ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447590

NON-STATIONARY BANDITS WITH PERIODIC BEHAVIOR: HARNESSING RAMANUJAN PERIODICITY TRANSFORMS TO CONQUER TIME-VARYING CHALLENGES

Parth Thaker, Vineet Gattani*, Vignesh Tirukkonda*, Pouria Saidi, Gautam Dasarathy*
ECEE, Arizona State University

ABSTRACT

In traditional multi-armed bandits (MAB), a standard assumption is that the mean rewards are constant across each arm, a simplification that can be restrictive in nature. In many real-world settings, the rewards exhibit a periodic pattern on which traditional MAB algorithms would fail. This paper addresses the problem of regret minimization when the mean rewards change periodically. To this end, we propose an approach that utilizes the Ramanujan periodicity transform to estimate the support of the periods efficiently and, furthermore, use this information to minimize regret.

***Index Terms*—** Non-stationary Bandits, Periodic Bandits, Ramanujan Transform, Regret minimization

1. INTRODUCTION

Sequential decision-making under uncertainty is crucial in a wide variety of fields. Ideally, given ample time, one would exhaustively sample all available options before making decisions. However, in modern problems which present the decision maker with an enormous number of choices, such an approach is infeasible. The *Multi-armed bandits* (MAB) framework [1] addresses this by efficiently identifying optimal options in minimal time. Central to MABs is the Exploration-exploitation dilemma: one must balance exploring unknown choices and exploiting the best-known option. Given its strong theoretical foundations and its efficacy in a wide range of domains like recommendation systems, clinical trials, and online advertising, this framework and its variants have received much attention in recent years [1–3].

A key limitation of this framework, however, is its traditional reliance on stationarity of the underlying "reward" distribution. Real-world applications often exhibit non-stationarity. Introducing non-stationary reward distributions complicates matters due to potential erratic patterns. Although there have been endeavors to address this (see, e.g., [4–6]), formulating a universal learning policy for non-stationarity remains challenging.

* Equal contributions. This work was supported by the National Science Foundation under award number CCF-2048223, and the Office of Naval Research (ONR) under award number N00014-21-1-2615.

In this paper, we focus on *Periodic Bandits*, a class of non-stationary bandits that are characterized by a periodic pattern in their rewards. Such periodicity is common in a range of real-world scenarios, such as cell-tower congestion, advertisement trends, and behavior of electronic systems reliant on discharging power sources. Ignoring these patterns can result in highly suboptimal decisions [7]. Incorporating periodicity into multi-armed bandit algorithms enables one to make decisions that align more closely with the natural rhythms and temporal variations present in the problem domain.

Research such as [8] has addressed seasonal reward shifts, while [9] leverages historical data for sudden changes. Other studies, like [10], focus on regime-switching rewards, while [11] considers rewards based on auto-regressive models. [12] integrates periodicity in Gaussian process bandits. Our work aligns most closely with [13], which combines Fourier analysis with a confidence-bound-based learning procedure to learn the periods and minimize the regret.

This paper proposes a tractable methodology for tacking the periodic bandit framework. To this end, we utilize the framework of Ramanujan Periodicity Transforms (RPT) to estimate the length of the period and identify the fundamental periods if the signal is a combination of two or more periodic signals. The authors in [14, 15] introduced the notion of RPT and showed that one can utilize RPTs to estimate the underlying period of a periodic signal. In addition, the authors demonstrated that RPT-based methods are more robust in the presence of noise and showed the advantages of RPTs over the classical DFT-based techniques [14]. RPTs have been used in practice such as detecting periodicity in visually evoked potentials in brain-computer interfaces [16] and detecting the tandem DNA repeats [17] and have shown promising results.

Contributions. The main contributions of this work are the following.
a) We propose an online learning algorithm called Bandit Tracking System via Ramanujan Periodic transform (BTS-RaP) for non-stationary environments with seasonal patterns and unknown periods.
b) We propose the use of RPT dictionaries to estimate length of periods across different arms which are known to overcome the limitations of DFT-based technique.
c) Using computer simulations we show that BTS-RaP algorithm can achieve sublinear regret.

2. RAMANUJAN PERIODICITY TRANSFORMS

In this section, we briefly review the structure of the RPT dictionary, and their applicability to estimate the period of a periodic signal.

2.1. RPT dictionaries

RPT dictionaries are constructed based on the properties of Ramanujan sums, defined as [18]

$$c_p(n) = \sum_{\substack{k=1 \\ (k,p)=1}}^{p} \exp(j2\pi kn/p), \tag{1}$$

where (k, p) is the greatest common divisor of k and p. $\mathbf{c}_p$ indicates the vector form of $c_p(n)$, and $\mathbf{c}_p^{(i)}$ shows the circularly shifted version of $\mathbf{c}_p$ with step size i. For each value p construct a $p \times \phi(p)$ submatrix $\mathbf{C}_p$ as follows

$$\mathbf{C}_p = \left[\mathbf{c}_p \quad \mathbf{c}_p^{(1)} \quad \dots \quad \mathbf{c}_p^{(\phi(p)-1)}\right], \tag{2}$$

where $\phi(p)$ is the Euler totient function (the number of integers that are co-prime to p). The author in [15] showed that the $\phi(p)$ columns in $\mathbf{C}_p$ are linearly independent. Thus, one can construct the RPT dictionary $\mathbf{K}$ in three consecutive steps.

i) Build all the submatrices $\mathbf{C}_p$ for every $p \in \mathbb{P}$, where $\mathbb{P} = \{1, 2, \dots, P_{\max}\}$ and $P_{\max}$ is the largest possible period in the signal.

ii) Build the $L \times \phi(p)$ submatrices $\mathbf{R}_p$, by periodically extending all the columns of $\mathbf{C}_p$ to length L.

iii) Concatenate the matrices $\mathbf{R}_p$ as

$$\mathbf{K} = \left[\mathbf{R}_1 \quad \mathbf{R}_2 \quad \dots \quad \mathbf{R}_{P_{\max}}\right]. \tag{3}$$

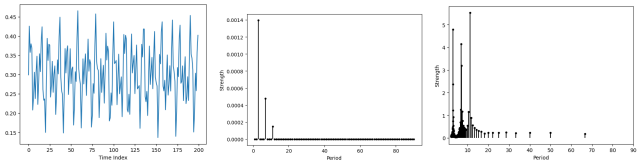
Therefore, denoting $\phi(P_{\max}) = \sum_{p=1}^{P_{\max}} \phi(p)$, the size of the dictionary is $L \times \phi(P_{\max})$.

2.2. Period estimation using RPT dictionary

Discrete periodic signals can be expressed using the RPT dictionary in a noise-free setup as:

$$\mathbf{y} = \mathbf{K}\mathbf{x} \tag{4}$$

where $\mathbf{y}$ is the vector form of the periodic signal with period p, $\mathbf{K}$ is the RPT dictionary introduced in section 2, and $\mathbf{x}$ is the sparse representation of the periodic signal under the RPT dictionary. Given a sufficiently long vector $\mathbf{y}$, vector $\mathbf{x}$ exhibits a sparse structure and its non-zero values correspond to the sub-matrices in $\mathbf{K}$, that have periodic columns with periods



(a) (b) (c)

Fig. 1. (a) A noisy period 231 time series signal with that was generated as sum of period 3, 7 and 11 signals. The strength vs period plot for the solutions of the convex problem (5) using (b) Ramanujan basis, and (c) DFT basis.

q_i that are divisors of p, or $q_i|p$. Therefore, it is possible to estimate the period of a periodic signal by first recovering the sparse representation of the signal under the RPT dictionary. Then, the support set of the signal identifies the divisors of the underlying period of the signal. The support set of a sparse vector are a set of indices that contain the location of the non-zero values of the vector. Finally, the estimate of the period is equal to the least common multiplier (LCM) of the divisors from the recovered support set. One can recover the support of the sparse vector $\mathbf{x}$ using sparse recovery algorithms [19]. In this work, we adopt the proposed approach in [14] and solve the following minimization program:

$$\min \|\mathbf{D}\mathbf{x}\|_2 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{K}\mathbf{x} \tag{5}$$

where, $\mathbf{D}$ is a diagonal penalty matrix with i-th entry on the diagonal being equal to p_i^2, where p_i is the period of the i-th column of the dictionary $\mathbf{K}$. An illustration of this method is presented in Fig. 1. In Part (a), we observe an incomplete segment of a signal with a period of 231, which has been affected by noise. This 231-period signal was constructed by combining three periodic signals with underlying periods of 3, 7, and 11. Following [14], we compute the energy corresponding to each subvector in $\mathbf{x}$ as follows and plot the strength vs. period. Part (b) illustrates the strength vs. period after solving (5). The strength at each period p is defined as:

$$E(p) = \sum_{k=P+1}^{P+\phi(p)} |x(k)|^2, \quad P = \sum_{d=1}^{p-1} \phi(d). \tag{6}$$

Similarly, in part (c) the periodogram displays the strength of the different period components in the signal. It is evident that, RPT basis is robust towards estimating the fundamental periods of a given signal.

3. PROBLEM SETUP

Consider a multi-armed bandit setting with $\mathcal{K}$ being the set of all arms such that mean of each arm $i \in \mathcal{K}$ is represented

by function $\mu_i : \mathbb{N} \to [a, b] \quad \forall i \in [K]$ such that $\mu_i[t + T_i] = \mu_i[t]$ for some unknown $T_i \in \mathbb{N}$. Throughout the paper, we sometimes refer $\mu_i[t]$ as $\mu_{t,i}$. At each round, the learner chooses an arm $a_t \in \mathcal{K}$ to sample and observes a noisy reward

$$r_{t,i} = \mu_{t,i} + \eta_{t,i}\,,$$

where, $\{\eta_{i,t}\}_{i,t}$ are i.i.d. noise samples from a σ^2- sub-Gaussian distribution. The goal of the problem is to minimize the regret up to a known time horizon T defined as,

$$\mathcal{R}(T) = \sum_{t=0}^{T} \left(\max_{i \in \mathcal{K}} \mu_{t,i} - \mu_{t,a_t} \right)\,, \tag{7}$$

where, the decision maker chooses an arm a_t at time step t. The aim of the work is to propose an algorithm to minimize regret as mentioned in (7). This can only be obtained if the decision maker chooses an arm that is optimal *for every time* t. If, at the current time instant, one of the arms is optimal, it is not necessary that the previously chosen arm will be optimal again at the next time step, which is well-suited for handling time-varying reward changes. The notion in (7) is different than the standard notion of regret [20], which focuses on selecting the one optimal choice for every time-step t.

3.1. Baseline method

Recently [13] proposed for addressing periodic bandits a two-stage approach which provides a sub-linear regret that scales as $\mathcal{O}\left(\sqrt{T \sum_{i=1}^{n} T_i}\right)$, where T_i is the period of arm i. The authors first propose to use (DFT) to estimate the length of the periods T_i's. Since the mean of arm i returns to the same value every T_i steps, the authors propose that for every arm i, the number of 'effective arms' is T_i (1 arm for every step until time reaches T_i). Therefore, we end up with $\hat{d} := \sum_k \hat{T}_k$ effective arms (unique mean rewards) to learn. In the second stage of the algorithm, the authors utilize the estimated number of effective arms to implement UCB-based approach to minimize regret (7). We refer this as MAB-UCB.

This approach suffers drawbacks due to the utilization of DFT as well as two distinct stages leading to sample inefficiency. We address this by using RPT and merging the two stages into one main algorithm thereby painting optimal sample efficiency.

4. PROPOSED APPROACH: BTS-RAP (BANDIT TRACKING SYSTEM)

We first provide an overview of the linear bandits and then show how it connects to RPT-based reward representation.

4.1. Linear bandits

Linear bandits [3] have emerged as a powerful and versatile tool in the field of bandit research literature. These algorithms are particularly well-suited for scenarios where the relationship between actions and rewards can be approximated linearly. Let the arm set be defined by the set $\mathcal{K}$. In a linear bandit setup, every arm is associated with a feature vector $\mathbb{R}^d$ such that $d < n$. On sampling the arm i at time t, the reward observed satisfies the relation $r_{t,i} = \langle \mathbf{a}_i, \boldsymbol{\theta}^* \rangle + \eta_t$, where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ dented the unknown reward parameter, $\mathbf{a}_i \in \mathbb{R}^d$ denote the feature vector associated with arm i and η_t is the i.i.d. Gaussian noise realized from a σ^2-subgaussian distribution at time t.

Due to the low dimensional structure of the linear bandit problem, it has been proven, both theoretically and experimentally that the regret upper bound scales as $\mathcal{O}(\sqrt{dT})$ (sublinear in time T), where d is the feature dimensionality. Note that for the case of stationary linear bandit, the regret takes the form as defined below,

$$\mathcal{R}_{LB}(T) = \sum_{t=0}^{T} \left(\max_{k \in \mathcal{K}} \langle \mathbf{a}_k, \boldsymbol{\theta}^* \rangle - \langle \mathbf{a}_t, \boldsymbol{\theta}^* \rangle \right) \tag{8}$$

Taking a step further [21] showcases that the regret upper bound can be further tightened to nearly $\mathcal{O}(\sqrt{s_0 T})$, where $s_0 : \|\boldsymbol{\theta}^*\|_0 \leq s_0$ is the support of $\boldsymbol{\theta}^*$.

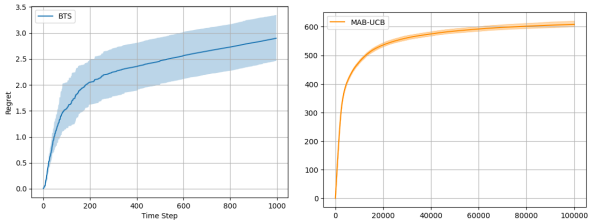
4.2. Connection to RPT decomposition

Let $\mathbf{K}$ be the RPT dictionary and $\mathbf{x}_i$ be the corresponding sparse vector associated with the arm i with support set S_i. We can map our problem setup to linear bandits as follows:

i) Construct the block diagonal matrix $\mathbf{K}_\mathcal{K} = \text{diag}(\mathbf{K}, \mathbf{K}, \dots)$, where each $\mathbf{K} \in \mathbb{R}^{T \times \phi(P_{\max})}$ are blocks on the diagonal and constructed as per Equation (3). For an arm i, the arm features, at time t is $\mathbf{a}_{t,i} = \mathbf{K}_\mathcal{K}[i * T + t]$, which is the $(i * T + t)^{th}$ row of $\mathbf{K}_\mathcal{K}$.

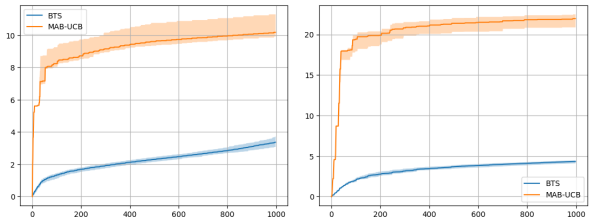
ii) The unknown feature vector $\boldsymbol{\theta}^*$ is a vector stack of $\mathbf{x}_i$, where $\mathbf{x}_i$ is the true solution to the minimization problem (5) in the noiseless case. The reward is obtained as $r_{t,i} = \langle \mathbf{a}_{t,i}, \mathbf{x}_i \rangle$.The pseudocode of the proposed algorithm is provided in Algorithm 1 Following directly from [21], we can provide the following theoretical backing to BTS-RaP:

Theorem 1. *Let $\mathbf{x}_i$ be the sparse representation of a periodic signal under the RPT dictionary with support set $\mathcal{S}_i$ that has $|\mathcal{S}_i|$ nonzero values, for all $\mathbf{x}_i \;\; i \in \mathcal{K}$, then the regret of Bandit Tracking System (BTS-RaP) is upper bounded by $\mathcal{O}(\sqrt{T \sum_{i \in \mathcal{K}} |\mathcal{S}_i|})$.*



(a) (b)

Fig. 2. Regret $\mathcal{R}$ vs time t plots on two armed periodic bandits setting for (a) BTS-RaP and (b) MAB-UCB. Rewards of each arm is generated as per Equation (10).



(a) (b)

Fig. 3. Regret $\mathcal{R}$ vs time t plots on two armed periodic bandits setting for MAB-UCB and BTS-RaP. Rewards of each arm is generated based on (9) with $\{p_1, p_2\}$ taking values (a) $\{7, 3\}$, (b) $\{9, 11\}$

Algorithm 1 Bandit Tracking System (BTS-RaP)

1: Given T, K arms, form $\mathbf{K} \in \mathbb{R}^{T \times \phi(P_{\max})}$ pull each arm once and form the observation vector $\boldsymbol{\mu}_i \ \forall i \in \mathcal{K}$
2: Create mini-dictionaries $\mathbf{K}_i$ for $i \in \mathcal{K}$ which will grow as arms get pulled. Initially each $\mathbf{K}_i$ is of size $1 \times \phi(P_{\max})$
3: Initialize support for each arm $\mathbf{x}_i$ for all $i \in \mathcal{K}$
4: **for** $t = K+1 \ldots T$ **do**
5: Choose arm $a_t = \arg\max_{k \in \{1,\ldots,K\}} \langle \mathbf{K}[t], \mathbf{x}_i \rangle + \sqrt{2\,\alpha\, \ln t / n_{t-1,i}}$, where, $\mathbf{K}[t]$ is the t^{th} row of $\mathbf{K}$
6: Append the row $\mathbf{K}[t]$ to the mini-dictionary $\mathbf{K}_{a_t}$
7: Observe $r_{t,a_t} = \mu_{t,a_t} + \eta_t$ and append this observation to $\boldsymbol{\mu}_{a_t}$
8: Solve : $\min \|\mathbf{D}\mathbf{x}_{a_t}\|_2$ s.t. $\boldsymbol{\mu}_{a_t} = \mathbf{K}_{a_t}\mathbf{x}_{a_t}$ to return updated $\mathbf{x}_{a_t}$
9: **end for**
10: **return** $\{a_t\}_{t=K+1}^{T}$

5. SIMULATION RESULTS

We consider a two-armed bandit setup with three different experiments. In the first and second experiments, we represent the means of the two arms by:

$$\mu_1(t) = c + \sin\left(\frac{2\pi t}{p_1}\right), \ \mu_2(t) = c + \sin\left(\frac{2\pi t}{p_2}\right), \quad (9)$$

where, $t = \{1, 2, \ldots, T\}$ and c is some positive scalar. For the first experiment the tuple $\{p_1, p_2\}$ take the values $\{9, 11\}$ and for the second one it takes the values $\{7, 3\}$.

For the third experiment, we consider periodic mixtures to generate rewards as follows,

$$\mu_1(t) = c + \sum_{i=1}^{3} \sin\left(\frac{2\pi t}{p_i}\right), \ \mu_2(t) = c + \sin\left(\frac{2\pi t}{p}\right), \quad (10)$$

where, for the first arm the periods are $\{p_1, p_2, p_3\} = \{3, 7, 11\}$ and second arm period is $p = 9$. In the first two experiments (Figures 2(a), (b)), we see that our proposed algorithm BTS-RaP outperforms MAB-UCB. While plotting the regret of MAB-UCB we do not consider the stage one (estimation of period) cost. One advantage of using RPT is that even if the period is large, we can still estimate it using RPT with fewer samples, sometimes, even when we have incomplete period length signal as illustrated in Figure 1. While MAB-UCB does achieve sub-linear regret it does so at a very slow pace compared to BTS-RaP. This is because we are selecting an optimal arm from a set of $\sum_k T_k$ arms and not 2 as stated in the problem. This increases the complexity of the MAB problem and is reflected in the regret curve.

The real issue is revealed in the third experiment where one of the arms rewards is a combination of sum of smaller periodic signals (7,3,11). Therefore, the resulting signal is a 231 length period signal. The second arm is a single period signal with period 9. So effectively, MAB-UCB algorithm has 240 effective arms to select from. Whereas, BTS-RaP effectively learns non-zero coordinates of the support vector $\mathbf{x}$ associated with each arm. This vector as highlighted earlier is sparse and the regret scales with the sum ℓ_0-norm of this support vector. Therefore, as seen in Figure 3, BTS-RaP achieves minimum regret quickly and MAB-UCB has to run for significantly longer time ($\sim 100\times$) to start learning the periodic pattern.

6. CONCLUSION

In this paper, we consider bandits that exhibits periodicity. We incorporated the periodic structure of the rewards and proposed an algorithm to minimize the regret. To this end, we utilized the newly introduced, Ramanujan-based periodicity estimation techniques to sequentially update the estimate of the periods of each arm, and subsequently select the best arm at each time step. Our results indicates that our RPT-based method dubbed BTS-RaP, can achieve sub-linear regret.

7. REFERENCES

[1] Aleksandrs Slivkins et al., "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[2] John Langford and Tong Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," *Advances in neural information processing systems*, vol. 20, 2007.

[3] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, Eds. 2011, vol. 24, Curran Associates, Inc.

[4] Omar Besbes, Yonatan Gur, and Assaf Zeevi, "Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards," *Stochastic Systems*, vol. 9, no. 4, pp. 319–337, 2019.

[5] Aurélien Garivier and Eric Moulines, "On upper-confidence bound policies for non-stationary bandit problems," *arXiv preprint arXiv:0805.3415*, 2008.

[6] Aleksandrs Slivkins and Eli Upfal, "Adapting to a changing environment: the Brownian restless bandits.," in *COLT*, 2008, pp. 343–354.

[7] Jenely Villamediana, Inés Küster, and Natalia Vila, "Destination engagement on facebook: Time and seasonality," *Annals of Tourism Research*, vol. 79, pp. 102747, 2019.

[8] Giuseppe Di Benedetto, Vito Bellini, and Giovanni Zappella, "A linear bandit for seasonal environments," 2020.

[9] Gerlando Re, Fabio Chiusano, Francesco Trovò, Diego Carrera, Giacomo Boracchi, and Marcello Restelli, "Exploiting history data for nonstationary multi-armed bandit," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*. Springer, 2021, pp. 51–66.

[10] Xiang Zhou, Yi Xiong, Ningyuan Chen, and Xuefeng Gao, "Regime switching bandits," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4542–4554, 2021.

[11] Qinyi Chen, Negin Golrezaei, and Djallel Bouneffouf, "Dynamic bandits with an auto-regressive temporal structure," 2023.

[12] Hengrui Cai, Zhihao Cen, Ling Leng, and Rui Song, "Periodic-gp: Learning periodic world with gaussian process bandits," 2021.

[13] Ningyuan Chen, Chun Wang, and Longlin Wang, "Learning and optimization with seasonal patterns," *CoRR*, vol. abs/2005.08088, 2020.

[14] S. V. Tenneti and P. P. Vaidyanathan, "Nested periodic matrices and dictionaries: New signal representations for period estimation," *IEEE Transactions on Signal Processing*, vol. 63, no. 14, pp. 3736–3750, 2015.

[15] P. P. Vaidyanathan, "Ramanujan sums in the context of signal processing—part I : Fundamentals," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4145–4157, 2014.

[16] P. Saidi, A. Vosoughi, and G. K. Atia, "Detection of brain stimuli using Ramanujan periodicity transforms," *Journal of Neural Engineering*, vol. 16, no. 3, pp. 036021, 2019.

[17] S. V. Tenneti and P. P. Vaidyanathan, "Detecting tandem repeats in DNA using Ramanujan filter bank," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2016, pp. 21–24.

[18] S. Ramanujan, "On certain trigonometrical sums and their applications in the theory of numbers," *Trans. Cambridge Philosoph. Soc*, vol. XXll, no. 13, pp. 259–276, 1918.

[19] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[20] Sébastien Bubeck, Ofer Dekel, Tomer Koren, and Yuval Peres, "Bandit convex optimization: $\sqrt{T}$ regret in one dimension," in *Conference on Learning Theory*. PMLR, 2015, pp. 266–278.

[21] Yining Wang, Yi Chen, Ethan X. Fang, Zhaoran Wang, and Runze Li, "Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection," 2020.