Experimental Design for Regret Minimization in Linear Bandits

Andrew Wagenmaker* **Julian Katz-Samuels*** **Kevin Jamieson**
Paul G. Allen School of Computer Science and Engineering, University of Washington
{ajwagen,jkatzsam,jamieson}@cs.washington.edu, *denotes equal contribution

Abstract

In this paper we propose a novel experimental design-based algorithm to minimize regret in online stochastic linear and combinatorial bandits. While existing literature tends to focus on optimism-based algorithms–which have been shown to be suboptimal in many cases–our approach carefully plans which action to take by balancing the tradeoff between information gain and reward, overcoming the failures of optimism. In addition, we leverage tools from the theory of suprema of empirical processes to obtain regret guarantees that scale with the Gaussian width of the action set, avoiding wasteful union bounds. We provide state-of-the-art finite time regret guarantees and show that our algorithm can be applied in both the bandit and semi-bandit feedback regime. In the combinatorial semi-bandit setting, we show that our algorithm is computationally efficient and relies only on calls to a linear maximization oracle. In addition, we show that with slight modification our algorithm can be used for pure exploration, obtaining state-of-the-art pure exploration guarantees in the semi-bandit setting. Finally, we provide, to the best of our knowledge, the first example where optimism fails in the semi-bandit regime, and show that in this setting our algorithm succeeds.

1 INTRODUCTION

Multi-armed bandits have received much attention in recent years as they serve as an excellent model for developing algorithms that adeptly deal with the exploration-exploitation tradeoff. In this paper, we consider the stochastic linear bandit problem in which there is a set of arms $\mathcal{X} \subset \mathbb{R}^d$ and an unknown parameter $\theta_* \in \mathbb{R}^d$. An agent plays a sequential game where at each round t she chooses an arm $x_t \in \mathcal{X}$ and receives a noisy reward whose mean is $x_t^\top \theta_*$. The goal is to maximize the reward over a given time horizon T. An important special case of stochastic linear bandits is the combinatorial setting where $\mathcal{X} \subset \{0,1\}^d$, which can be used to model problems such as finding a shortest path in a graph or the best weighted matching in a bipartite graph. We consider both the *bandit feedback* setting, where the agent receives a noisy observation of $x_t^\top \theta_*$, and the *semi-bandit feedback* setting, where the agent receives a noisy observation of $\theta_{*,i}$ for each i with $x_{t,i} = 1$.

Existing regret minimization algorithms for linear bandits suffer from several important shortcomings. First, they typically rely on naive union bounds, which yield regret guarantees scaling as either $\mathcal{O}(d\sqrt{T})$ or $\mathcal{O}(\sqrt{d\log(|\mathcal{X}|)T})$. Such union bounds ignore the geometry present in the problem and, as such, can be very wasteful. As the union bound often appears in the confidence interval within the algorithm, this is not simply an analysis issue—it can also affect real performance. Second, in the moderate, non-asymptotic time regime, existing algorithms tend to rely on the principle of *optimism*—pulling only the arms they believe may be optimal. Algorithms relying on this principle are very myopic, foregoing initial exploration which could lead to better long-term reward and instead focusing on obtaining short-term reward, leading to suboptimal long-term performance. This is a well-known effect in the bandit setting but, as we show, is also present in the semi-bandit setting.

In this paper, we develop an algorithm overcoming both of these shortcomings. Rather than employing a naive union bound, we appeal to tools from empirical process theory for controlling the suprema of a Gaussian process, allowing us to obtain confidence bounds that are geometry-dependent and potentially much tighter. In addition, our algorithm relies on care-

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

ful planning to balance the exploration-exploitation tradeoff, taking into account both the potential information gain as well as the reward obtained when pulling an arm. This planning allows us to collect sufficient information for good long-term performance without incurring too much initial regret and, to the best of our knowledge, is the first planning-based algorithm in the linear bandit setting that provides finite-time guarantees.

We emphasize that we are interested in the non-asymptotic regime and aim to optimize the whole regret bound, including lower-order terms. While several recent works achieve instance-optimal regret, they suffer from loose lower-order terms which dominate the regret for small to moderate T. Our results aim to minimize such terms through employing tighter union bounds. We summarize our contributions:

- We develop a single, general algorithm that achieves a state-of-the-art finite-time regret bound in stochastic linear bandits, in combinatorial bandits with bandit feedback, and in combinatorial bandits with semi-bandit feedback. In addition, our framework is general enough to extend to settings as diverse as partial monitoring and graph bandits.
- We show that in the combinatorial semi-bandit regime, our algorithm is computationally efficient, relying only on calls to a linear maximization oracle, and state-of-the-art, yielding a significant improvement on existing works in the non-asymptotic time horizon regime.
- We give the first example for combinatorial bandits with semi-bandit feedback that shows that optimistic strategies such as UCB and Thompson Sampling can do arbitrarily worse than the asymptotic lower bound, and show that our algorithm improves on optimism in this setting by an arbitrarily large factor.
- As a corollary, we obtain the first computationally efficient algorithm for pure exploration in combinatorial bandits with semi-bandit feedback, and achieve a state-of-the-art sample complexity.

This work can be seen as obtaining *problem-dependent minimax* bounds—minimax bounds that depend on the arm set but hold for all values of the reward vector—and are similar in spirit to the bounds on regret minimization in MDPs given by Zanette and Brunskill [2019]. For some favorable arm sets $\mathcal{X}$, our bounds are tighter than prior $\mathcal{X}$-independent minimax bounds by large dimension factors. To the best of our knowledge, we are the first to obtain such geometry-dependent minimax bounds for linear bandits.

2 PRELIMINARIES

Let $\text{diam}(\mathcal{X}) = \max_{x,y\in\mathcal{X}} \|x-y\|_2$ denote the diameter of $\mathcal{X} \subseteq \mathbb{R}^d$. $\text{diag}(X)$ will refer to the operator which sets all elements in a matrix X not on the diagonal to 0. $\widetilde{\mathcal{O}}(.)$ hides logarithmic terms. $\triangle_{\mathcal{X}} := \{a \in \mathbb{R}^{|\mathcal{X}|} : \|a\|_1 = 1,\, a_i \geq 0\, \forall i\}$ denotes the simplex over $\mathcal{X}$. We use $\lambda \in \triangle_{\mathcal{X}}$ to refer to probability distributions over $\mathcal{X}$ and λ_x to denote the probability on $x \in \mathcal{X}$. We let $\tau \in [0,\infty)^{|\mathcal{X}|}$ refer to allocations over $\mathcal{X}$ and, similarly, τ_x to denote the weight on $x \in \mathcal{X}$. We will somewhat interchangeably use τ to refer to the vector in $\mathbb{R}^{|\mathcal{X}|}$ and the sum of its elements, $\sum_{x\in\mathcal{X}} \tau_x$, but it will always be clear from context which we are referring to. If $x \in \{0,1\}^d$, we will often write $i \in x$ for $x_i = 1$ and $i \notin x$ for $x_i = 0$. Throughout, we will let d denote the dimension of the ambient space and $k = \max_{x\in\mathcal{X}} \|x\|_1$.

We are interested primarily in regret minimization in linear bandits. Given some set $\mathcal{X} \subseteq \mathbb{R}^d$, at every timestep we choose $x_t \in \mathcal{X}$ and receive reward $x_t^\top \theta_*$, for some unknown $\theta_* \in \mathbb{R}^d$. We will define regret as:

$$\mathcal{R}_T = T \max_{x\in\mathcal{X}} x^\top \theta_* - \sum_{t=1}^{T} x_t^\top \theta_*$$

Throughout, we assume that $\theta_* \in [-1,1]^d$. We consider two observation models: semi-bandit feedback and bandit feedback. In the bandit feedback setting, at every timestep we observe:

$$y_t = x_t^\top \theta_* + \eta_t$$

where $\eta_t \sim \mathcal{N}(0,1)$. In the semi-bandit feedback setting, we assume that our bandit instance is combinatorial, $\mathcal{X} \subseteq \{0,1\}^d$, and at every timestep we observe:

$$y_{t,i} = \theta_{*,i} + \eta_{t,i}, \quad \forall i \in x_t$$

where $\eta_t \sim \mathcal{N}(0, I)$. Note that, while we assume Gaussian noise for simplicity, all our results will hold with sub-Gaussian noise [Katz-Samuels et al., 2020].

In the bandit setting, after T observations, our estimate of θ_* will be the standard least squares estimate:

$$\hat{\theta} = \left(\sum_{t=1}^{T} x_t x_t^\top\right)^{-1} \sum_{t=1}^{T} x_t y_t$$

In the semi-bandit setting, we will estimate θ_* coordinate-wise, forming the estimate:

$$\hat{\theta}_i = \frac{1}{T_i} \sum_{t=1, x_{t,i}=1}^{T} y_{t,i}$$

where T_i is the number of times $x_{t,i} = 1$. We denote:

$$A_{\text{band}}(\lambda) = \sum_{x\in\mathcal{X}} \lambda_x x x^\top, A_{\text{semi}}(\lambda) = \text{diag}\left(\sum_{x\in\mathcal{X}} \lambda_x x x^\top\right)$$

For convenience we assume the optimal arm is unique and denote it by x_*. As is standard, we denote the gap of arm x by $\Delta_x := \theta_*^\top(x_* - x)$. We denote the minimum gap as $\Delta_{\min} = \min_{x\in\mathcal{X}\,:\,\Delta_x>0}\Delta_x$ and the maximum gap by $\Delta_{\max} = \max_{x\in\mathcal{X}}\Delta_x$.

In the combinatorial setting, $|\mathcal{X}|$ can often be exponentially large in the dimension, making computational efficiency non-trivial since $\mathcal{X}$ cannot be efficiently enumerated. As such, much of the literature on combinatorial bandits has focused on obtaining algorithms that rely only on an argmax oracle:

$$\text{ORACLE}(v) = \arg\max_{x\in\mathcal{X}} x^\top v$$

Efficient argmax oracles are available in many settings, for instance finding the minimum weighted matching in a bipartite graph and finding the shortest path in a directed acyclic graph.

3 MOTIVATING EXAMPLES

Before presenting our algorithm and main results, we present several examples that motivate the necessity of planning and the wastefulness of naive union bounds, and illustrate how our algorithm is able to make improvements in both these aspects.

First, we show that an optimistic strategy cannot be optimal for combinatorial bandits with semi-bandit feedback. Consider a generic optimistic algorithm that maintains an estimate $\widehat{\theta}_t$ of θ at round t and selects the maximizer of an upper confidence bound, $x_t = \arg\max_{x\in\mathcal{X}} x^\top\widehat{\theta}_t + \text{CB}(x, \{x_s\}_{s=1}^{t-1})$. We make two assumptions on the confidence bound $\text{CB}(\cdot,\cdot)$. First, we assume that $\mathbb{P}[\exists t \le T, \exists x \in \mathcal{X} : |x^\top(\widehat{\theta}_t - \theta)| > \text{CB}(x, \{x_s\}_{s=1}^{t-1})] \le 1/T$. Second, we assume that the confidence bound is at least as good as a confidence bound formed from taking the least squares estimate

$$\text{CB}(x, \{x_s\}_{s=1}^{t-1}) \le \sqrt{\alpha \|x\|^2_{(\sum_{s=1}^{t-1} x_s x_s^\top)^{-1}} \log(T)}$$

where $\alpha > 0$ is a universal constant. We call this algorithm the *generic optimistic algorithm* and let $\mathcal{R}_T^{\text{optimism}}$ denote its regret. Then we have the following.

Proposition 1. *Fix any $m \in \mathbb{N}$ and $\epsilon \in (0,1)$. Then there exists a $\mathcal{O}(m)$-dimensional combinatorial bandit problem with semi-bandit feedback where:*

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathcal{R}_T^{\text{optimism}}]}{\log(T)} = \Omega\left(\frac{m}{\epsilon}\right).$$

and Algorithm 1 has expected regret bounded as, for any T:

$$\mathbb{E}[\mathcal{R}_T] \le \mathcal{O}\left(\min\left\{\frac{\sqrt{m}\log(T)}{\epsilon^2}, \frac{m\log(T)}{\epsilon}\right\}\right).$$

Thus, treating ϵ as a constant, the asymptotic regret of the generic optimistic algorithm is loose by a square root dimension factor, and Algorithm 1 in the current paper improves over optimism by an arbitrarily large factor. As it also relies on the principle of optimism, albeit in a randomized fashion, Thompson Sampling will be suboptimal by this same factor on this instance. A similar instance can also be found in the bandit feedback setting. The improvement in Algorithm 1 is due to its ability to pull informative but suboptimal arms if the information gain outweighs the regret incurred, reducing the cumulative regret. Optimistic algorithms, in contrast, will only pull arms they believe may be optimal, and so do not effectively take into account the information gain which, in some cases, causes them to be very suboptimal.

To illustrate the improvement we gain by applying a less naive union bound, we will consider the following combinatorial class:

$$\mathcal{X} = \left\{x \in \{0,1\}^{m+n} : \sum_{i=1}^{m} x_i = k, \sum_{i=m+1}^{m+n} x_i = \ell\right\}$$

where $d = n + m$. This class corresponds to the Cartesian product of a Top-k problem on dimension m and a Top-ℓ problem on dimension n. As we will show, the minimax regret of Algorithm 1 scales with $\bar{\gamma}(A)$, a measure of the Gaussian width of $\mathcal{X}$, as defined below in (1). In contrast, algorithms that apply naive union bounds have regret that scales either with $(m+n)\log|\mathcal{X}|$ or $(m+n)^2$. The following proposition illustrates the improvement in scaling we are able to obtain, as well as the subtle dependence of minimax regret on the geometry of $\mathcal{X}$.

Proposition 2. *For $\mathfrak{f} \in \{\text{band}, \text{semi}\}$, on the product of Top-$k$ instances described above, we have:*

$$\bar{\gamma}(A_{\mathfrak{f}}) \le \mathcal{O}(km + \ell n), \quad \log|\mathcal{X}| \ge \Omega(k+\ell)$$

This implies there exist settings of m, n, k, and ℓ such that the regret of Algorithm 1 with either bandit feedback or semi-bandit feedback will be bounded:

$$\mathbb{E}[\mathcal{R}_T] \le \widetilde{\mathcal{O}}\left(d^{1/2}\sqrt{T}\right)$$

while algorithms employing naive union bounds will achieve regret bounds scaling at best as:

$$\mathbb{E}[\mathcal{R}_T] \le \widetilde{\mathcal{O}}\left(d^{2/3}\sqrt{T}\right).$$

In the appendix we discuss in more detail how the regret scales for specific algorithms in this setting. The regret bound we present for our algorithm in Proposition 2 is in fact state-of-the-art—all other existing algorithms will incur the larger dimension dependence.

4 EXPERIMENTAL DESIGN FOR REGRET MINIMIZATION

4.1 Gaussian Width

Before introducing our algorithm, we present a final concept critical to our results. For a fixed θ_*, let $\mathcal{X}_\epsilon = \{x \in \mathcal{X} \ : \ \Delta_x \leq \epsilon\}$, then, for $\mathfrak{f} \in \{\text{band}, \text{semi}\}$:

$$\bar{\gamma}(A_\mathfrak{f}) = \sup_{\epsilon > 0} \inf_{\lambda \in \triangle_{\mathcal{X}_\epsilon}} \mathbb{E}_\eta \left[\sup_{x \in \mathcal{X}_\epsilon} x^\top A_\mathfrak{f}(\lambda)^{-1/2} \eta \right]^2 \quad (1)$$

Intuitively, $\bar{\gamma}(A_\mathfrak{f})$ is the largest Gaussian width of any subset of $\mathcal{X}$ formed by taking all $x \in \mathcal{X}$ with gap bounded by ϵ. The following results are helpful in giving some sense of the scaling of $\bar{\gamma}(A_\mathfrak{f})$.

Proposition 3. *For any $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathfrak{f} \in \{\text{band}, \text{semi}\}$, we have:*

$$\bar{\gamma}(A_\mathfrak{f}) \leq c \min\{d \log |\mathcal{X}|, d^2\}.$$

Proposition 4. *If $\mathcal{X} \subseteq \{0,1\}^d$, $k = \max_{x \in \mathcal{X}} \|x\|_1$, and $\mathfrak{f} \in \{\text{band}, \text{semi}\}$, then, for $d \geq 3$:*

$$\bar{\gamma}(A_\mathfrak{f}) \leq cdk \log d.$$

Note that these upper bounds are often loose. The following results shows that, in some cases, we pay a d instead of dk.

Proposition 5. *There exists a combinatorial bandit instance in $\mathbb{R}^d$ with $k = \sqrt{d}$ where:*

$$\bar{\gamma}(A_{\text{semi}}) \leq cd \log(d).$$

The Gaussian width is critical in avoiding wasteful union bounds, allowing instead for geometry-dependent confidence intervals. The following confidence interval will form a key piece in our analysis.

Proposition 6 (Tsirelson-Ibragimov-Sudakov Inequality [Katz-Samuels et al., 2020, Tsirelson et al., 1976])**.** *Consider playing arm x τ_x times, where τ is an allocation chosen deterministically. Assume $\mathfrak{f} \in \{\text{band}, \text{semi}\}$ is set to correspond to the type of feedback received and let $\hat{\theta}$ be the least squares estimate of θ_* from these observations. Then, simultaneously for all $x \in \mathcal{X}$, with probability at least $1 - \delta$:*

$$|x^\top(\hat{\theta} - \theta_*)| \leq \mathbb{E}_{\eta \sim \mathcal{N}(0,I)} \left[\sup_{x \in \mathcal{X}} x^\top A_\mathfrak{f}(\tau)^{-1/2} \eta \right] + \sqrt{2 \sup_{x \in \mathcal{X}} \|x\|^2_{A_\mathfrak{f}(\tau)^{-1}} \log(2/\delta)}.$$

4.2 Algorithm Overview

We next present our algorithm, RegretMED, in Algorithm 1. Inspired by several recent algorithms achieving asymptotically optimal regret [Lattimore and Szepesvari, 2017], at every epoch our algorithm finds a new allocation by solving an experimental design problem (2). This minimizes an upper bound on the regret incurred in the epoch while ensuring the allocation produced will explore enough to improve the estimates of the gaps for each arm, thereby balancing exploration and exploitation and allowing us to obtain a tight bound on finite-time regret. We apply the TIS inequality to bound the estimation error of our gaps, which motivates the constraint in (2). Critically, this yields a regret bound scaling with the Gaussian width of the action set.

We define $\text{SPARSE}(\tau, n) : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}^{|\mathcal{X}|}$ to be a function taking as input an allocation and returning a new allocation that is n sparse and approximating the solution to (2). So long as $n \geq d + 1$ in the semi-bandit setting and $n \geq d^2 + d + 1$ in the bandit setting, it is possible to find a distribution α that is n sparse and will achieve the same value of the constraint and objective of (2), see Lemma 1. $\text{MINGAP}(\theta, \mathcal{X})$ takes as input an estimate of θ and returns the gap between the

Algorithm 1 **R**egret **M**inimizing **E**xperimental **D**esign: **RegretMED**

Input: Set of arms $\mathcal{X}$, largest gap $\Delta_{\max}$, confidence δ, total time T, feedback type $\mathfrak{f} \in \{\text{band}, \text{semi}\}$
$\hat{\theta}_0 \leftarrow 0, x_1 \leftarrow 0, \hat{\Delta}_x \leftarrow 0, \ell \leftarrow 1$
while total pulls less than T **do**
  $\epsilon_\ell \leftarrow \Delta_{\max} 2^{-\ell}$
  Let τ_ℓ be a solution to:

$$\begin{aligned} &\arg\min_{\tau} \sum_{x \in \mathcal{X}} 2(\epsilon_\ell + \hat{\Delta}_x)\tau_x \\ &\text{s.t. } \mathbb{E}_\eta \left[\max_{x \in \mathcal{X}} \frac{(x_\ell - x)^\top A_\mathfrak{f}(\tau)^{-1/2} \eta}{\epsilon_\ell + \hat{\Delta}_x} \right] \\ &\quad + \sqrt{2 \sup_{x \in \mathcal{X}} \frac{\|x\|^2_{A_\mathfrak{f}(\tau)^{-1}}}{(\epsilon_\ell + \hat{\Delta}_x)^2} \log(2\ell^3/\delta)} \leq \frac{1}{128} \end{aligned} \quad (2)$$

  if $\sum_{x \in \mathcal{X}} (\epsilon_\ell + \hat{\Delta}_x)\tau_{\ell,x} > T\epsilon_\ell$ **then**
    break
  end if
  $\alpha_\ell \leftarrow \text{SPARSE}(\tau_\ell, n_\mathfrak{f})$
  Pull arm x $\lceil \alpha_{\ell,x} \rceil$ times, compute $\hat{\theta}_\ell$
  $x_{\ell+1} \leftarrow \arg\max_{x \in \mathcal{X}} x^\top \hat{\theta}_\ell$, $\hat{\Delta}_x \leftarrow \hat{\theta}_\ell^\top (x_{\ell+1} - x)$
  if $\text{MINGAP}(\hat{\theta}_\ell, \mathcal{X}) > 2\epsilon_\ell$ **then**
    break
  end if
  $\ell \leftarrow \ell + 1$
end while
Pull $\hat{x} = \arg\max_{x \in \mathcal{X}} x^\top \hat{\theta}_{\ell-1}$ for all remaining time

best and second best arms in $\mathcal{X}$ with respect to this θ. It is possible to compute this quantity efficiently with only calls to a linear maximization oracle (see Appendix C).

While Algorithm 1 takes as input $\Delta_{\max}$, we require this only to simplify the analysis. In practice, we can use an upper bound instead without changing the final regret of our algorithm by more than a logarithmic factor. Since $\Delta_{\max} \leq \sqrt{d}\text{diam}(\mathcal{X})$, an upper bound can be obtained without knowledge of θ_*.

Key Theoretical Tools: We briefly describe the key theoretical tools employed by RegretMED. First, we note that an experimental design based algorithm is novel in the setting of regret minimization. As we have shown, this approach allows us to perform properly on challenging instances by explicitly balancing the information gain and reward, while also yielding a computationally feasible solution in the semi-bandit regime. Our second innovation is the use of the TIS inequality to obtain tight concentration bounds. While we are not the first to utilize this in the linear bandit setting [Katz-Samuels et al., 2020], it previously was only utilized in the best arm identification setting, and our work therefore shows how it can be applied in the regret minimization setting as well. The use of the TIS inequality yields two important improvements over more naive union bounds. First, it provides tighter confidence intervals in the non-asymptotic time regime and therefore yields improved regret bounds. Second, as we will see, it allows us to write the constraint for our experiment design problem (2) in a form that is linear in the the decision variable. This allows us to reduce solving the optimization to calls of a linear maximization oracle, and is a key piece in showing our algorithm is computationally efficient.

4.3 Main Regret Bound

We now state our main regret bound. Define

$$\ell_{\max}(T) := \log_2\left(\frac{\max_{x\in\mathcal{X}}\|x\|_2}{\min_{x\in\mathcal{X}}\|x\|_2}\left(\Delta_{\max}\sqrt{T}+3\right)\right) = \mathcal{O}(\log(T))$$

and $\ell_{\max}(\theta_*) := \lceil \log(4\Delta_{\max}/\Delta_{\min})\rceil$. Let $n_{\text{band}} = d^2 + d + 1, n_{\text{semi}} = d + 1$.

Theorem 1. *With $\mathfrak{f} \in \{\text{band}, \text{semi}\}$ set to correspond to the type of feedback received, Algorithm 1 will have gap-dependent regret bounded, with probability $1-\delta$, as:*

$$c_1\Delta_{\max}\ell_{\max}(\theta_*)^2(d+n_{\mathfrak{f}}) + \frac{c_2\left(\bar{\gamma}(A_{\mathfrak{f}})\ell_{\max}(\theta_*)^2 + d\log(\ell_{\max}(\theta_*)/\delta)\right)}{\Delta_{\min}}$$

and minimax regret bounded as:

$$c_1\Delta_{\max}\ell_{\max}(T)^2(d+n_{\mathfrak{f}}) + \ell_{\max}(T)\sqrt{c_2(\bar{\gamma}(A_{\mathfrak{f}})\ell_{\max}(T)^2 + d\log(\ell_{\max}(T)/\delta))T}$$

for absolute constants c_1 and c_2.

The proof of this result is deferred to Appendix B. See Section 6 and Table 1 for a summary of how this bound scales in particular settings of interest. As a brief comparison, in the semi-bandit feedback setting, considering expected regret, we obtain a leading term of order $\mathcal{O}\left(\frac{d\log(T)}{\Delta_{\min}}\right)$, which matches the lower bound [Degenne and Perchet, 2016], while the previous state-of-the-art scaled as $\mathcal{O}\left(\frac{d\log^2(k)\log(T)}{\Delta_{\min}}\right)$ [Perrault et al., 2020a]. Algorithm 1 is then the first algorithm to achieve the lower bound for arbitrary combinatorial structures. In the bandit feedback setting our minimax regret scales as $\widetilde{\mathcal{O}}(\sqrt{(\bar{\gamma}(A_{\text{band}})+d)T})$ while LinUCB obtains regret scaling as $\widetilde{\mathcal{O}}(d\sqrt{T})$ [Abbasi-Yadkori et al., 2011]. Proposition 3 shows that we are never worse than the LinUCB regret and, as Proposition 2 shows, we can sometimes be much better. In Appendix A, we present a modified algorithm which avoids the factors of $\ell_{\max}(T)$ on the leading term of the minimax regret, although it suffers from several other shortcomings.

4.4 Computationally Efficient Algorithm

While Algorithm 1 can be run in settings where $\mathcal{X}$ is enumerable, it becomes computationally infeasible for very large $\mathcal{X}$, as (2) cannot be solved via a linear maximization oracle. In place of (2), consider instead solving:

$$\arg\min_{\tau} \sum_{x\in\mathcal{X}} 2(\epsilon_\ell + \hat{\Delta}_x)\tau_x \quad (3)$$
$$\text{s.t. } \mathbb{E}_\eta\left[\max_{x\in\mathcal{X}} \frac{(x_\ell - x)^\top A(\tau)^{-1/2}\eta}{\epsilon_\ell + \hat{\Delta}_x}\right] \leq \frac{1/128}{1+\sqrt{\pi\log(2\ell^3/\delta)}}$$

As we show in Theorem 4, we can solve this problem with a computationally feasible algorithm in the semi-bandit feedback regime. Running this modified version of Algorithm 1, we obtain the following regret bound.

Theorem 2. *Assume $\mathfrak{f} \in \{\text{band}, \text{semi}\}$ is set to correspond to the type of feedback received. Consider running Algorithm 1 but now setting τ_ℓ to be an approximate solution to (3). Then with probability at least $1-\delta$, the gap-dependent regret will be bounded as:*

$$c_1\Delta_{\max}\ell_{\max}(\theta_*)^2(d+n_{\mathfrak{f}}) + \frac{c_2\bar{\gamma}(A_{\mathfrak{f}})\log(\ell_{\max}(\theta_*)/\delta)\ell_{\max}(\theta_*)^2}{\Delta_{\min}}$$

and the minimax regret will be bounded as:

$$c_1\Delta_{\max}\ell_{\max}(T)^2(d+n_{\mathfrak{f}}) + \ell_{\max}(T)^2\sqrt{c_2\bar{\gamma}(A_{\mathfrak{f}})\log(\ell_{\max}(T)/\delta)T}$$

for absolute constants c_1, c_2.

In the semi-bandit setting, we can apply Theorem 4 to compute an approximate solution to (3) in polynomial time, as described below. See Section 6 and Table 1 for an in-depth discussion of how our result compares to existing works.

Note that the minimax regret guarantees given in Theorems 1 and 2 depend on θ_* through $\bar{\gamma}(A_{\mathfrak{f}})$ and $\Delta_{\max}$. This dependence can be removed by simply taking a supremum of $\bar{\gamma}(A_{\mathfrak{f}})$ over θ_* and using the upper bound $\Delta_{\max} \le \sqrt{d}\mathrm{diam}(\mathcal{X})$. While we state our regret bounds in high probability, expected regret bounds can also be obtained by setting $\delta = 1/T$.

4.5 Pure Exploration with Semi-Bandit Feedback

Although our algorithm is designed to minimize regret, a slight modification gives a computationally efficient algorithm for best arm identification in the semi-bandit feedback setting. In particular, instead of (2) consider solving:

$$\arg\min_{\tau} \sum_{x\in\mathcal{X}} \tau_x \quad (4)$$
$$\text{s.t. } \mathbb{E}_\eta\left[\max_{x\in\mathcal{X}} \frac{(x_\ell - x)^\top A_{\mathfrak{f}}(\tau)^{-1/2}\eta}{\epsilon_\ell + \tilde{\Delta}_x}\right] \le \frac{1/128}{1+\sqrt{\pi\log(2\ell^3/\delta)}}$$

Then we have the following.

Theorem 3. *Define*

$$\rho^* := \inf_{\lambda\in\triangle}\sup_{x\in\mathcal{X}\setminus\{x_*\}} \frac{\|x_* - x\|^2_{A_{\mathrm{semi}}(\lambda)^{-1}}}{[\theta_*^\top(x_* - x)]^2}$$
$$\gamma^* := \inf_{\lambda\in\triangle}\mathbb{E}_\eta\left[\sup_{x\in\mathcal{X}\setminus\{x_*\}} \frac{(x_* - x)^\top A_{\mathrm{semi}}(\lambda)^{-1/2}\eta}{\theta_*^\top(x_* - x)}\right]^2$$

Let $\delta \in (0,1)$. Run Algorithm 1 but replace (2) with (4) and omit the break on line 7. Invoke Theorem 4 to efficiently find an approximate solution to (4). Then, with probability $1-\delta$, the algorithm will terminate after collecting at most:

$$c\Big([\gamma^* + \rho^*]\log(\ell_{\max}(\theta_*)/\delta) + d\Big)\ell_{\max}(\theta_*)$$

samples and we will have $\hat{x} = x_*$.

We state and prove a lower bound for this problem in the appendix, Theorem 6, which shows that this sample complexity is near-optimal. To the best of our knowledge, this is the first general, computationally efficient, and near optimal algorithm for pure exploration with semi-bandit feedback.

4.6 Optimization

In this section, we provide a polynomial-time algorithm for solving (3) in the semi-bandit feedback setting. The generic optimization problem can be written as follows for a fixed $T \in \mathbb{N}$, $\bar{x} \in \mathcal{X}$, $\bar{\theta} \in \mathbb{R}^d$, and $\beta > 0$:

$$\min_{\tau\in[T],\lambda\in\triangle_{\mathcal{X}}} \tau\sum_{x\in\mathcal{X}} \bar{\theta}^\top(\bar{x} - x)\lambda_x + \tau\beta \quad (5)$$
$$\text{s.t. } \mathbb{E}_\eta\left[\max_{x\in\mathcal{X}} \frac{(\bar{x} - x)^\top A_{\mathrm{semi}}(\lambda)^{-1/2}\eta}{\beta + \bar{\theta}^\top(\bar{x} - x)}\right] \le \sqrt{\tau}C.$$

The following result shows that there exists a polynomial-time algorithm that finds an approximately optimal solution, i.e., it is within a constant approximation factor of the optimal solution.

Theorem 4. *Let* OPT *be the optimal value of (5). There exists an Algorithm that returns $(\bar{\tau}, \bar{\lambda})$ such that $\bar{\lambda} \in \triangle_{\mathcal{X}}$, $\bar{\tau} \le 2T$, and, with probability at least $1 - \delta - \frac{1}{2^d}$:*

$$\bar{\tau}\sum_{x\in\mathcal{X}} \bar{\theta}^\top(\bar{x} - x)\bar{\lambda}_x + \bar{\tau}\beta \le 4\mathrm{OPT} + 2$$
$$\mathbb{E}_\eta\left[\max_{x\in\mathcal{X}} \frac{(\bar{x} - x)^\top A_{\mathrm{semi}}(\bar{\lambda})^{-1/2}\eta}{\beta + \bar{\theta}^\top(\bar{x} - x)}\right] \le \sqrt{\bar{\tau}}C.$$

Furthermore, the number linear maximization oracle calls is polynomial in $(d, \beta, T, \log(1/\delta))$.

We briefly sketch the algorithmic approach. We recast (5) as a series of feasibility problems and employ the Plotkin-Shmoys-Tardos reduction of convex feasibility programs to online learning to solve each of these feasibility programs using the multiplicative weights update algorithm. To employ this reduction, we fix τ and develop a solver for the Lagrangian of (5), $\mathcal{L}(\kappa;\lambda)$, which we show to be convex and strongly-smooth in λ over a carefully constructed subset of the simplex $\tilde{\triangle}_{\mathcal{X}} \subset \triangle_{\mathcal{X}}$. We solve $\min_{\lambda\in\tilde{\triangle}_{\mathcal{X}}}\mathcal{L}(\kappa;\lambda)$ by employing stochastic Frank-Wolfe, which maintains sparse iterates to overcome the challenge posed by the exponential number of variables in $\mathcal{L}(\kappa;\lambda)$. Evaluating the gradient requires computing for $\eta \sim N(0, I)$

$$\arg\max_{x\in\mathcal{X}} \frac{(\bar{x} - x)^\top A_{\mathrm{semi}}(\lambda)^{-1/2}\eta}{\beta + \bar{\theta}^\top(\bar{x} - x)},$$

which can be solved using only linear maximization oracle calls via the binary search procedure from Katz-Samuels et al. [2020]. The proof of this result and full algorithm is given in Section D.

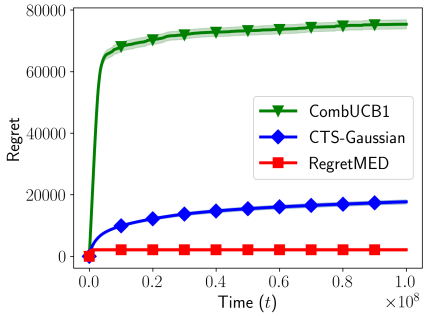

Figure 1: Resource allocation example with $d = 5$.

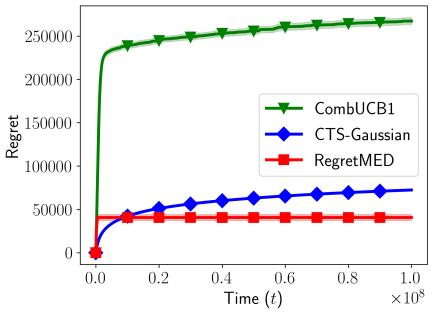

Figure 2: Resource allocation example with $d = 25$.

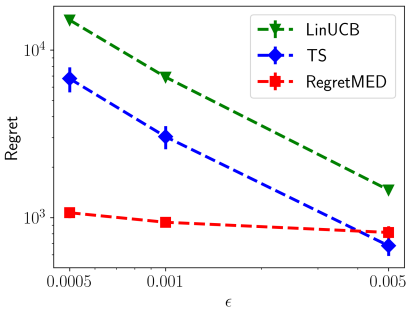

Figure 3: End of Optimism example varying ϵ.

Rounding: The allocation τ_ℓ is not integer, so must be rounded. Naive rounding could incur problematically large regret, so we instead seek a sparse allocation, which will allow us to round without incurring significant regret. Recalling that $n_{\text{band}} = d^2 + d + 1, n_{\text{semi}} = d + 1$, we have:

Lemma 1. *Given τ_ℓ a solution to (2) or (3), there exists an n_{f}-sparse α_ℓ which achieves the same value of the constraint and objective of (2) or (3), respectively. Furthermore, in the semi-bandit setting, if we run the procedure of Theorem 4 to find an approximate solution to (3), we can compute α_ℓ in time* $\text{poly}(d, \Delta_{\min}, T, 1/\delta)$.

We prove this result and state how this rounded distribution can be computed in Appendix E.

5 EXPERIMENTAL RESULTS

We next present experimental results for RegretMED in both the semi-bandit and bandit feedback settings. Every point in each plot is the average of 50 trials. The error bars indicate one standard error.

Semi-Bandit Feedback: We compare the computationally efficient version of RegretMED against CombUCB1 [Kveton et al., 2015] and CTS-Gaussian [Perrault et al., 2020a], a formulation of Thompson Sampling in the semi-bandit setting. As a test instance, we consider a resource allocation problem where an agent is tasked with maximizing profit subject to production cost. In particular, assume there are d buyers, each offering a different price for a good. At each timestep the agent can sell to any number of them, but incurs an additional production cost for each item they sell. The agent observes a noisy realization of the price the buyer they sold to is willing to pay and of the production cost. In particular, if at time t we sell to k buyers $x_{t_1}, \ldots, x_{t_k}$, we will pay production costs $y_1, \ldots, y_k$, where y_i is the production cost of producing the ith good. We can model this problem with $\mathcal{X} \subseteq \mathbb{R}^{2d}$, $\theta_{*,1:d}$ corresponding to the prices each buyer will pay, and $\theta_{*,d+1:2d}$ corresponding to the costs, y_i.

We illustrate the result in Figures 1 and 2 for different values of d. In both cases, RegretMED yields a significant improvement over CTS-Gaussian and CombUCB1. Note that $|\mathcal{X}|$ is growing exponentially in d and for $d = 25$ we have $|\mathcal{X}| \approx 3 \cdot 10^7$. In all experiments we set $\delta = 1/T$.

Bandit Feedback: In the bandit setting, we compare against LinUCB [Abbasi-Yadkori et al., 2011] and Thompson Sampling. For Thompson Sampling we use the Bayesian version. We run on the instance described in Lattimore and Szepesvari [2017]. In particular, in this instance $\theta_* = e_1 \in \mathbb{R}^2$ and $\mathcal{X} = \{e_1, e_2, x\}$ where $x = [1 - \epsilon, 8\epsilon]$. We set $\delta = 1/T$ and, for each experiment, use $T = 25/\epsilon^2$, which is the natural scaling for the problem since, as shown in Lattimore and Szepesvari [2017], optimistic algorithms will require on order $1/\epsilon^2$ pulls to determine x is suboptimal. For completeness, in Appendix H we include the plots of regret against time for each point in this figure.

As Figure 3 illustrates, the performance of RegretMED is almost unaffected by the choice of ϵ, while the performance of both TS and LinUCB degrades significantly. Optimistic algorithms are suboptimal on this instance as they do not pull the suboptimal but informative arm, e_2. Our results indicate that RegretMED is able to overcome this difficulty by continuing to pull e_2 even when it has been determined suboptimal, recognizing the information gain outweighs the regret incurred.

6 DISCUSSION AND PRIOR ART

Linear Bandits with Bandit Feedback: Several of the most well-studied algorithms for regret minimization in stochastic linear bandits with bandit feedback are LinUCB [Abbasi-Yadkori et al., 2011], action elimination, and LinTS [Lattimore and Szepesvári, 2020]. LinUCB achieves regret of $\widetilde{\mathcal{O}}(d\sqrt{T})$, action elimination

	Lower Bound	Prior Art	Theorem 1	Theorem 2 (Efficient)
Semi-Bandit	$\Theta\left(\frac{d\log(T)}{\Delta_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\log^2(k)\log(T)}{\Delta_{\min}}+\frac{dk^2\Delta_{\max}}{\Delta^2_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\log(T)}{\Delta_{\min}}+\frac{\bar{\gamma}(A_{\text{semi}})}{\Delta_{\min}}+dk\right)$	$\widetilde{\mathcal{O}}\left(\frac{\log^2(k)\bar{\gamma}(A_{\text{semi}})\log(T)}{\Delta_{\min}}+dk\right)$
Bandit	$\Theta\left(\frac{d\log(T)}{\Delta_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\log(T)+d\log(\lvert\mathcal{X}\rvert)}{\Delta_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d\log(T)+\bar{\gamma}(A_{\text{band}})}{\Delta_{\min}}\right)$	(Not Efficient)

Table 1: Gap-dependent expected regret guarantees in bandit and semi-bandit feedback settings. Note that lower bounds stated hold only for specific instances (e.g. standard multi-armed bandits with equal gaps).

has regret bounded as $\widetilde{\mathcal{O}}(\sqrt{dT\log(|\mathcal{X}|)})$, and Thompson Sampling has (frequentist) regret of $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. Both LinUCB and action elimination rely on wasteful union bounds—LinUCB union bounds over every direction in $\mathbb{R}^d$, incurring an extra factor of $\sqrt{d}$, while action elimination union bounds over every arm without regard to geometry, incurring an extra $\log(|\mathcal{X}|)$. By leveraging tools from empirical process theory, we develop bounds that depend on the fine-grained geometry of $\mathcal{X}$. Indeed, as already stated, our algorithm achieves an expected regret of $\widetilde{\mathcal{O}}(\sqrt{\bar{\gamma}(A_{\text{band}})T})$ which, by Proposition 3, is at least as good as, and in some cases much better than the bounds of LinUCB and action elimination (see Proposition 2). Our bound can be seen as similar in spirit to the problem-dependent minimax bound for regret minimization in MDPs given in Zanette and Brunskill [2019].

Combinatorial Bandits with Semi-Bandit Feedback: Significant attention has been given to the combinatorial semi-bandit problem. Kveton et al. [2015] handles the case where noise is correlated between coordinates, and provides a computationally efficient algorithm with a regret bound of $\widetilde{\mathcal{O}}\left(\frac{dk\log(T)}{\Delta_{\min}}+dk\right)$. Degenne and Perchet [2016] builds on this, showing that if the noise is assumed to be uncorrelated between coordinates, the k on the leading term can be improved to a $\log^2(k)$. Although their algorithm is not computationally efficient, several subsequent works proposed efficient procedures that achieved similar regret bounds [Wang and Chen, 2018, Perrault et al., 2020a, Cuvelier et al., 2020].

We give the first upper bound on regret (Theorem 1) that matches the lower bound on the leading $\log(T)$ term. Prior works are loose by a factor of $\log^2(k)$ and, moreover, have large additive terms that dominate until $T \geq \widetilde{\mathcal{O}}(\exp(\frac{k^2\Delta_{\max}}{\log^2(k)\Delta_{\min}}))$, making their bounds essentially vacuous for all practical time regimes. Although our analysis of the computationally efficient algorithm does not match the lower bound, its leading term is $\frac{dk\log^3(k)\log(T)}{\Delta_{\min}}$ in the worst case, and, due to our smaller additive terms, our regret bound improves on the state of the art until $T \geq \widetilde{\mathcal{O}}(\exp(\frac{k\Delta_{\max}}{\log^3(k)\Delta_{\min}}))$. Furthermore, Proposition 5 implies that there exist instances where Theorem 2 matches the state-of-the-art in the leading term, up to a single $\log(k)$ factor. While we have assumed the noise between coordinates is uncorrelated, RegretMED extends to the case where it is correlated by using $A_{\text{cor}}(\lambda) = \Sigma \circ A_{\text{semi}}(\lambda)^{-1}A_{\text{band}}(\lambda)A_{\text{semi}}(\lambda)^{-1}$ for Σ an upper bound on the noise covariance and $\circ$ denoting element-wise multiplication.

While prior algorithms have tended to be based on the principle of optimism [Kveton et al., 2015, Combes et al., 2015, Degenne and Perchet, 2016, Wang and Chen, 2018, Perrault et al., 2020a], we have shown that optimistic strategies are asymptotically suboptimal (see Proposition 1), motivating our planning-based algorithm. Additional work includes [Chen et al., 2016, Talebi and Proutiere, 2016, Perrault et al., 2020b]. We summarize our results in Table 1.

Asymptotically Optimal Regret in Linear Bandits: Another related line of work focuses on asymptotic performance [Lattimore and Szepesvari, 2017, Combes et al., 2017, Hao et al., 2020, Degenne et al., 2020, Cuvelier et al., 2020]. In the bandit setting asymptotic lower bounds have been shown to scale as:

$$\min_{\tau} \sum_{x\in\mathcal{X}} \Delta_x \tau_x \quad \text{s.t.} \quad \|x\|^2_{A_{\text{band}}(\tau)^{-1}}/\Delta_x^2 \leq \frac{1}{2}, \forall x \neq x_*$$

While we do not claim RegretMED is asymptotically optimal, we note that the optimization we are solving (2) closely resembles the above optimization. Indeed, at the final epoch of RegretMED, our estimates of the gaps will be sufficiently accurate so as to ensure we are playing approximately the asymptotically optimal distribution. Furthermore, as Proposition 1 and Figure 3 show, RegretMED appears to be playing the asymptotically optimal strategy in situations where optimism fails. We leave a rigorous proof of the asymptotic qualities of RegretMED to future work.

Concurrent to this work, several works appeared which simultaneously achieve asymptotically optimal and sub-$\mathcal{O}(\sqrt{T})$ regret [Tirinzoni et al., 2020, Kirschner et al., 2020b]. In particular, Tirinzoni et al. [2020] achieves instance-optimal $\log T$ regret in finite time. We remark that their regret bound contains large additive terms which will dominate the leading $\log T$ term for moderate time horizons. Our primary concern is in this non-asymptotic regime, where the union bound applied is still significant, and we therefore see our work as complementary, addressing issues they do not

address.

Asymptotically optimal regret has been relatively unexplored in the semi-bandit setting. Following the acceptance of this work, a very recent work [Cuvelier et al., 2021] proposed a computationally efficient asymptotically optimal algorithm in the semi-bandit setting, which was the first of its kind. As with the bandit setting, our concern is with the non-asymptotic time regime, so this result is complementary to ours.

Stochastic Multi-Armed Bandits with Side Observations: In the stochastic multi-armed bandits with side observations problem, the agent is given a graph of n nodes where each node is associated with an independent distribution. When the agent pulls a node i, she observes and suffers its stochastic reward and she also observes the stochastic reward of any node with an edge connected to node i. Caron et al. [2012] proposed a UCB-like algorithm and Buccapatnam et al. [2014] used a linear programming solution to show that the regret scales with the minimum dominating set.

Using the design matrix $A_{\text{graph}}(\lambda) = \sum_{i=1}^n \lambda_i \sum_{(i,j)\in E} e_j e_j^\top$, where E denotes the edges in the graph, our algorithmic approach offers an explicit and natural way to model the tradeoff between estimated regret and information gain in this setting. In addition, our work suggests an algorithm for a novel extension of this problem where each node i is associated with a feature vector $x_i \in \mathbb{R}^d$ and the expected reward of i is $\theta_*^\top x_i$, that is, stochastic linear bandits with side observations.

Partial Monitoring: The partial monitoring problem [Cesa-Bianchi and Lugosi, 2006, Cesa-Bianchi et al., 2006, Bartók et al., 2011] is a generalization of the multi-armed bandit problem where now the learner is no longer able to directly observe the loss incurred, but only some function of it. The linear partial monitoring problem [Lin et al., 2014, Kirschner et al., 2020a] is a special case where the learner observes $y_t = z_{x_t}^\top \theta_* + \eta_t$, for some known z_x, but receives reward $x_t^\top \theta_*$, which is not observed. RegretMED directly generalizes to this setting if we employ the design matrix $A_{\text{pm}}(\lambda) = \sum_{x\in\mathcal{X}} \lambda_x z_x z_x^\top$. We leave a full investigation of this application to future work.

Pure Exploration in Multi-Armed Bandits: There has not been a significant amount of previous work on pure exploration combinatorial bandits with semi-bandit feedback. Chen et al. [2020] provide a general framework that subsumes combinatorial bandits with semi-bandit feedback but their algorithm is non-adaptive and suboptimal. Several special cases of pure exploration combinatorial bandits with semi-bandit feedback have been studied. Best arm identification (where $\mathcal{X} = \{e_1, \dots, e_d\}$) has received much attention [Even-Dar et al., 2006, Jamieson et al., 2014, Karnin et al., 2013, Kaufmann et al., 2016, Chen and Li, 2015]. The setting in Jun et al. [2016] subsumes the top-K problem, but their approach does not generalize to other combinatorial problem instances. Concurrent to this work, Jourdan et al. [2021] derived an asymptotically optimal best arm identification algorithm for the semi-bandit setting. We note that our result focuses on optimality in the finite-time regime, so our results our complementary.

Our work is also related to transductive linear bandits [Fiez et al., 2019]. In this problem, there are measurement vectors $\mathcal{X} \subset \mathbb{R}^d$, item vectors $\mathcal{Z} \subset \mathbb{R}^d$, and the agent at each round chooses $x_t \in \mathcal{X}$ and observes the realization of a noisy random variable with mean $x_t^\top \theta$ with the goal to identify $\arg\max_{z\in\mathcal{Z}} \theta^\top z$ as quickly as possible. Our work on combinatorial bandits with semi-bandit feedback can be straightforwardly extended to a generalization of transductive linear bandits that allows for multiple measurements at each round. More concretely, in this setting, the agent is given a collection of subsets of $\mathcal{X}$, $\mathcal{C} \subset 2^{\mathcal{X}}$, and at each round, she chooses a *set* of linear measurements $Y_t \subset \mathcal{X}$ where $Y_t \in \mathcal{C}$, and observes the realization of a noisy random variable with mean $x^\top\theta$ for each $x \in Y_t$. This generalization subsumes the work of Wu et al. [2015], which studies a version of this problem where $\mathcal{X} = \mathcal{Z} = \{e_1, \dots, e_d\}$.

Our algorithmic technique bridging empirical process theory and experimental design is inspired by the work on pure exploration combinatorial bandits in Katz-Samuels et al. [2020]. The semi-bandit feedback setting in the present paper poses a new and non-trivial computational challenge since, unlike in Katz-Samuels et al. [2020], the number of variables in the optimization is potentially exponential in the dimension.

Acknowledgements

AW is supported by an NSF GFRP Fellowship DGE-1762114. JKS is supported by an Amazon Research Award. The work of KJ is supported in part by grants NSF RI 1907907 and NSF CCF 2007036.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization ap-

proach. *Mathematical Programming*, pages 1–40, 2020.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 133–154, 2011.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300, 2014.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolo Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.

Lijie Chen and Jian Li. On the optimal sample complexity for best arm identification. *arXiv preprint arXiv:1511.03774*, 2015.

Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534, 2017.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Wei Chen, Yihan Du, and Yuko Kuroki. Combinatorial pure exploration with partial or full-bandit linear feedback. *arxiv Preprint arXiv:2006.07905v1*, 2020.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2017.

Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Statistically efficient, polynomial time algorithms for combinatorial semi bandits. *arXiv preprint arXiv:2002.07258*, 2020.

Thibaut Cuvelier, Richard Combes, and Eric Gourdin. Asymptotically optimal strategies for combinatorial semi-bandits in polynomial time. *arXiv preprint arXiv:2102.07254*, 2021.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, pages 2972–2980, 2016.

Rémy Degenne, Han Shao, and Wouter M Koolen. Structure adaptive algorithms for stochastic bandits. *arXiv preprint arXiv:2007.00969*, 2020.

Harold Gordon Eggleston. *Convexity*. Number 47. CUP Archive, 1958.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7 (Jun):1079–1105, 2006.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, pages 10667–10677, 2019.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

Kevin Jamieson. Some notes on multi-armed bandits. `https://courses.cs.washington.edu/courses/cse599i/20wi/resources/bandit_notes.pdf`. Accessed: October 14, 2020.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

Marc Jourdan, Mojmír Mutnỳ, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. *arXiv preprint arXiv:2101.08534*, 2021.

Kwang-Sung Jun, Kevin Jamieson, Robert Nowak, and Xiaojin Zhu. Top arm identification in multi-armed bandits with batch arm pulls. In *Artificial Intelligence and Statistics*, pages 139–148, 2016.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 1238–1246. JMLR Workshop and Conference Proceedings, May 2013.

Julian Katz-Samuels, Lalit Jain, Zohar Karnin, and Kevin Jamieson. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *arXiv preprint arXiv:2006.11685*, 2020.

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

Johannes Kirschner, Tor Lattimore, and Andreas Krause. Information directed sampling for linear partial monitoring. *arXiv preprint arXiv:2002.11182*, 2020a.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. *arXiv preprint arXiv:2011.05944*, 2020b.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543, 2015.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tian Lin, Bruno Abrahao, Robert Kleinberg, John Lui, and Wei Chen. Combinatorial partial monitoring game with linear feedback and its applications. In *International Conference on Machine Learning*, pages 901–909, 2014.

Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8307–8318, 2019.

James R Munkres. *Analysis on manifolds*. CRC Press, 2018.

Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:2006.06613*, 2020a.

Pierre Perrault, Michal Valko, and Vianney Perchet. Covariance-adapting algorithm for semi-bandits with application to sparse outcomes. volume 125 of *Proceedings of Machine Learning Research*, pages 3152–3184. PMLR, 09–12 Jul 2020b.

Mohammad Sadegh Talebi and Alexandre Proutiere. An optimal algorithm for stochastic matroid bandit optimization. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 548–556, 2016.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *arXiv preprint arXiv:2010.12247*, 2020.

Boris S Tsirelson, Ildar A Ibragimov, and VN Sudakov. Norms of gaussian sample functions. In *Proceedings of the Third Japan—USSR Symposium on Probability Theory*, pages 20–41. Springer, 1976.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122, 2018.

Yifan Wu, Andras Gyorgy, and Csaba Szepesvari. On identifying good options under combinatorially structured feedback in finite noisy environments. In *International Conference on Machine Learning*, pages 1283–1291, 2015.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.