sample based estimators are accurate (i.e., event A_t holds). Note that under these scenarios, we can improve our estimation accuracy by using the all sample estimator ($O(\sqrt{1/t})$ in Proposition 5), instead of relying only on the forced sample estimator ($O(\sqrt{1/\log t})$ in Proposition 3). Benefiting from the improved estimation accuracy, we can bound the cumulative regret for the third group by $(4Kbx_{\max} + \frac{C_0 s^2 \sigma^2 \lambda_{\max}}{2\lambda_{\min}^2} p^*) \log T$.

6. Experiments

We benchmark the MCP-Bandit algorithm to two bandit algorithms that are not specifically designed for high-dimensional settings (i.e., OLS-Bandit by Goldenshluger et al. 2013 and OFUL by Abbasi-Yadkori et al. 2011) and one bandit algorithm that is developed for high-dimensional problems (i.e., Lasso-Bandit by Bastani & Bayati 2015).

6.1. Synthetic Data

In the synthetic data experiment, we present a two-arm bandit setting with decision parameter $\boldsymbol{\beta}_i,\ i = 1, 2$. To simulate different sparsity level, we generate four possible covariates dimensions, $d = 10$, 10^2, 10^3, and 10^4, and keep the dimension for significant covariates unchanged $s = 5$. Other parameter combinations exhibit similar pattern and observations, and therefore omitted. In addition, we share the same parameter λ in both the Lasso-Bandit algorithm and the MCP-Bandit algorithm and select the unique parameter for the MCP-Bandit algorithm a at 2. We arbitrarily set the coefficients for significant covariates for the first arm to be $\boldsymbol{\beta}_1 = (1, 2, 3, 4, 5)$ and for the second arm to be $\boldsymbol{\beta}_2 = 1.1 \cdot \boldsymbol{\beta}_1$. The covariates are generated from $N(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$ and the random error ϵ follows $N(0, 1)$. For each covariates dimension, we generate an average of $10,000$ trials. Figure 1 shows the influence of the covariates dimension d and the sample size T on the cumulative regret for OFUL, OLS-Bandit, Lasso-Bandit, and MCP-Bandit algorithms.

We observe that the MCP-Bandit algorithm outperforms all other three benchmarks and has the lowest cumulative regret. The cumulative regret for all four algorithms increases in the covariates dimension d, but at different rate (see the left-hand-side of Figure 1). Comparing to OLS-Bandit (Goldenshluger et al., 2013) and OFUL algorithm in (Abbasi-Yadkori et al., 2011), Lasso-Bandit (Bastani & Bayati, 2015) and MCP-Bandit algorithms, both of which are designed for high-dimensional problems, have lower cumulative regret that increases in d at a slower rate. Further, the benefits of adopting the MCP-Bandit algorithm seem to increase in d, which confirms our theoretical findings: The

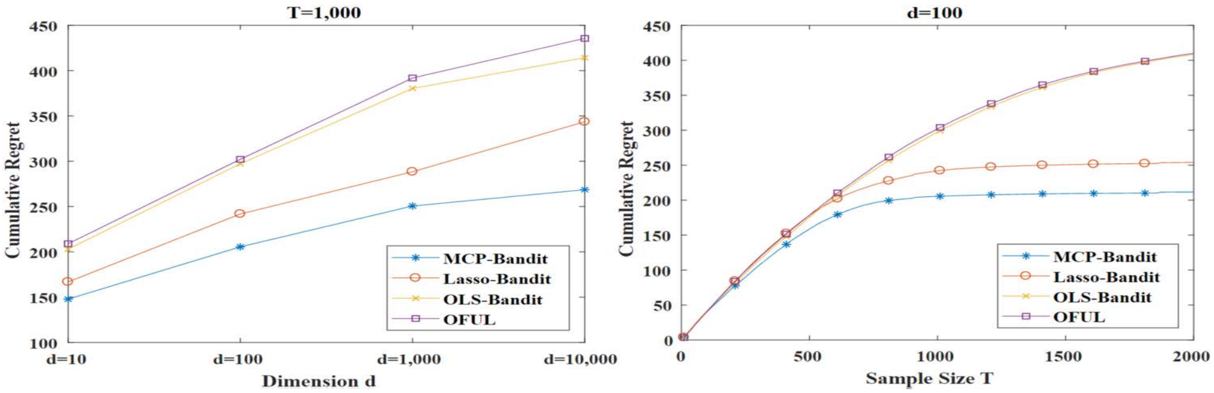

Figure 1. The influence of the covariates dimension d and the sample size T on the expected cumulative regret.

MCP-Bandit algorithm has a better dependence in d (e.g., $\log d$), than Lasso-Bandit (e.g., $\log^2 d$), OFUL, and OLS-Bandit (the last two algorithms have polynomial bounds in d).

The right-hand-side of Figure 1 reports the influence of sample size T on the cumulative regret. As MCP-Bandit provides the optimal time dependence under high-dimensional settings, MCP-bandit is guaranteed to strictly improve from Lasso-Bandit, especially when T is not too small. When there are insufficient samples, all algorithm fails to accurately learn covariates parameters vectors. As a result, all four algorithm perform equally poor under limited samples. As the sample size increases, the MCP-bandit algorithm immediately outperforms all other benchmarks. In the right-hand-side of Figure 1, the regret reduction of MCP-Bandit over Lasso-Bandit is significant ($> 1\%$) when T is larger than 35; the regret reduction improves in T and is stabilized around 16% after 175 samples. This observation also echoes our theoretical findings that the MCP-Bandit algorithm attains the optimal regret bound in sample size dimension ($O(\log T)$).

6.2. Warfarin Dosing Patient Data

The second experiment considers a health-care decision-making process in which physicians determine the optimal warfarin dosage for every incoming patient. The warfarin dosing patient data (Consortium et al. 2009), which is known to be dense (e.g., $\log T$ is not necessarily larger than s), contains approximately 100 detailed covariates for $5,700$ patients. Under this dataset, Bastani & Bayati (2015) demonstrate that the Lasso-Bandit algorithm outperforms many existing bandit algorithms, including OFUL_LS (Abbasi-Yadkori et al. 2011), OFUL-EG (Abbasi-Yadkori et al. 2012), and OLS-Bandit (Goldenshluger et al. 2013).

We apply the MCP-Bandit algorithm to the same warfarin dosing patient data to evaluate its performance in practical decision-making contexts where technique assumptions specified early may not hold. Figure 2 compares the average percentage of correct dosing decisions under the MCP-Bandit algorithm to those under the oracle case, OLS-Bandit, Lasso-Bandit, OFUL, and actual physicians' decisions. We observe that when the number of patients is not too small (i.e., great than 370 patients), the MCP-Bandit algorithm always outperforms all other benchmarks (e.g., the regret reduction ranges from 0% to 22.1%).

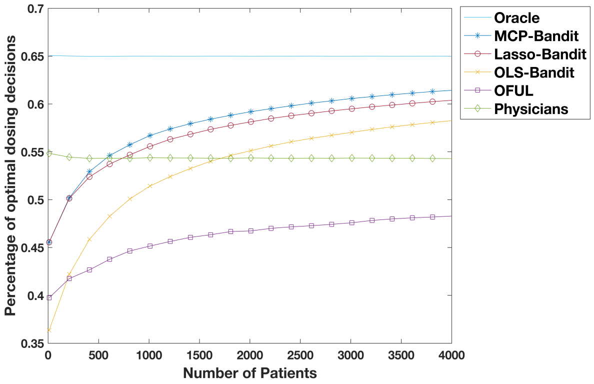

Figure 2. The percentage of optimal warfarin dosing decisions.

7. Conclusion

In this paper, we propose the MCP-Bandit algorithm for online learning and decision-making processes in high-dimensional data settings. To overcome the computational and statistical challenges associated with solving the MCP estimator under non-i.i.d. samples, we propose the 2sWL procedure and show that the MCP estimator solved by the 2sWL procedure matches the oracle estimator with high probability. We demonstrate that the cumulative regret of the MCP-Bandit algorithm over sample size T is bounded by $O(\log T)$, which is lowest theoretical bound for all possible algorithms. On the covariates dimension d and the number of significant covariates dimension s, the cumulative regret of the MCP-Bandit algorithm is bounded by $O(s^2(s + \log d))$, which is also a tighter bound than the other existing bandit algorithms. We show that the MCP-Bandit algorithm performs favorably in all our experiments, especially when the data sparsity level is high or when the sample size is not too small.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, volume 22, pp. 1–9, 2012.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pp. 127–135, 2013.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Bastani, H. and Bayati, M. Online decision-making with high-dimensional covariates. 2015.

Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pp. 208–214, 2011.

Consortium, I. W. P. et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*, 2009(360):753–764, 2009.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366, 2008.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Fan, J., Xue, L., and Zou, H. Strong oracle optimality of folded concave penalized estimation. *Annals of statistics*, 42(3):819, 2014.

Fan, J., Liu, H., Sun, Q., and Zhang, T. Tac for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *arXiv preprint arXiv:1507.01037*, 2015.

Goldenshluger, A., Zeevi, A., et al. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Liu, H., Yao, T., Li, R., and Ye, Y. Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical Programming*, pp. 134. doi: 10.1007/s10107-017-1114-y.

Liu, H., Yao, T., Li, R., et al. Global solutions to folded concave penalized nonconvex learning. *The Annals of Statistics*, 44(2):629–659, 2016.

Liu, H., Yao, T., Li, R., and Ye, Y. Folded concave penalized sparse linear regression: Sparsity, statistical performance, and algorithmic theory for local solutions. *Mathematical programming*, 166(1-2):207–240, 2017.

Rigollet, P. and Zeevi, A. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tsybakov, A. B. et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1): 135–166, 2004.

Wang, Z., Liu, H., and Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Yang, Y., Zhu, D., et al. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

Zhang, C.-H. et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2): 894–942, 2010.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.